BIOMEDICAL CONCEPT ASSOCIATION AND CLUSTERING USING WORD EMBEDDINGS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Setu Shah

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Electrical and Computer Engineering

December 2018

Purdue University

Indianapolis, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Xiao Luo, Co-Chair

Department of Computer Information and Graphics Technology

Dr. Mohamed El-Sharkawy, Co-Chair

Department of Electrical and Computer Engineering

Dr. Brian King

Department of Electrical and Computer Engineering

Approved by:

Dr. Brian King

Head of the Graduate Program

To my parents,

Sanjay Shah, and

Dr. Shobha Shah.

ACKNOWLEDGMENTS

Foremost, I would like to thank my mother Dr. Shobha Shah, and my father Sanjay Shah, for their unwavering support and unconditional love. Their belief in me and my abilities never stops astounding me.

To Dr. Xiao Luo for her help and guidance during the completion of this thesis. She has been instrumental in the conception, execution and publication of this work.

To Dr. Mohamed El-Sharkawy for serving as the co-chair on my committee, and for regularly helping keep my thesis on track.

To Dr. Brian King for being a part of my thesis committee and his assistance in planning my Master's coursework.

To Dr. Zina Ben Miled for being my research mentor and sharing her invaluable experiences and knowledge.

To Saravanan Kanakasabai, Ricardo Tuason and Gregory Klopper, from Indiana University Health, for helping me procure data for this research.

To Sherrie Tucker for all her time and assistance since before my admittance into the graduate program, and answering every small question with a smile.

To Srishti Chauhan for always standing by me, believing in me and helping me be the best person I can be.

To Bhavin Shah for making me realize the importance of writing in my life.

Last, but certainly not the least, to Abhijit Katkar, Aniket Udare, Shweta Daule and Tanmay Parashar for being my family away from my family.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

SYMBOLS

$\|A\|$	Number of items in set A
$\|\|A\|\|$	Absolute normalized value of vector A
$\|\|a - b\|\|$	Euclidean distance between a and b
CV	Concept vector
CM	Corpus matrix
DV	Document vector
$s_{i,j}$	Cosine similarity between word vectors WV_i and WV_j
$tfidf_{t,d,D}$	TF-IDF value for a concept t in document d of corpus D
WV	Word vector

ABBREVIATIONS

AI	Artificial Intelligence
BoW	Bag-of-Words
CBOW	Continuous Bag-of-Words
DB Index	Davies–Bouldin Index
EHR	Electronic Health Record
ELMo	Embeddings from Language Models
EMR	Electronic Medical Record
etc.	et cetera
GloVe	Global Vectors for word representations
HIPAA	Health Insurance Portability and Accountability Act
HITECH	The Health Information Technology for Economic and Clinical Health Act
IDF	Inverse Document Frequency
IRB	Institutional Review Board
IU Health	Indiana University Health
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	long short-term memory
MEDLINE	Medical Literature Analysis and Retrieval System Online
NLM	United States National Library of Medicine
NLP	Natural Language Processing
PCA	Principal Component Analysis
PMC	PubMed Central
RNN	recurrent neural network

SOM	Self-Organizing Map
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TREC	Text REtrieval Conference
t-SNE	T-Distributed Stochastic Neighbor Embedding
UMLS	Unified Medical Language System
US	United States
USPSTF	United States Preventive Services Task Force

GLOSSARY

one-hot vector	a vector with a single high (1) value
corpus	a set of documents

ABSTRACT

Shah, Setu. M.S.E.C.E., Purdue University, December 2018. Biomedical Concept Association and Clustering Using Word Embeddings. Major Professors: Xiao Luo and Mohamed El-Sharkawy.

Biomedical data exists in the form of journal articles, research studies, electronic health records, care guidelines, etc. While text mining and natural language processing tools have been widely employed across various domains, these are just taking off in the healthcare space.

A primary hurdle that makes it difficult to build artificial intelligence models that use biomedical data, is the limited amount of labelled data available. Since most models rely on supervised or semi-supervised methods, generating large amounts of pre-processed labelled data that can be used for training purposes becomes extremely costly. Even for datasets that are labelled, the lack of normalization of biomedical concepts further affects the quality of results produced and limits the application to a restricted dataset. This affects reproducibility of the results and techniques across datasets, making it difficult to deploy research solutions to improve healthcare services.

The research presented in this thesis focuses on reducing the need to create labels for biomedical text mining by using unsupervised recurrent neural networks. The proposed method utilizes word embeddings to generate vector representations of biomedical concepts based on semantics and context. Experiments with unsupervised clustering of these biomedical concepts show that concepts that are similar to each other are clustered together. While this clustering captures different synonyms of the same concept, it also captures the similarities between various diseases and the symptoms that those diseases are symptomatic of.

To test the performance of the concept vectors on corpora of documents, a document vector generation method that utilizes these concept vectors is also proposed. The document vectors thus generated are used as an input to clustering algorithms, and the results show that across multiple corpora, the proposed methods of concept and document vector generation outperform the baselines and provide more meaningful clustering. The applications of this document clustering are huge, especially in the search and retrieval space, providing clinicians, researchers and patients more holistic and comprehensive results than relying on the exclusive term that they search for.

At the end, a framework for extracting clinical information that can be mapped to electronic health records from preventive care guidelines is presented. The extracted information can be integrated with the clinical decision support system of an electronic health record. A visualization tool to better understand and observe patient trajectories is also explored. Both these methods have potential to improve the preventive care services provided to patients.

1. INTRODUCTION

In recent years, active research in the biomedical domain has generated a massive number of documents and articles. Biomedical domain is among the most popular areas of research with a large amount of textual data being generated. At the same time, a lot of previously performed research is also being digitized and made available online. Despite active research, there are relatively few studies which are implemented in medical practice because it is not easy for medical practitioners to go through every form of new literature published.

There is also a large amount of biomedical information stored in electronic health records of patients. The usage of electronic health records throughout the world has only been on the rise [1]. The patient data stored in electronic health records is confidential and governed by laws like HIPAA that restrict access [2]. This makes it imperative that all of the data be de-identified and scrubbed of personal information before it is made available for research purposes.

At the same time, artificial intelligence (AI) research has also been growing in the last two decades. The area of artificial intelligence research that focuses on processing textual data, like the one from biomedical journals, articles and EHRs, is called natural language processing (NLP). Natural language processing focuses on getting computers to understand, analyze and process natural human language. NLP techniques may use small or large amounts of data, based on the context and type of application. While NLP started off as a way to get computers to understand the structure of human language and parts of speech, recent advancements are focused on more complicated problems like automating analysis of large amounts of textual data, providing answers to questions, language models that generate language, and

using vectors as a representation of text. Like other AI models, NLP models may be supervised, unsupervised or semi-supervised. And like other AI models, NLP models are also computationally expensive.

With all this data, there is a continuous need for development of techniques to discover, search, access and share knowledge. Even though a lot of the data is readily available, the amount of data that is utilized for improving clinical outcomes remains relatively low. The availability of biomedical data, a growing interest in AI-related applications and reducing price of computation, there has brought about a spurt in biomedical AI research. Research that is at the intersection of healthcare and artificial intelligence is a lucrative opportunity for researchers because it helps in solving complex problems that can improve the quality of life of patients. Research also focuses on developing systems that can aid doctors, and medical professionals into performing their jobs better. This process has brought out some fundamental challenges in the biomedical space which make it necessary to modify existing general-purpose solutions to fit the biomedical domain. Some of the challenges faced by researchers being:

1. Unavailability of large amounts of labelled data
2. Multiple semantically-equivalent representations of the same concepts
3. Stringent regulations on storing and using patient data
4. Difficulty in evaluating, comparing and reproducing solutions across datasets

In this thesis, I present my research in trying to solve problems 1 and 2 noted above. My work focuses on using unsupervised methods to generate mathematical representations of biomedical concepts.

A discussion about word embeddings, the current state-of-art in word embeddings, how they are generated and used, is presented in Chapter 2. Chapter 3 provides details about the implementation of the different clustering algorithms used, and the

evaluation and visualization metrics. A brief introduction of the external tools used in this tool for pre-processing the text input is presented in Chapter 4. Chapter 5 contains information about the datasets used, and the pre-processing performed.

A discussion about the representation of biomedical concepts, and how they were used in this work is presented in Chapter 6. Clustering results of disease and symptom concepts is also presented in Chapter 6. Chapter 7 discusses the generation of document vectors for biomedical documents, a proposed document weighting scheme, document clustering and the results of these experiments. Experiments performed towards improving patient care are presented in Chapter 8. Chapter 9 summarizes the major contributions of this work, and Chapter 10 provides a direction for future work in the area.

2. WORD EMBEDDINGS

Word embeddings are a mathematical representation of every word in the text. This mathematical representation may be a binary representation making an affirmation, an integer representation where each word is represented by an integer, or a complex vector that represents various properties of the word. While the concept of word embeddings has existed for a long time, modern word embeddings have roots in term frequency [3] (Section 2.2), singular value decomposition [4], latent semantic analysis [5] and latent dirichlet allocation [6].

Recent advances in natural language processing have elaborated this concept of word embeddings by using word vectors where each word is represented by a high-dimensional word vector. These word vectors are based on co-occurrences of words and phrases of the corpus used to generate these word vectors. These co-occurrences are converted into a real number vector representation by using a probability model. These word vectors are generated by using fairly straight-forward neural networks with multiple layers, in an unsupervised way, by using large amounts of textual data.

Artificial intelligence applications rarely use non-numeric inputs, and thus converting characters and strings from their raw form to a numbers becomes necessary. Because of this limitation, it becomes necessary to convert the text input to a numeric form. The strengths of word embeddings have made them a very lucrative initial step in all types of machine learning pipelines. In a lot of complex deep neural networks, word embeddings are used as inputs instead of raw text. This necessitates the conversion of word vectors to phrase, sentence and even paragraph vectors.

Word embeddings have become popular because they have been observed to capture not just the probability distributions of word appearances, but also the semantics and context of words within the original corpus. This was discovered by Mikolov, et al. in their research describing how word vectors are also representative of linguistic

properties of the corpus [7]. Their work noted that word embeddings carry forward the relationships of the real world into the continuous vector space. They present examples like Equation 2.1 where addition and subtraction operations on word vectors show how closely related the word vectors are:

$$king - man + woman \approx queen \tag{2.1}$$

Similar experiments were performed on different verb tense pairs and country-capital pairs. Dimensionality reduction is applied on these word vectors, and their plot in 3D space (male-female and verb-tense examples) or 2D space (country-capital examples) are shown in Figure 2.1 [8]. All of these show the effectiveness of word embeddings in understanding the structure of natural language and learning from it.

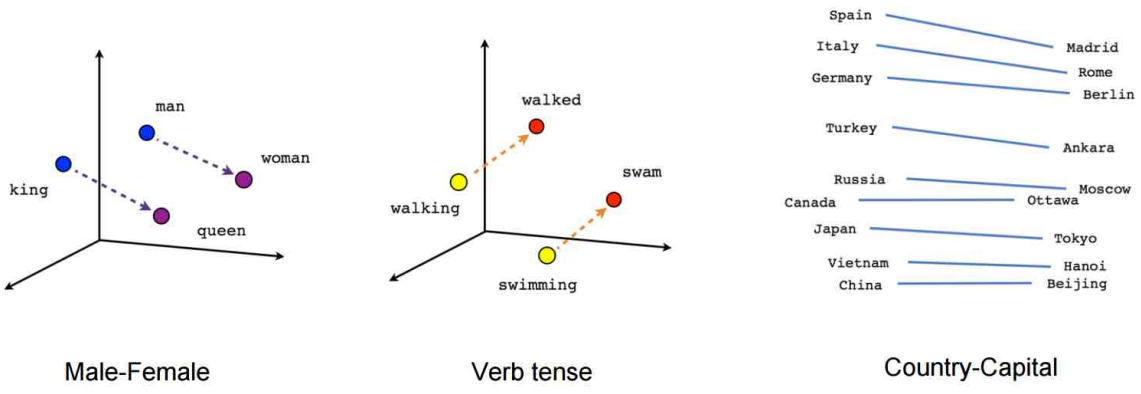

Fig. 2.1.: Relationships between word vectors [8].

Alongside these results, the simple technique to train and generate word vectors is also a significant reason behind their success. In general, word embeddings are generated by training a simple neural network consisting of one input layer, one hidden layer and one output layer to perform a certain task. The task that the neural network is trained on is usually completely unrelated to the final task to perform. However, the weights, specifically the input-hidden layer weights, learned by the network as a part of the training process are used as 'word vectors'.

The forward and backward propagation of the neural network training process helps train the network and thus generate the values for the weights. Usually, the training process for word embeddings does not continue for a large number of epochs, but the input corpus size is large. Word vectors also improve if the word appears repeatedly within the corpus, and the words that are infrequent are pruned from the dictionary. Similarly, extremely frequent words (articles, prepositions, etc.) are also detrimental to the generation of word vectors, and are usually ignored.

The following sections describe the most popular methods for generating word embeddings.

2.1 Bag of Words

Table 2.1.: Example of Bag-of-Words dictionary.

Word	ID	Count
John	1	1
likes	2	2
to	3	1
travel	4	1
Jane	5	1
is	6	1
fond	7	1
of	8	1
traveling	9	1
She	10	1
also	11	1
music	12	1
and	13	1
art	14	1

The earliest and simplest forms of representing text in the form of embeddings is by using a bag-of-words model that uses a numeric count for the number of times each word appears in the query string [9]. The first step towards creating a BoW embedding is to convert the given corpus into a dictionary such that each word corresponds to its frequency in the document. A BoW dictionary is generally created from a corpus, and applied at the sentence or document-level to generate vector representations. It is important to note here that the BoW representation does not change with the sequence of words, but only relies on the absolute frequency of the terms.

For example, let the below be an example corpus.

John likes to travel.
Jane is fond of traveling.
She also likes music and art.

The BoW dictionary would be as given in Table 2.1, and an enconding of the sentence 'John also likes art.' is,

$$[1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1]$$

where every index represents the ID of every word in the dictionary (Table 2.1), and the integer value represents the frequency of the word in the query string.

Bag-of-Words are often also generated for n-grams – a combination of n words that appear sequentially – to explore the most frequenct words and phrases from the corpora.

2.2 Term Frequency

Term frequency (TF) is a simple form of creating word embeddings for a given document [3]. The number of occurrences of a term in the document is called the term frequency. In a word embedding model that uses term frequency, the number of times the word appears in the document is used as the weight of the term.

Documents in a corpus are coded by the number of times each word appears in them, in a mapping between each word and its frequency, without focusing on the semantics or context of the word. Term frequency is very similar to the bag-of-words approach discussed in Section 2.1, with the only difference being the sequence of appearance of the words in the document is maintained in the output vector, and the words that do not appear are skipped.

The term frequency for a term t in document d is given as,

$$tf_{t,d} = f_{t,d} \tag{2.2}$$

where $f_{t,d}$ is the frequency of the term t in document d. Thus, term frequency is different for different pairs of terms and documents in the same corpus.

A TF representation of the query string 'John also likes art.' with the same dictionary as shown in Table 2.1 is,

$$[(1,1),(11,1),(2,1),(14,1)]$$

2.3 Term Frequency-Inverse Document Frequency

Term frequency gives frequent words like 'the', 'a', 'an', etc. higher weights, thus biasing the weighing scheme without actually improving the results. A solution to this was proposed by Sparck Jones by proposing an inverse relationship between the term frequency and the number of documents a term appears in [10].

Inverse document frequency for a term t in a corpus of N documents is given as Equation 2.3.

$$idf_{t,D} = \log \frac{|D|}{|\{d \in D; t \in d\}|} \tag{2.3}$$

where D indicates the number of documents containing the term t, and $|\{d \in D; t \in d\}|$ is the number of documents in the corpus for which $tf_{t,d} \neq 0$ (Equation 2.2). It is important to note that the $idf_{t,D}$ value is the same for all the D documents in the corpus that contain the term t and does not change depending on the term frequency.

If a term that does not exist in the corpus is queried, the denominator becomes 0 which causes a division-by-zero error. To avoid the possibility of the denominator becoming 0, 1 is added to the denominator, and the inverse document frequency equation becomes,

$$idf_{t,D} = \log \frac{|D|}{1 + |\{d \in D; t \in d\}|} \tag{2.4}$$

The equation for term frequency-inverse document frequency (TF-IDF) is a multiplication of Equation 2.2 and Equation 2.4. The formula used to calculate the TF-IDF weight of a term t in a document d in a corpus of size $|D|$, where t is present in D documents is given as,

$$tfidf_{t,d,D} = f_{t,d} \cdot \log \frac{N}{1 + |\{d \in D; t \in d\}|} \tag{2.5}$$

This combination of TF-IDF is widely used to weigh terms in various corpora.

2.4 Word2Vec

Mikolov et al. presented their work of generating and calculating word vectors and called their approach 'Word2Vec' [7]. In their work, they presented results of using a recurrent neural network language model to generate word embeddings. The simplest form of the network consisted of 3 layers, an input layer, a hidden layer and an output layer. The input also consisted of a hidden layer that is carried forward from the previous execution of the neural network. A framework of the recurrent neural network is shown in Figure 2.2.

In Figure 2.2, $w(t)$ is the input vector of vocabulary size N with a one-hot encoding, i.e. only one value is set to 1, others are set to 0. The hidden layer $s(t)$ is a vector of dimension D and the output layer $y(t)$ is a vector of dimension N. $s(t-1)$ represents the hidden layer from the previous iteration. U is a matrix of dimension $N \times D$ with weights for each input word in the vocabulary to the hidden layer. On the other hand, V is a matrix of dimension $D \times N$ with weights for the connection between the hidden layer and the output layer. The output is a log-likelihood of each

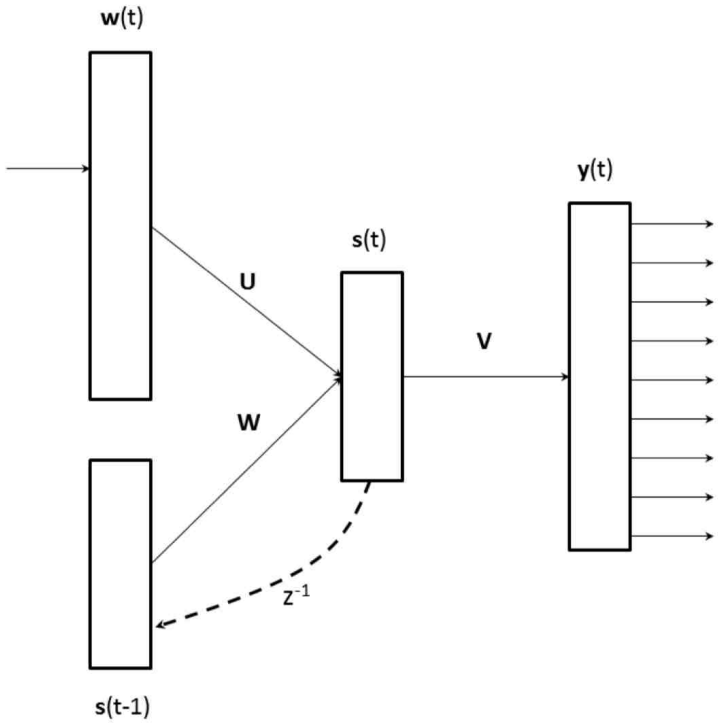

Fig. 2.2.: Framework of a recurrent neural network language model [7].

possible word in the vocabulary to be the next possible word. The RNN is trained with back-propagation to maximize this log-likelihood of the output layer for the correct next word in the sentence [7].

The word vectors from this model were derived from the U matrix – the input-hidde layer weights – and used for evaluation. The results showed that a continuous vector space representation dervied from a language model captures the linguistic regularities well. The work presented as a part of [7], paved the way for further exploration of word embeddings and Word2Vec was presented by modifying the training procedures [11]. The training architectures presented were: 1) Continuous Bag-of-Words and, 2) Skip-gram [11].

2.4.1 Continuous Bag-of-Words

Continuous Bag-of-Words (CBOW) is a sliding window approach that uses $2k$ one-hot encoded input context vectors of the vocabulary length (k before the query word, k after the query word) for a window size of k.

These are used as an input to the hidden layer, and the output is a vector of vocabulary length. The output vector for each word, given the context window k is calculated using Equation 2.6.

$$\frac{1}{T}\sum_{t=1}^{k} logp(w_t|w_{t-k},\ldots,w_{t+k}) \tag{2.6}$$

where T is the vocabulary size, k is the window size, and w_t is the context word.

$softmax$ is applied to the output vector to generate the output vector. The expected output vector is a one-hot encoded output context vector.

The CBOW framework is shown in Figure 2.3.

2.4.2 Skip-gram

Skip-gram is a reverse continuous bag-of-words approach where the input to the network is a one-hot encoded input context vector for the query word and the output is a vector of the length of the vocabulary, showing the possible probabilities of the j context words surrounding it. In the case of skip-gram, the output values are probabilities.

The Equation 2.7 is used to calculate the output probabilities for each word in the vocabulary.

$$argmax_\theta \frac{1}{T}\sum_{t=1}^{T}\sum_{j\in c, j\neq 0} logp(w_{t+j}|w_t;\theta) \tag{2.7}$$

where θ is a chosen parameter, and w_t is the context word.

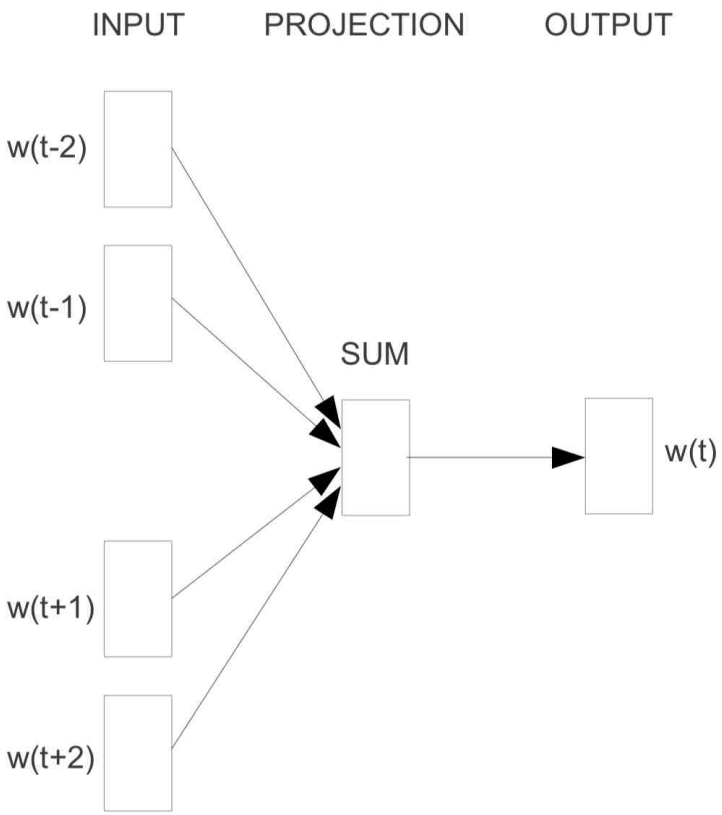

Fig. 2.3.: Framework of the Word2Vec Continuous Bag-of-Words model [11].

The *argmax* function is used to calculate the maximum value of the log-likelihood for each of the words in the vocabulary. These probabilities are compared with the actual context words in the sentence that surround the query word, and are used for back-propagation. The skip-gram framework used in the Word2Vec algorithm is shown in Figure 2.4.

Mikolov et al. trained a word embeddings model on the Google News dataset of ∼100 billion words and phrases, but the vocabulary was reduced to ∼300 million words and phrases based on frequency of words. The model was trained using CBOW and skip-gram and compared [11]. The skip-gram model was later made available on their website [12] [13].

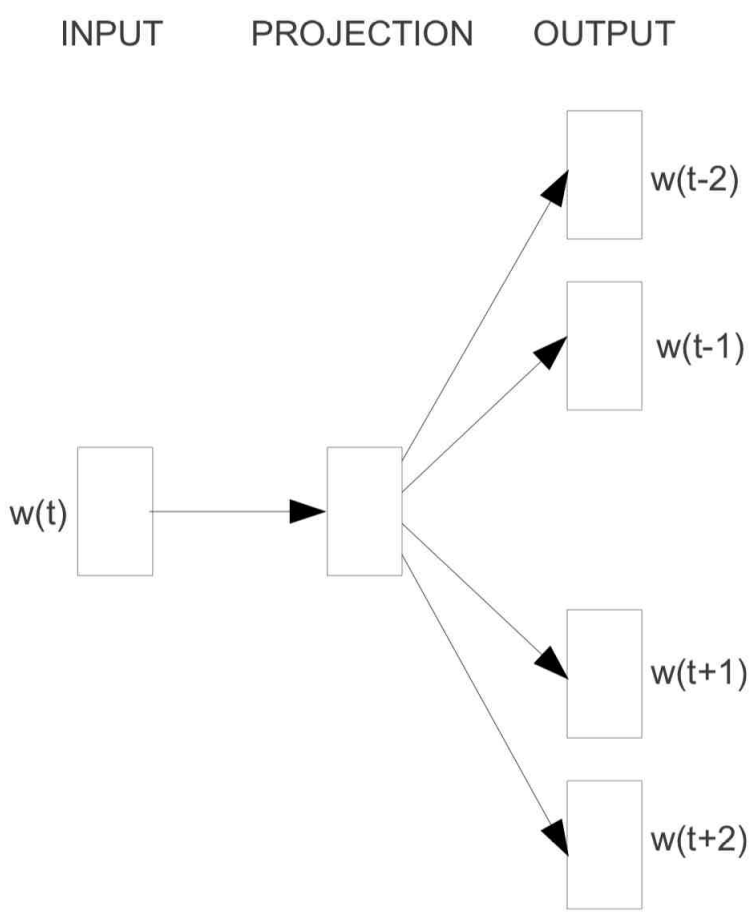

Fig. 2.4.: Framework of the Word2Vec skip-gram model [11].

Since the introduction of Word2Vec, word embeddings quickly rose in popularity in the natural language processing domain. Word embeddings were tested and used in various fields including, but not limited to part-of-speech tagging [14], question-answering [15], classification [16], entity recognition [17], word analogy tasks [7], neural machine translation [18], and especially in language models [19].

There has also been a lot of research into using word embeddings in various fields that utilize NLP in some form, like cyber-security [20], biomedical [21], sentiment classification [22] and parsing [23].

Word embeddings have also been studied to see how many properties of linguistics they carry forward. There has been a lot of research in trying to identify the properties of language that are also seen in the word embeddings. Things like gender and racial bias have been heavily studied and it has been observed that word embeddings learn those exact biases that are present in the original corpora that they are trained from.

At the same time, this also brought about a new beginning towards creating 'better' word embeddings that learn faster, learn from smaller amounts of data, run faster, can be monitored better, and more holistically represent the context.

2.5 GloVe

Global Vectors for word representations (GloVe [24]) was proposed by Pennington et al. and attempts to explain the regularities in the vector space for word vectors [25]. They further describe the model properties required to generate these regularities in the word vectors.

The authors of [25] go on to explain that there is little difference between count-based and prediction-based architectures, since fundamentally, they both probe co-occurrences. The resulting log-bilinear regression model proposed uses a word-word co-occurrence matrix instead of a individual context windows. This model learns word representations in an unsupervised way and outperforms Word2Vec's skip-gram model in tasks like word analogy, word similarity and named entity recognition.

2.6 FastText

Bojanowski et al. identified a limitation of Word2Vec (Section 2.4) and GloVe (Section 2.5), whereby the morphological structure of the words and sentences is ignored to create word embeddings. Instead, word embeddings are created off of the sequence of words, with little to no pre-processing performed. To tackle this, the authors proposed a new method of creating word vectors that relied on representing words as a bag-of-character n-grams [26] [27]. The word vector is then aggregated

for these character n-grams. Another advantage of this type of a training procedure is that such word embeddings are capable of producing word vectors for words that were not seen during the training.

2.7 Embeddings from Language Models

One of the major limitations of word embeddings was that they are not context dependent. For a neural network learning word embeddings, the word 'tears' that appears in the sentences 'He shed tears' and 'She tears the paper' is identical, even though they are very different in connotation. While it is easy for a human to identify the occurrence in the first sentences as the noun implying the action of crying, and the second as a verb, it is a particularly tricky challenge to teach neural networks this without labelled data.

Such examples introduced the need for word embeddings that learn from context. A solution to this was proposed by Peters et al. using deep neural networks that rely on language models. Because this method relied on learning word embeddings from a language model with modifications, they call their approach ELMo [28] [29]. The proposed model relies on syntax, semantics and language polysemy. The advantage of such a model is that it generates a different vector for each word based on the context and the sentence it appears in. Because of this, the same word in two different sentences does not have the same word vector, which was the case in all the previously word embeddings. However, this comes at a cost: computation. The model relies on bi-directional long short-term memory neural networks and thus induces additional overhead in the creation of word vectors.

3. CLUSTERING, EVALUATION AND VISUALIZATION

Clustering is an important step of the proposed work. Clustering algorithms like self-organizing maps (Section 3.1.1) and k-means clustering (Section 3.1.2) enable grouping of concepts, documents and patient records.

Evaluation of the clustering performed is performed to understand the quality of the clustering performed. Internal evaluation metrics like Davies-Bouldin index (Section 3.2.1), and external evaluation metrics like F-measure (Section 3.2.3) and purity (Section 3.2.2) are used.

Visualizing the results of the generated clusters allows for the evaluation of the clustering performance in an observable way. Visualization is performed by using visualization tools like U-matrix and hit histogram (Section 3.3.1) or by using dimensionality reduction algorithms like Principal Component Analysis (Section 3.3.2) and T-distributed Stochastic Neighbor Embedding (Section 3.3.3), and then using scatter plots on 2-dimensional data.

3.1 Clustering

Clustering is the process of grouping sets of objects together based on the similarity or differences between objects. Objects that appear in the same group have a degree of similarity, whereas those in different groups have a higher degree of difference. Each group of data objects is called a 'cluster'.

Clustering algorithms, in general, are unsupervised and rely on mathematical representations of data. Clustering is performed by repeated iterations that group and re-group data until stability or maximum number of pre-defined iterations are reached. Clustering can be generalized as a multi-class classification problem. Descriptions of the clustering algorithms used in this work are presented in the following subsections.

3.1.1 Self Organizing Map

Self-Organizing Map (SOM) is a type of neural network that is used for document clustering and visualization [30]. SOM implements a topologically ordered display of the data to facilitate understanding structures in the input data set. It is also readily explainable and easy to visualize. Visualization of multi-dimensional data is one of the main applications of SOM [31]. These features make SOM an appropriate choice as a clustering algorithm for this work.

A basic SOM consists of M neurons located on a low dimensional grid (typically 2 dimensional) [31]. The algorithm responsible for the formation of the SOM involves three basic steps after initialization - sampling, similarity matching, and updating. These three steps are repeated until formation of the feature map is complete. Each neuron i has a d-dimensional prototype weight vector $W_i = W_{i1}, W_{i1}, ..., W_{id}$. Given X is a d-dimensional input vector, the algorithm can be summarized as follows:

1 **Initialization:** Choose random values to initialize all the neuron weight vectors $W_i(0)\ \forall\ i = 1, 2, \ldots, M$ where M is the total number of neurons in the map.

2 **Sampling:** Draw a sample data X from the input space with a uniform probability.

3 **Similarity Matching:** Find the best matching unit (BMU) or winner neuron of X, denoted here by b which is the closest neuron (map unit) to X in the criterion of minimum Euclidean distance, at time step n (n^{th} training iteration).

$$b = \arg\min_i ||X - W_i(n)|| \forall i = 1, 2, \ldots, M \tag{3.1}$$

4 **Updating:** Adjust the weight vectors of all neurons by using the Equation 3.2, so that the best matching unit and its topological neighbors are moved closer to the input vector X in the input space.

$$W_i(n+1) = W_i(n) + \eta(n) \cdot h_{b,i}(n) \cdot (X - W_i(n)) \tag{3.2}$$

where $\eta(n)$ denotes the learning rate and $h_{b,i}(n)$ is the suitable neighborhood kernel function centered on the winner neuron. The distance kernel function can be, for example, Gaussian:

$$h_{b,i}(n) = e^{-\frac{||r_b - r_i||^2}{2\sigma^2(n)}} \tag{3.3}$$

where r_b and r_i denote the positions of neuron b and i on the SOM grid and $\sigma(n)$ is the width of the kernel or neighborhood radius at step n. $\sigma(n)$ decreases monotonically along the steps as well. The initial value of neighborhood radius $\sigma(0)$ should be fairly wide to avoid the ordering direction of neurons to change discontinuously. $\sigma(0)$ can be properly set to be equal to or greater than half the diameter of the map. Equation 3.4 gives the initial value of the neighborhood radius for a map of size a by b.

$$\sigma(0) = \frac{\sqrt{a^2 + b^2}}{2} \tag{3.4}$$

5 **Continuation:** Continue with steps 2-4 until no noticeable changes in the feature map are observed or a pre-defined maximum number of iterations is reached.

The results of an SOM can be directly projected onto a two-dimensional space for visualization. For this reason, SOM is also a dimensionality reduction algorithm. The most common visualization techniques used for visualizing and evaluating SOM results are the U-matrix and Hit histogram, which are further detailed in Section 3.3.1.

3.1.2 k-means

k-means clustering algorithm [32] is straight-forward to implement and can be applied to large and high dimensional data sets. It has been successfully used in various application domains, such as text mining, computer vision and so on [33] [34]. k-means clustering algorithm tries to assign the data in the data set to one of the

predefined number of clusters. The aim is to minimize the sum of distances of each point within the cluster to the cluster center. Given $x = x_1, x_2, \ldots, x_d$ is a set of d-dimensional input vector, and $C = C_1, C_2, \ldots, C_k$ is a set of randomly initialized k centers with d-dimensions, the algorithm is summarized as follows:

1 **Assignment of cluster centers:** Assign each data point x_i to the cluster C_j whose Euclidean distance from the cluster center is minimum of all the cluster centers.

$$C_j = \{x_i : ||x_i - C_j|| \leq ||x_i - C_i||, \forall i, 1 \leq i \leq k\} \tag{3.5}$$

2 **Update cluster centers:** Set the new center of each cluster to the mean of all data points belonging to that cluster.

$$\mu_i = \frac{1}{|C_i|} \sum_{j \in C_i} x_j, \forall i \tag{3.6}$$

3 **Repeat:** The steps 1-2 are repeated until convergence or until a pre-defined maximum number iterations is reached.

k-means algorithm does not have the property of projecting the input vectors to a low dimensional space for cluster visualization. The techniques used to reduce to visualize the clustering output to a low dimensional space is described in Section 3.3.2 and Section 3.3.3.

3.2 Evaluation

Different evaluation metrics rely on different parts of the data, with some metrics depending more heavily on intra-cluster distances, and others consider inter-cluster distances more significantly.

Evaluation metrics that rely on properties of the data that are used for clustering are called internal evaluation metrics. These metrics do not require external labels and rely on the data itself. Often, these metrics also rely on the centroid of the clusters. If the clustering algorithm does not specify the centroids, it can be quickly calculated by taking a mean of the cluster members' values.

External evaluation metrics rely on more than the data that is available to the clustering. These metrics require the 'ground truth' labels describe the ideal result of the clustering.

It should be noted that all of the evaluation results are calculated for the clustering results, before any visualization or dimensionality reduction algorithms are applied.

3.2.1 Davies–Bouldin Index

Davies–Bouldin Index (DB index) [35] is built from the idea that cluster members should have high similarity, whereas those in different clusters should have low similarity. DB index is calculated as,

$$DB - Index = \frac{1}{C} \sum_{i=1}^{C} \max_{j,j \neq i} \frac{SD_i + SD_j}{||CL_i - CL_j||} \tag{3.7}$$

where C is the total clusters, SD_i is the standard deviation of the distance of samples in a cluster to the respective cluster centroid, and $||CL_i - CL_j||$ is the Euclidean distance between centroids CL_i and CL_j.

The more distinct the clusters are from each other, smaller the Davies–Bouldin index value is. Thus, the lower the value of DB index, the better the clustering. An issue of using DB index as a validation metric is that it does not necessarily imply the best information retrieval. Another drawback being the equal weight given to each cluster, which skews the result towards larger clusters in dataset with unequal classes.

3.2.2 Purity

Purity is a basic external evaluation metric that verifies the association of a clustering result with the label [36]. For each cluster, the most frequent class is assumed to be the class of the cluster. All the members of the cluster that belong to this cluster are correct, while those that belong to any other class are incorrect.

Purity is a sum over all clusters of the count of the maximum membership class. For N datapoints clustered into C clusters, with D possible classes, purity can be expressed as,

$$Purity = \frac{1}{N} \sum_{c \in C} \max_{d \in D} |c \cap d| \tag{3.8}$$

However, purity does not provide a complete picture of the accuracy of the system, and is thus usually complimented with F-measure (Section 3.2.3), which builds on the concept and provides a better evaluation metric for labelled data.

3.2.3 F-measure

F-measure (also called F-score, F_1-score or F1-measure) is an external clustering evaluation metric that relies on the true positives, true negatives, false positives and false negatives of a binary classification to determine the accuracy of the classification [37]. since clustering is an n-class classification problem, F-measure can be scaled to suit more classes and accurately measure the results, provided the labels are available.

F-measure is formed with its 2 components, precision and recall. Precision is calculated as the ratio of true positives to the total membership for each cluster (Equation 3.9). Recall, on the other hand, is the ratio of true positives for each cluster to all the datapoints that should be in that cluster (Equation 3.10).

$$precision = \frac{tp}{tp + fp} \tag{3.9}$$

$$recall = \frac{tp}{tp + fn} \tag{3.10}$$

In Equations 3.9 and 3.10, tp stands for the number of true positives, tn for the number of true negatives, fp for the number of false positives, and fn for the number of false negatives.

In terms of clustering, precision and recall are calculated separately for each cluster (or class), and then averaged together. Precision is similar to purity (Section 3.2.2), but with one major difference: precision is calculated separately for each cluster, but purity is calculated for the whole result.

F-measure is calculated using precision and recall as,

$$f = \left(\frac{precision^{-1} + recall^{-1}}{2} \right)^{-1} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.11}$$

Equation 3.11 can be simplified in terms of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) as,

$$f = 2 \cdot \frac{tp}{2 \cdot tp + fp + fn} \tag{3.12}$$

3.3 Visualization

Clustering algorithms provide a 'grouping' of the data in the form of clusters. However, they do not provide a visualization that can be used to decipher the clustering results.

Clustering algorithms like self-organizing maps (Section 3.1.1) reduce the dimensions on the clustered data to (typically) two dimensions, allowing for them to be easily visualized. At the same time, data that is stored as a part of the SOM training in the form of weights contains more information about the sample space and data set. All of these can be plot and interpreted in the form of U-matrix and Hit histogram (Section 3.3.1) to visually evaluate the results of the SOM clustering.

On the other hand, clustering algorithms like k-means (Section 3.1.2) provide a classification of the data set on the same dimension as the input set. These clustering results are applied dimensionality reduction algorithms like Principal Component Analysis (PCA) (Section 3.3.2) and T-distributed Stochastic Neighbor Embedding (t-SNE) (Section 3.3.3) to reduce the dimension of the data set and the cluster centroids to 2D. The dimension reduced data and centroids can then be plot onto a 2D or 3D space.

3.3.1 U-Matrix and Hit Histogram

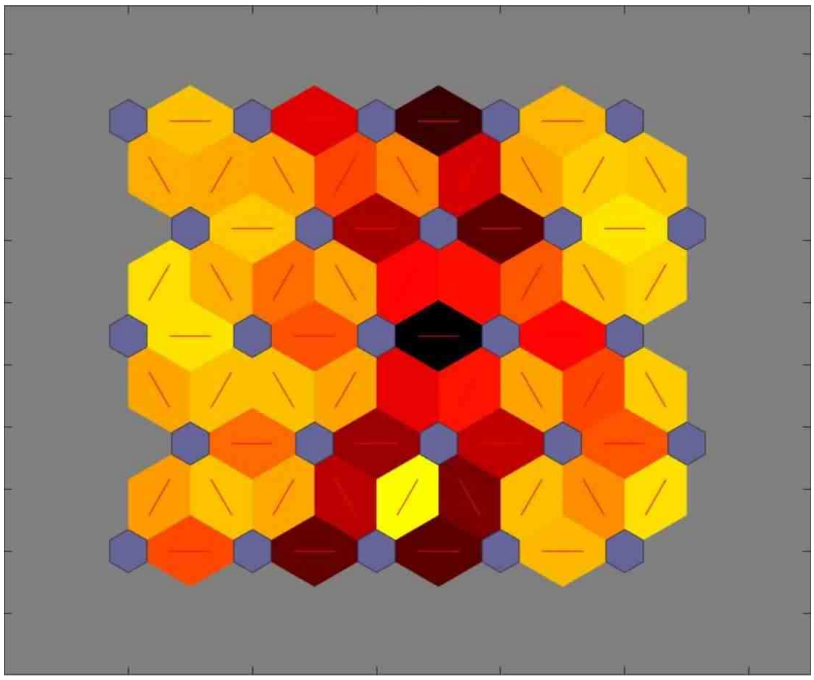

Fig. 3.1.: Sample U-matrix of a trained SOM.

The most commonly used visualization techniques of SOM are the U-Matrix and Hit histogram. The U-matrix holds all distances between neurons and the immediate neighbor neurons [31]. Figure 3.1 shows a sample U-matrix of a trained self-organizing map on an input data set that has two clusters. The lighter the color in the hexagon connecting any two neurons, the smaller is the distance between them. From the U-matrix, two large light regions can be visualized. One is towards the left, while the other is to the right. These regions present the two clusters obtained on training the input data set. The U-matrix gives a direct visualization of the number of clusters and their distribution.

The hit histogram of the input data set on the trained map provides a visualization that details the distribution of input data across the clusters. Each input data instance in the data set can be projected to the closest neuron on a trained SOM map. The closest neuron is called the best matching unit (BMU) of the input data instance.

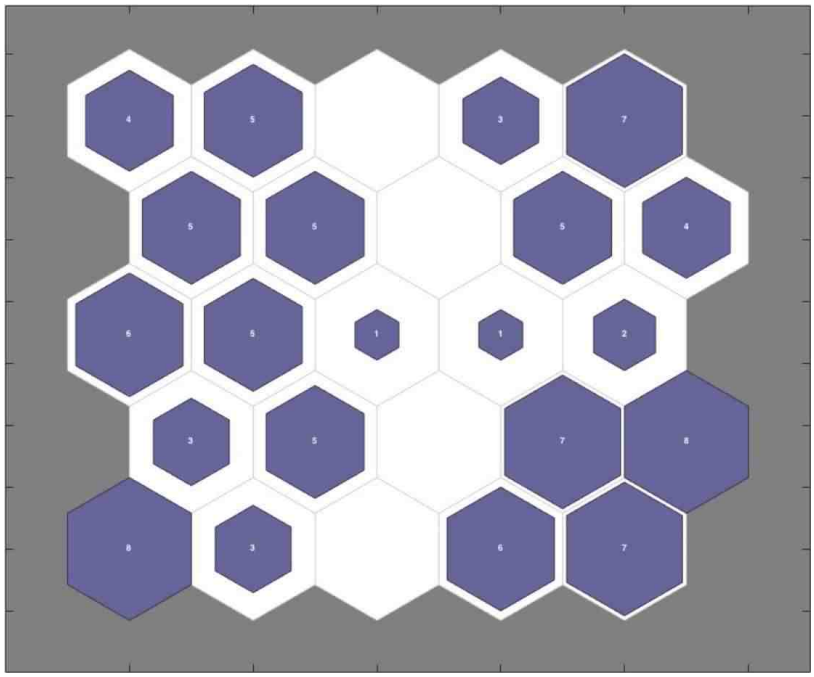

Fig. 3.2.: Sample Hit histogram of a trained SOM.

The hit histogram is constructed by counting the number of hits each neuron receives from the input data set. Figure 3.2 shows the hit histogram of a sample input data set on a trained SOM. Each hexagon represents one neuron on the map. The size of the marker indicates the number of hits the neuron receives. Thus, a larger marker is representative of a larger number of hits on that neuron. Based on the hit histogram, it is visualized that most of the input data hits neurons in the left and right regions. These two regions correspond to the two clusters on the U-matrix shown in Figure 3.1.

3.3.2 Principal Component Analysis

Principal Component Analysis (PCA) uses an orthogonal linear transformation to convert a set of correlated observations into uncorrelated principal components [38]. Principal components are a set of variables that are linearly uncorrelated.

Applying principal component analysis to a dataset brings out the variations of datapoints and draws an emphasis on the patterns in the dataset. Thus, principal component analysis is very frequently used as a tool for dimensionality reduction and visualization of large data sets.

The first co-ordinate in the new system contains the variable from the transformation with the greatest variance, the second co-ordinate contains the variable with the second greatest variance, and the last co-ordinate in the new system contains the variable with the least variance across the transformed data set.

When using PCA for dimensionality reduction to reduce the data set to d dimensions, only the top d-dimensions from the transformed data set are chosen, and the others are dropped. This means the top-d variables with the highest variances are chosen, while the other variables with lower variances are truncated.

Dimensionality reduction with PCA is tricky when considering high dimensions as it may lead to a loss of large amounts of valuable data in the truncated variables. However, PCA serves as a reliable step in the dimensionality reduction process by aiding the removal of variables that contain low variances. Thus, PCA suppresses some noise without severely distorting the distances between data points.

In this work, PCA is used as an intermediary step while reducing the dimensionality of the high-dimensional k-means output data. The output of PCA is used as an input to the t-SNE algorithm, which further reduces the dimensions and provides a data set reduced to 2-dimensions, which can be visualized using scatter plots.

3.3.3 T-Distributed Stochastic Neighbor Embedding

T-Distributed Stochastic Neighbor Embedding (t-SNE) was proposed by Maaten et al. [39] and is employed to visualize the clustering results. The t-SNE algorithm is used for dimensionality reduction. The algorithm minimizes the sum of the Kullback-Leibler divergences of all data points in the original dimensional space and the mapping space.

Given a set of N inputs $x_1, \ldots, x_N$, the goal of t-SNE is to learn a d-dimensional map. The algorithm can be summarized as [39]:

1 **Pairwise affinities:** The affinity of datapoint x_i and datapoint x_j is based on the probability of x_i picking x_j as its neighbor. This probability depends on the variance of the Gaussian distribution centered at x_i, σ_i. The pairwise affinities can be mathematically represented as:

$$p_{j|i} = \frac{\exp\left(\frac{-||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-||x_i - x_k||^2}{2\sigma_i^2}\right)} \quad \forall i, \forall j \tag{3.13}$$

2 **Pairwise similarities:** The pairwise similarities for each (i, j) pair are calculated as:

$$p_{ij} = \begin{cases} \frac{p_{j|i} + p_{i|j}}{2N} & i \neq j \\ 0 & i = j \end{cases} \tag{3.14}$$

3 **Initial Solution:** An initial solution is sampled as $Y^{(0)} = y_1, \ldots, y_N$ such that $y_i \in \mathbb{R}^d$.

4 **Compute Low-Dimensional Affinities:** The affinities of the output variables are calculated by the Equation 3.15:

$$q_{ij} = \begin{cases} \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}} & i \neq j \\ 0 & i = j \end{cases} \tag{3.15}$$

5 **Gradient:** The gradient of the Kullback-Leibler divergence between p_{ij} and q_{ij} is calculated as:

$$\frac{\delta C}{\delta y_i} = 4 \cdot \sum_j (p_{ij} - q_{ij}) \cdot (y_i - y_j) \cdot (1 + ||y_i - y_j||^2)^{-1} \tag{3.16}$$

6 **Update Solution:** The solution is updated as:

$$Y^{(t)} = Y^{(t-1)} + \eta \cdot \frac{\delta C}{\delta y_i} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)}) \tag{3.17}$$

7 Repeat: Steps 4-6 are repeated until convergence or until a maximum number of iterations are reached.

Thus, the locations of the input datapoints X on the d-dimensional map are determined by minimizing the Kullback-Leibler divergence (Equation 3.16). This minimization is performed by using gradient descent. The output of the t-SNE algorithm, Y is a map that correlates the similarity of the input datapoints in the low-dimensional space. The output Y can then be visualized using a scatter plot.

It must be noted that the computational cost of t-SNE is high when the original dimensionality of the data is high. To speed up the process, Principle Component Analysis (Section 3.3.2) is used to reduce the dimensionality to to a lower space before t-SNE technique is applied to convert the lowered dimensional representation to a two-dimensional map.

4. INFORMATION EXTRACTION

The data sets used as a part of this work are unlabeled and provide no information between medically relevant and medically irrelevant parts of a sentence. While training word embeddings and using word vectors does not require any labels, it is necessary to figure out which words and phrases are relevant for further downstream tasks (Chapters 7 and 8).

4.1 UMLS MetaMap

The United States National Library of Medicine (NLM) maintains a collection of many popular vocabularies as a part of the Unified Medical Language System (UMLS) [40]. UMLS provides a mapping between various terms and ontologies across different vocabularies. This mapping through UMLS aims to provide a comprehensive and exhaustive ontological representation of all of the biomedical concepts defined in any of the vocabularies to make the vocabularies inter-operable. The UMLS vocabularies are organized under different knowledge sources:

1 **Metathesaurus:** A large biomedical thesaurus that spans across UMLS vocabularies. Vocabularies that are a part of Metathesaurus include SNOMED CT, RxNorm, MeSH, ICD-10, etc. [41].

2 **Semantic Network:** A categorization framework for concepts within UMLS. The Semantic Network also includes the 'semantic types' that biomedical concepts are grouped into, and the relationships between these semantic types [42].

3 **Specialist Lexicon:** A set of many independent NLP tools designed to link common English vocabulary with biomedical data. These tools are available as a set of Java programs [43].

In 2001, Dr. Alan Aronson and the NLM introduced MetaMap, a new tool to provide a mapping from any text to biomedical concepts available in UMLS [44] [45] [46]. MetaMap provides a correlation between input text and its representation in UMLS. Since 2001, MetaMap is actively developed and maintained with bi-annual releases.

MetaMap uses natural language processing and computational linguistic techniques like parsing, phrase matching and word sense disambiguation to evaluate input text and provide matches to biomedical lexicon from the UMLS Metathesaurus.

MetaMap is available to try interactively through the NLM's website or can be used locally with a command line interface or a Java API. MetaMap is extremely configurable and can be configured to suit the requirements for the application.

MetaMap takes text as an input and outputs either a human-readable formatted text or XML or JSON strings. For every word / phrase match from the input text, MetaMap provides a list of output parameters in the output [47]. Some of these output parameters are described in Table 4.1.

'Candidate Matched' and 'Candidate Preferred' output parameters from Table 4.1 are biomedical normalized representations of every phrase matched. The matched concept is the closest representation as it appears in the original text. Whereas, the preferred concept is the normalized representation of the biomedical concept in the Metathesaurus. For example, for an input of 'lung cancer', the candidate matched is 'lung cancer', but the preferred concepts are 'malignant neoplasm of the lung' and 'carcinoma of lung'. The preferred candidate, thus, helps in normalizing the concepts across the dataset.

The 'Semantic Type' output parameter of of MetaMap is among the most valuable to this work. It provides one or more semantic types that the candidate concept is categorized into. Some of the semantic types used are:

1. Disease or Syndrome

2. Neoplastic Process

Table 4.1.: Description of some of MetaMap's output parameters.

MetaMap tag / key	**Description**
Candidate Matched	The Metathesaurus concept candidate the phrase was matched to.
Candidate Preferred	The Metathesaurus preferred concept candidate for the phrase.
Candidate Score	The negative score of the concept match.
CUI	The concept unique identifier ID.
Words Matched	The words from the original text that were matched to this candidate.
Semantic Type	The semantic type of the concept from the categories defined in the Semantic Network.
Sources	The vocabularies from Metathesaurus that the candidate concept is available in.

3. Sign or Symptom
4. Age Group
5. Clinical Attribute
6. Organism Attribute
7. Clinical Drug
8. Pharmacologic Substance
9. Qualitative Concept
10. Quantitative Concept
11. Temporal Concept

Given the following paragraph from United States Preventive Services Task Force's 'Lung Cancer: Screening' guidelines as an input [48], Table 4.2 shows a formatted subset of the results of running MetaMap on the above sentence.

> The USPSTF recommends annual screening for lung cancer with low-dose computed tomography (LDCT) in adults aged [...].

In the scope of this work, MetaMap was used to extract biomedical concepts from all of the different datasets described in Chapter 5. Throughout the work, MetaMap was configured to only show outputs from a reduced set of semantic types. The disease concepts were extracted by using the 'Disease or Syndrome' and 'Neoplastic Process' semantic types, and symtoms concepts are extracted using the 'Sign or Symptom' semantic type. For all the concepts identified, the words matched were replaced by the candidate preferred.

A local instance of MetaMap is run using the command line and output format is set to JSON. The JSON output was piped through a Python wrapper and parsed further [49].

4.2 Stanford CoreNLP

Stanford CoreNLP is a suite of linguistic tools developed at Stanford University's Natural Language Processing group [50]. CoreNLP contains NLP tools to perform part-of-speech tagging, named entity recognition, dependency parsing, open information extraction, tokenization, etc. Except English, the CoreNLP suite is also available in Arabic, Chinese, French, German and Spanish.

CoreNLP was used as a part of this work for pre-processing some of the data in applications in Chapter 8. CoreNLP is available online as a demo [51], but is recommended to be used locally for longer using a local instance of the Java API. A Python wrapper for CoreNLP's Java API developed by the NLP group was also used in this project [52].

Table 4.2.: Results of some of MetaMap's output parameters for an example.

Words Matched	**Candidate Matched**	**Candidate Preferred**	**CUI**	**Semantic Type**	**Candidate Score**
recommends	recommends	Recommendation	C0034866	Idea or Concept	-1000
annual	Annual	Annual	C0332181	Temporal Concept	-593
screening, for, lung, cancer	Screening for Lung Cancer	Screening for malignant neoplasm of lung	C0281477	Diagnostic Procedure	-926
low, dose	Low dose	Low dose	C0445550	Quantitative Concept	-612
computed, tomography	Computed Tomography	X-Ray Computed Tomography	C0040405	Diagnostic Procedure	-778
adults	Adults	Adult	C0001675	Age Group	-581
aged	Aged	age	C0001779	Organism Attribute	-1000
aged	aged	Old age	C1999167	Population Group	-1000
aged	Aged	Elderly (population group)	C0001792	Population Group	-1000

The tools from CoreNLP used in this work, and their descriptions are:

1 **Part-of-Speech Tagging:** The part-of-speech (POS) tagger is one of the initial tools developed for CoreNLP, and is also a crucial component [53]. The POS tagger is a standard log-linear tagger that tags every word in the text with the tags from the Penn Treebank [54].

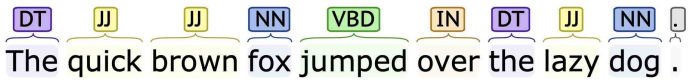

Fig. 4.1.: An example result of CoreNLP's part-of-speech tagger.

2 **Dependency Parser:** The dependency parser in CoreNLP uses a neural network dependency parses [55]. The parser parses every part of a sentence and describes links them between different parts of the sentences based on subject, object, modifiers, roots, together with the part-of-speech tags.

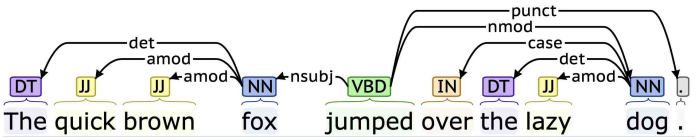

Fig. 4.2.: An example result of CoreNLP's dependency parser.

3 **Named Entity Recognizer:** The named entity recognition identifies names of locations, people, organizations, etc. and miscellaneous entries like money, percentages, numbers, dates and timestamps [56].

Fig. 4.3.: An example result of CoreNLP's named entity recognizer.

4 **Open Information Extraction:** The Open Information Extraction system in CoreNLP identifies and extracts relation tuples and binary relationships from text, without a pre-defined schema [57].

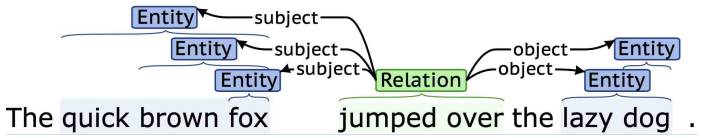

Fig. 4.4.: An example result of CoreNLP's Open Information Extraction system.

5. DATA SOURCES

Large repositories of freely medical articles and journals are available through the National Institute of Health's PubMed service [58] [59]. PubMed articles that include access to full-text are included as a part of PubMed Central [60], whereas the database that contains the citations and abstracts for articles are a part of MEDLINE [61]. PubMed's MEDLINE, the largest biomedical abstracts and citations database, has more than 26 million articles with thousands more added daily. MEDLINE documents are also tagged and indexed with Medical Subject [62] [63].

Different subsets of PubMed have been developed and used for various types of biomedical research over the years. The datasets described in Section 5.1 were generated for varied research purposes, and span more than 30 years. Section 5.1.3 describes the TREC Genomics dataset that was developed as a part of National Institute of Standards and Technology's Text REtrieval Conference series. PubMed Central – Open Access (5.1.1) contains a subset of PubMed Central that is made available openly for any research [60]. Ohsumed text collection (Section 5.1.2) contains a dataset developed in late the 1990s for text categorization tasks.

These biomedical document corpora have been used repeatedly in the literature for previous attempts at text classification and clustering, and are thus ideal datasets for experimenting applications of research presented in this thesis.

Electronic Health Records (EHRs), also called Electronic Medical Records (EMRs), are a method of collecting and storing patient information. Concisely, Electronic Health Records are patient health records that are stored electronically and available on a variety of devices. Most electronic health records are stored online and available through web portals of hospitals.

In recent years, Electronic Health Record systems have recently been adopted by many countries [1]. Electronic Health Records are filled up by doctors, physicians, nurses and therapists, and also used extensively for billing purposes. EHRs contain extremely useful information which can help in improving the quality of patient care.

The information available in EHRs, however, contains identifiable patient information. This is why such information must be dealt with extreme care to protect confidentiality. As a part of this work, patient data from 500 patients over 15 years was used to show a proof-of-concept of how natural language processing techniques can be used towards improving patient care. Section 5.2 contains information about the EHR data used.

5.1 Biomedical Documents

To evaluate the proposed biomedical document clustering framework (Chapter 7), three datasets of biomedical document collections from the NLM's PubMed are used [58]. The three datasets selected are very different from each other with respect to the categories of documents, the size of corpora, the documents within the corpora, etc.

One of these is a labeled dataset while the other two are unlabeled. This distinction between datasets helps in evaluating the performance of the framework across datasets of different sizes. The performance of the framework on the labeled dataset helps establish a baseline for comparison with other methods. Since most of the corpora are unlabeled and large, testing the framework on varying sizes of datasets helps in estimating its performance as the size of datasets scale up.

For all of the datasets, the focus of this work is on document-level clustering based on concepts of diseases. Thus, the first step is extracting the disease concepts by using UMLS MetaMap (Chapter 4, Section 4.1). For the extracted concepts, we also present the distribution of concepts across the corpus, as well as the span of the concepts across the number of words in the respective sections.

In this research, only content in the 'Title' and 'Abstract' sections of the documents are input to proposed framework. Since we only use these short sections of the documents, there are some documents from which no disease-related concepts are identified by MetaMap. Such documents are excluded and not considered for document vector generation (Chapters 6 and 7).

5.1.1 PubMed Central – Open Access

PubMed Central – Open Access is an unlabeled subset of over 1 million articles from the total collection of articles in PubMed Central [60] [64]. The PubMed Central – Open Access data set has been widely used in many research projects to examine tasks of biomedical clustering and classification [34] [65]. PubMed Central is also a part of the training corpus for the word embeddings model used in this research [21].

For this research, 600 articles were randomly selected from the 'A-B' subset which includes articles from journals whose names start with letter 'A' or 'B'. The number of selected articles from each journal is shown in Table 5.1.

After retrieving disease concepts from the 600 documents in the dataset using MetaMap, 658 unique concepts of diseases were identified. Figure 5.1 shows the distribution of these concepts based on the number of words in each concept. It can be seen that around 20% of the concepts identified are of single word length, whereas another 50% of the concepts are two words in length. The concepts of length three words are almost 20%, and the number of concepts greater than or equal to four words in length are around 9%.

Figure 5.2 shows that about 73% of the concepts extracted appear in only 1 document from the PubMed Central – Open Access dataset, and 23.4% appear in 2-5 documents. Just over 3% of concepts appear in 6-14 documents, and only 0.6% of the concepts appear in more than 15 documents. These document frequencies of concepts show that the concept appearance through the corpus is much more sparse

Table 5.1.: Journal-wise distribution of the documents in the PubMed Central – Open Access dataset.

Name of journal	**Document Count**
American Journal of Hypertension	13
Augmentative and Alternative Communication	2
Ancient Science of Life	3
Bioinformatics and Biology Insights	45
Allergy and Asthma Proceedings	28
BoneKEy Reports	4
Anesthesia, Essays and Researches	135
Biological Trace Element Research	31
Bone Marrow Research	1
Brain and Language	1
American Journal of Physiology, Endocrinology and Metabolism	11
Aphasiology	3
Annals of Rehabilitation Medicine	323

than that of the TREC dataset (Section 5.1.3). Such sparsity of concepts occurs usually in large, randomized and diverse corpora because of the uneven distribution of concepts.

5.1.2 Ohsumed Collection

The Ohsumed text collection [66] is a subset of MEDLINE [61] from 1987-1991 [67]. The subset of the Ohsumed collection used here includes the abstracts of approximately 2400 articles. These articles are related to cardiovascular diseases.

Table 5.2.: Category-wise distribution of the abstracts selected from Ohsumed collection subset.

Category	**Number of documents**
Bacterial Infections and Mycoses	100
Virus Diseases	94
Parasitic Diseases	65
Neoplasms	152
Musculoskeletal Diseases	92
Digestive System Diseases	111
Stomatognathic Diseases	100
Respiratory Tract Diseases	115
Otorhinolaryngologic Diseases	125
Nervous System Diseases	103
Eye Diseases	98
Urologic and Male Genital Diseases	106
Female Genital Diseases and Pregnancy Complications	106
Cardiovascular Diseases	108
Hemic and Lymphatic Diseases	104
Neonatal Diseases and Abnormalities	100
Skin and Connective Tissue Diseases	102
Nutritional and Metabolic Diseases	102
Endocrine Diseases	95
Immunologic Diseases	108
Disorders of Environmental Origin	108
Animal Diseases	92
Pathological Conditions, Signs and Symptoms	121
Total	**2407**

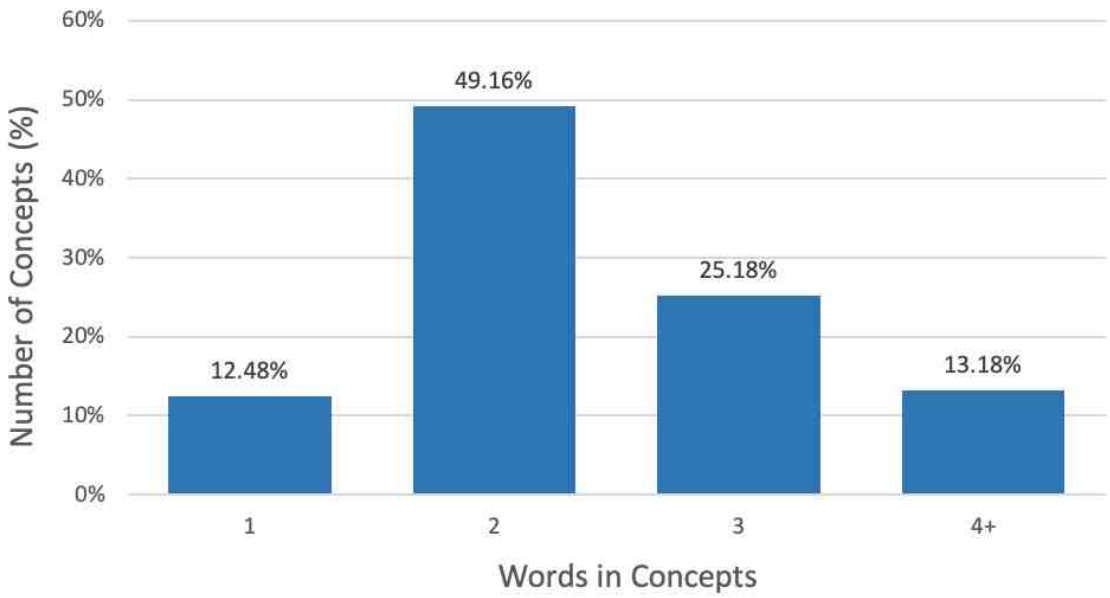

Fig. 5.1.: Distribution of the number of words over every concept identified after processing the PubMed Central – Open Access dataset with MetaMap.

In total, 3649 concepts of diseases are identified and extracted from just over 2400 abstracts. These abstracts are categorized into a category hierarchy which has 23 top level categories and the documents from each category, are given in Table 5.2. However, since all documents are related to cardiovascular diseases, these categories are not labels but categories to which the diseases and documents refer to. The original category labels of the Ohsumed Collection are not assigned based on the concepts of diseases, thus, these labels are used to evaluate the clustering performance.

Figure 5.4 shows the distribution of these concepts based on the number of words in the concepts. From the 3649 concepts extracted, 20% concepts are single word, just over 50% concepts consist of two words. About 20% concepts have 3 words, and about 9% concepts have more than 4 or more words. This shows the distribution of concepts is fairly similar across datasets (Figures 5.5 and 5.1).

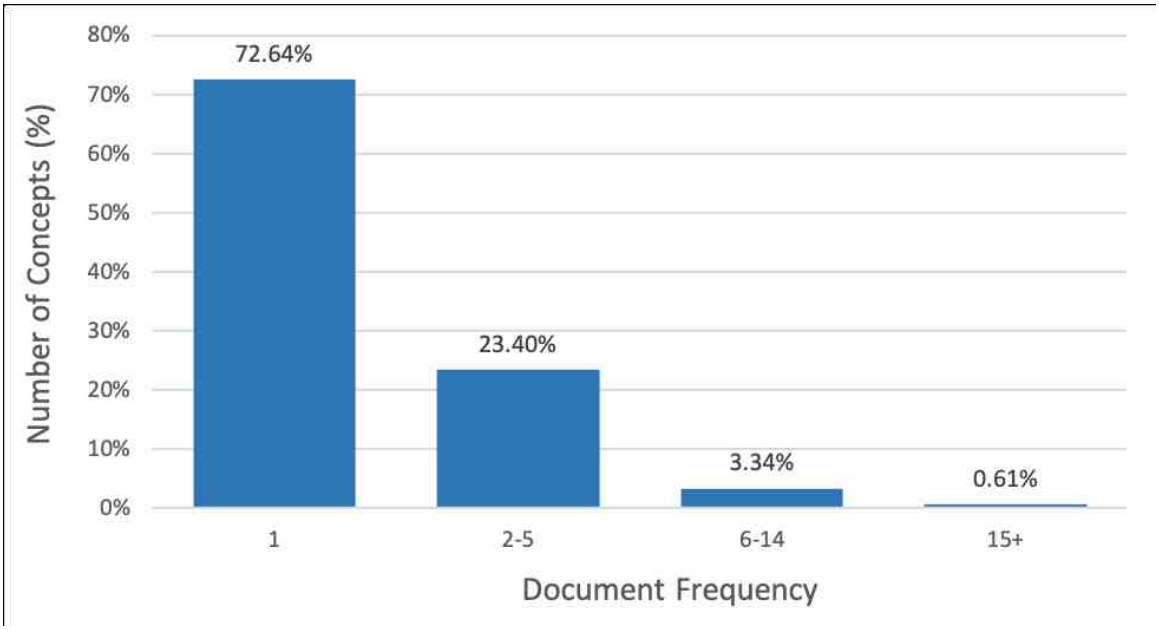

Fig. 5.2.: Distribution of the identified disease concepts over document frequency for the PubMed Central – Open Access dataset.

Similarly, the distribution of concepts across documents is also fairly similar across datasets, with almost 60% concepts appearing in only 1 document as seen in Figure 5.4. 34.2% of the concepts appear in 2-5 documents, and 5.1% concepts are present in 6-14 documents. Only 1.4% of the concepts are present in more than 15 documents. Although the total number of concepts of diseases extracted from the Ohsumed collection is large, the distribution of concepts across documents is very similar to that of PubMed (Figure 5.2).

5.1.3 TREC 2005 Genomics

This corpus consists of a subset of documents from MEDLINE extracted as a part of the TREC 2005 Genomics Track [68] [69]. This dataset includes documents that were used for the Genomics TREC 2005 challenge. The documents contain information about correlations between genes and their mutations, and relate them to the diseases they cause.

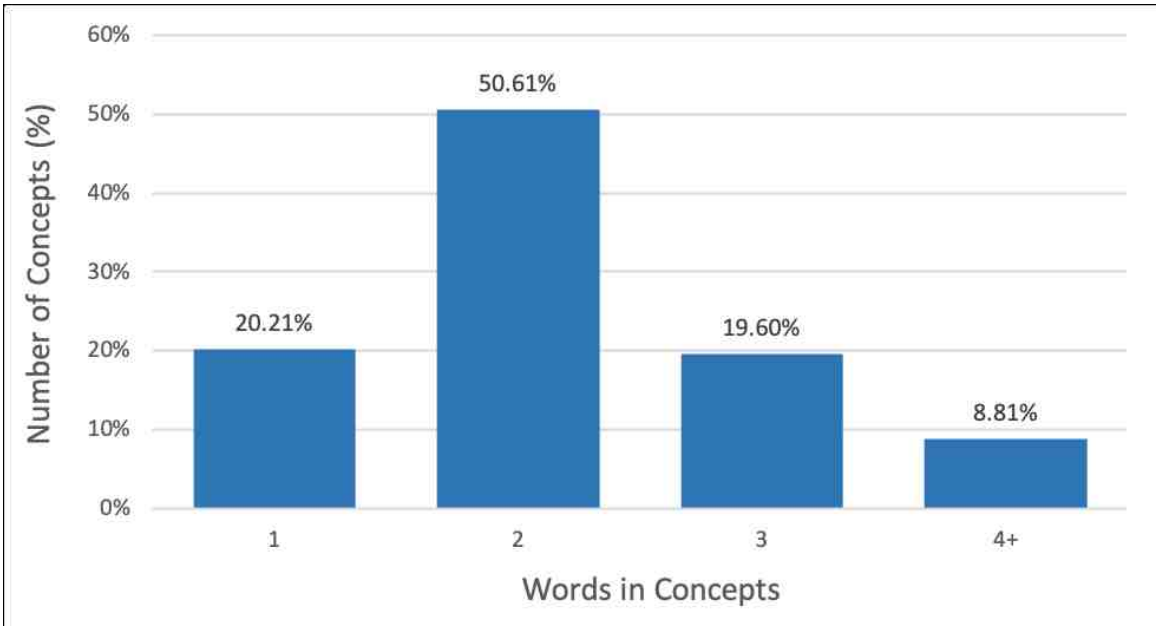

Fig. 5.3.: Distribution of the number of words over every concept identified after processing the OHSUMED dataset with MetaMap.

Table 5.3.: Overview of the TREC genomics dataset.

Disease	**Number of documents**
Multiple Sclerosis	554
Mad Cow Disease	447
Alzheimer's Disease	1201
Colon Cancer	567
Parkinson's Disease	769
Cerebral Amyloid Angiopathy	482
Breast Cancer	458

The documents are grouped by diseases, and thus every document's label is a disease. MetaMap is used to label the concepts of diseases in each of these documents (Chapter 4, Section 4.1). However, not all documents have the concept of the disease described in the 'Title' or 'Abstract'. Such documents are removed from the dataset and not used.

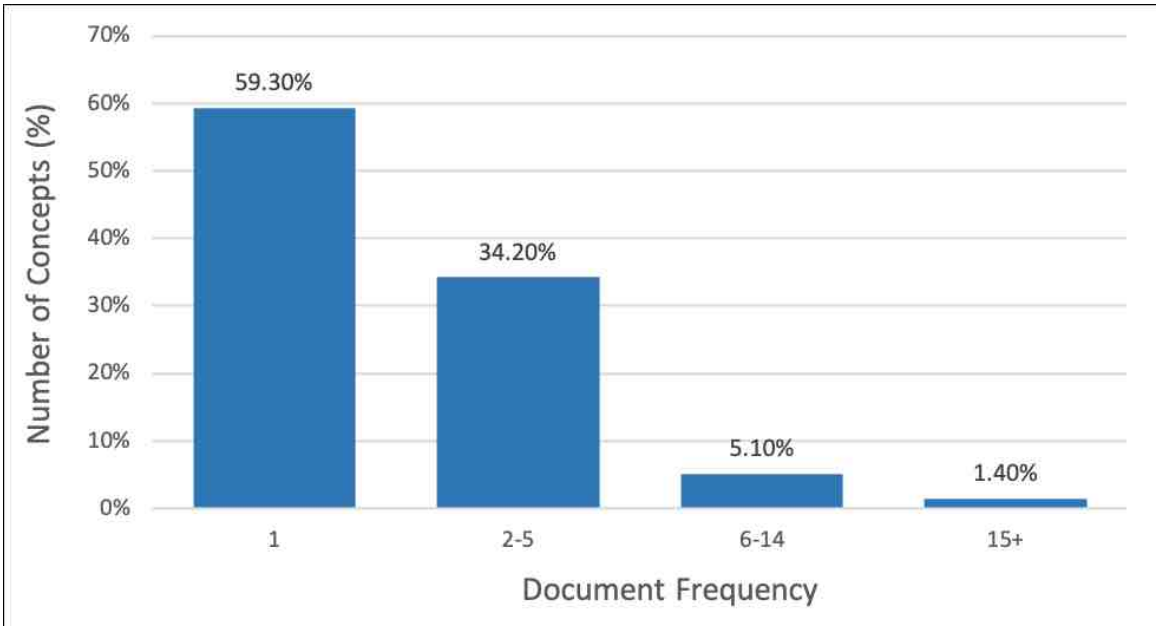

Fig. 5.4.: Distribution of the identified disease concepts over document frequency for the OHSUMED dataset.

The categories – document labels – used and the corresponding number of documents in each category from this dataset are detailed in Table 5.3.

By pre-processing the TREC dataset containing 4478 documents with MetaMap (Chapter 4.1), 2693 concepts were extracted.

Figure 5.5 shows the distribution of these concepts based on the number of words in each concept. About 12.5% of the concepts have one word, about 50% of the extracted concepts of diseases contain two words, 25% have three words, and 13.2% of them have more than four words.

This shows that a lot of concepts related to 'diseases' cannot be represented by a single word, and must be represented by phrases. This is also why individual *word* vectors are not sufficient in representing the frequencies of these multi-word concepts in documents.

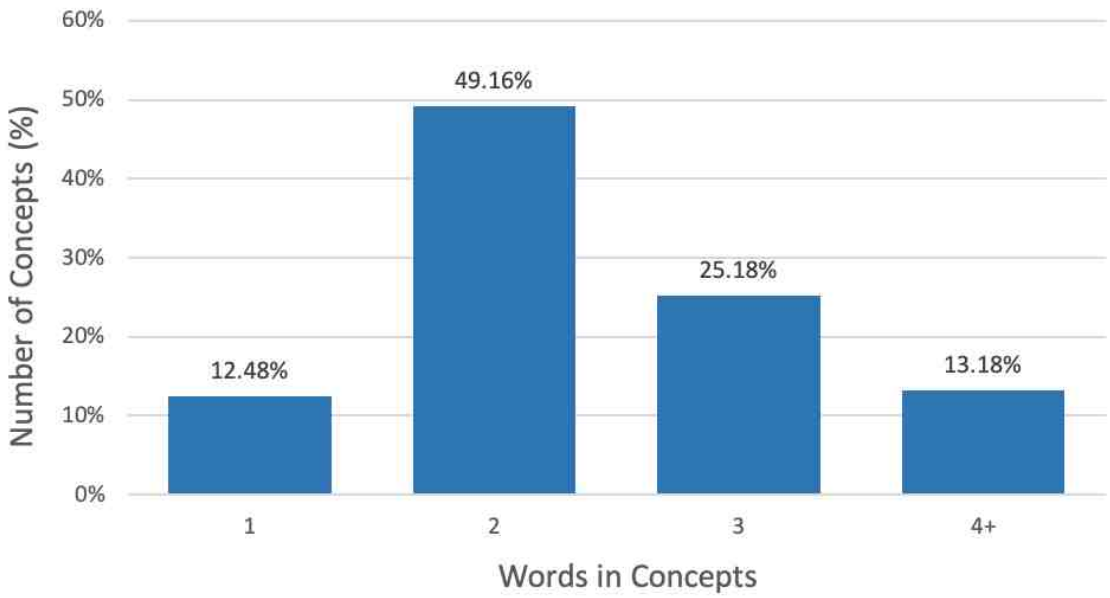

Fig. 5.5.: Distribution of the number of words over every concept identified after processing the TREC dataset with MetaMap.

Figure 5.6 shows the distribution of concepts by document frequency. Over 57% of the concepts have document frequency 1, and only 7% of the concepts have document frequency over 10. Since the content is fairly short, not many concepts of disease with same words occur in more than 3 documents. Because of the distinct labels provided as a part of the dataset, all of those labels appear in approximately as many documents as extracted of that label.

The concept-document distribution for TREC Genomics 2005 is different from Figures 5.2 and 5.4 because of the larger number of normalized concepts in the labelled TREC dataset.

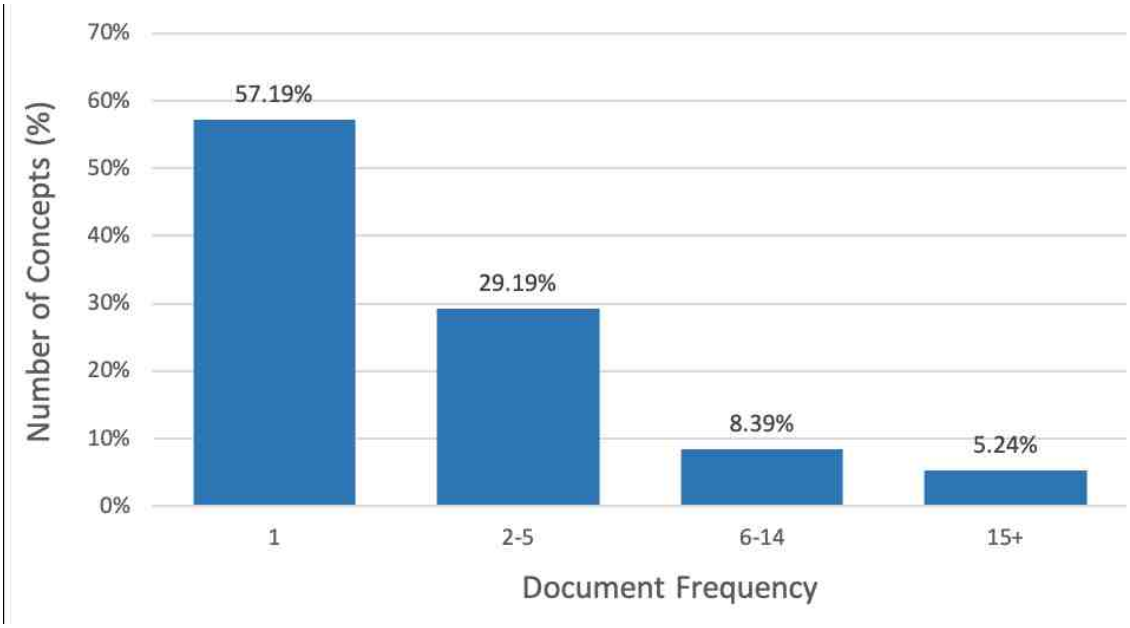

Fig. 5.6.: Distribution of the identified disease concepts over document frequency for the TREC 2005 Genomics dataset.

5.2 Electronic Health Records

The Health Information Technology for Economic and Clinical Health Act, HITECH Act, enacted in 2009 pushed for the promotion and expansion of adopting technology in health information. The HITECH Act pushed for healthcare reform that helped in paving the way towards a more broadly available and accessible EHR systems. The Act also defined three stages of meaningful use of EHRs.

- **Stage 1** contained objectives like recording demographics, maintaining medications and allergy lists, computerizing medication orders, implementing drug allergy checks, recording vital signs, record smoking status, etc.
- **Stage 2** included provisions for letting patients view, download and transmit information, sharing patient information across systems, organizations and patients.

- **Stage 3** is to improve population health outcomes, improve clinical outcomes, improve security of healthcare systems and gain more robust research data on health systems.

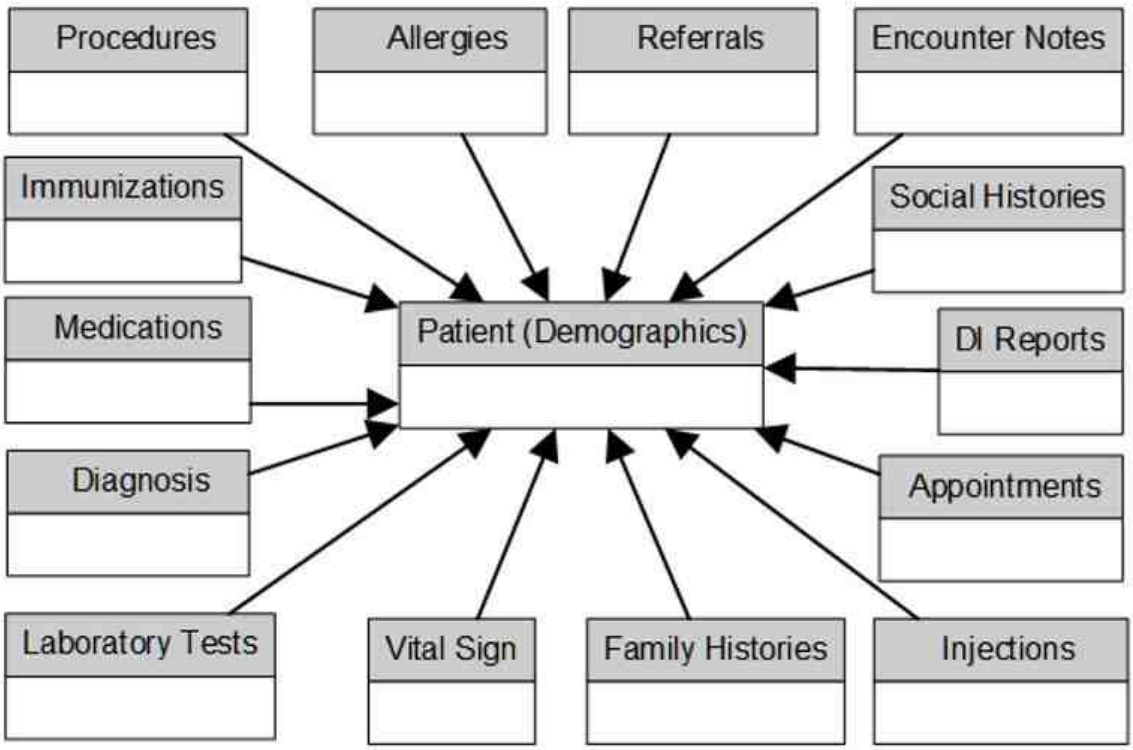

Fig. 5.7.: A sample framework of an Electronic Health Record system.

Figure 5.7 shows the primary functional modules in a typical EHR system. Many of these modules, such as medication and diagnosis, contain structured data which comprises of defined data types and is often ready to use for data mining applications.

The encounter notes (also called clinical notes) is a major component of the EHR and also includes large amounts of unstructured data. These unstructured data are mostly text written or dictated by physicians or nurses. Arguably, it is an important part of the patient's medical records. Sondhi et al. demonstrated the importance of mining the clinical notes by detecting the symptoms of Congestive Heart failure

(CHF) [70]. Previous studies also demonstrated the need for extracting clinical signs and symptoms from patient medical records and further analyzing their associations with specific diseases [71] [72].

In collaboration with Indiana University Health (IU Health), the patient data for 500 adult patients was spanning 2003 to 2017. The approval of the Institutional Review Board (IRB) was received for this study.

Some of the 500 patients have over 10 years of medical history. The dataset provided included patient data from the Diagnosis, Social History, Family History, Laboratory Results, Medications, Demographics, Vitals and Imaging modules of the EHR. The dataset also included patient notes.

Table 5.4.: Top 10 most frequent diagnoses appearing in the IU Health EHR dataset.

Diagnosis in EHR Chart	**Patient Count**
Essential (primary) hypertension	298
Hyperlipidemia unspecified	239
Unspecified Essential Hypertension	209
Atherosclerotic heart disease of native coronary artery without angina pectoris	201
Other and Unspecified Hyperlipidemia	186
Heart failure unspecified	172
Type 2 diabetes mellitus without complications	154
Cough	153
Shortness of breath	153
Congestive Heart Failure Unspecified	137

The data available in the 'Diagnosis' module of the EHR contains diagnoses made and entered into the EHR by the physician. The data from this module is also used for the billing purposes. After analysis, we found that all 500 patients had more than one diagnosis. Table 5.4 lists the most frequent diagnoses as extracted

from the diagnosis module of the EHR system, and the number of patient's records those concepts appeared in. For this dataset, some patients had hypertension and/or hyperlipidemia. It was also noticed that 'cough' and 'shortness of breath' are both found in the diagnosis module with associated ICD codes.

Table 5.5.: Top 10 most frequent social history entries appearing in the IU Health EHR dataset.

Social History in EHR Chart	**Patient Count**
Tobacco amount per day	95
Work/School description	60
Tobacco number of years	50
Complex Living Situation	48
Tobacco started at age	29
Tobacco stopped at age	28
Caffeine intake	19
Tobacco total pack years	16
Number of current partners	11
Alcohol amount average	11

The data in the 'Social History' module contains information about the social history of the patient. This includes information about the patient's tobacco consumption, caffeine consumption, living conditions, work conditions, alcohol consumption, sexual activity, etc. From the 500 patients, there is information about 204 patients' social history. All of the entries in this module are prefixed with *'SHX'*. The most frequent entries in the 'Social History' module are shown in Table 5.5, without the prefix.

'Family History' module of the EHR contains information about significant medical history of the patient's family. Out of 500, 282 patients have family history data. Table 5.6 contains the top 10 family history concepts that appeared in the EHR dataset from IU Health, and their frequencies.

Table 5.6.: Top 10 most frequent social history entries appearing in the IU Health EHR dataset.

Family History in EHR Chart	**Patient Count**
Diabetes mellitus type 2	63
Breast cancer	49
Stroke	42
Hypertension	42
Heart disease..	40
High blood pressure..	37
Heart attack..	37
Hypertension..	24
Cancer of colon	24
Coronary artery disease..	23

'Laboratory Results' and 'Medications' are other modules for which the EHR data was available, with data for 494 patients and 498 patients in them. However, the data from these modules was not used for this research and was only observed. Thus, summaries of this data is not provided.

The data also included 154,738 notes from the 'Clinical Notes' module of the EHR, for these 500 patients. These notes were most significant and are the focus of most of the work in Chapter 6, Section 6.3 and Chapter 8, Section 8.4. An example note is presented in Figure 5.8.

Associated Diagnoses: None .
Subjective:
11/30/15: 80 who presented to the hospital with 3 days history of fever and cough. She was diagnosed with CAP and was started on antibiotics. Unfortunately, she had a significant episode of hypoxemia and had to be intubated. Pinkish frothy sputum was reported after intubation. Patient has a remote history of smoking.
……
11/30/2015 06:00 Transparent Physical Examination General: intubated and sedated. Eye: Pupils are equal, round and reactive to light, Extraocular movements are intact. HENT: intubated and sedated. Neck: Supple, No lymphadenopathy. Respiratory: bilateral rales. Cardiovascular: Normal rate, Regular rhythm, No murmur. Gastrointestinal: Soft, Non-distended. Musculoskeletal: intubated and sedated. Integumentary: Warm, Dry. Neurologic: intubated and sedated. Results Review Labs Last 24 Hrs SELECT Labs ONLY
……
12/01/2015 06:52 - XR Chest PA AP Portable IMPRESSION: Diffuse bilateral airspace opacities. Interval improvement. Impression and Plan 1- Acute respiratory failure 2- Bilateral infiltrate: pulm edema vs. worsening pneumonia vs. alveolar hemorrhage (bloody sputum and HB dropped 2 grs) 3- Pneumonia 4- COPD: seen on CT chest 2014 5- Troponin elevation: troponin went up to 2 due to her respiratory failure. However, her echo is very suggestive of CAD. Appreciate cardiology. 6- CHF: sudden bilateral infiltrates and high troponin Plan Increase diuresis US of left chest and tap if needed bronch……

Fig. 5.8.: A sample note from the IU Health EHR system.

Most than 160 patients have 5 and 105 clinical notes, whereas only 40 of the 500 patients have more than 805 clinical notes over the period of 15 years. This means that most of the patients have very few encounters recorded. This amounts to lesser information in the EHR in every module. Because of this sparsity of information it becomes imperative to extract information that is not stored in the structured EHR modules.

The unstructured clinical notes need to be examined and information is extracted from notes. Information pertinent to patient history that is not present in the other EHR modules, like symptoms, but is of extreme importance to make clinical decisions should be extracted from clinical notes. Moreover, information extracted from the

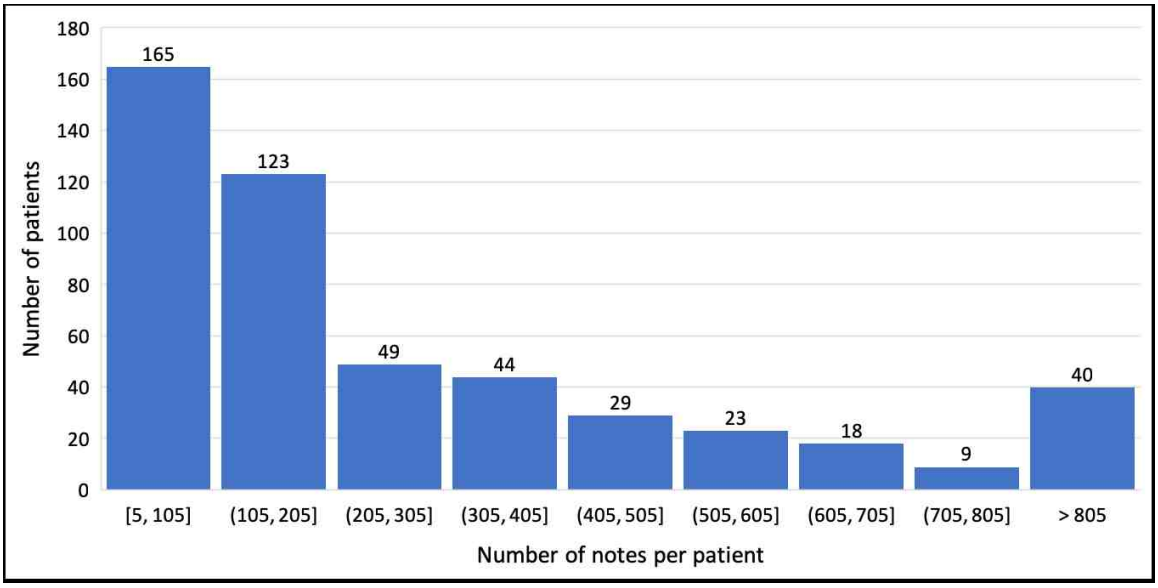

Fig. 5.9.: A histogram showing the distribution of clinical notes per patient in the IU Health EHR dataset.

notes can also help in auto-filling other modules of the notes and check for their completeness. Thus, information extraction (Chapter 4) from the unstructured clinical notes is important.

As a part of this work, research was performed using word embeddings for concepts and diseases extracted from the clinical notes, to examine associations between clinical symptoms and identify patterns in their occurrence. Other research performed includes clustering of the extracted concepts of symptoms and diseases (Chapter 6, Section 6.3). Preliminary work was also done towards identifying patient populations that are at a high risk of contracting diseases, based on their EHR records and preventive care guidelines. The identified patient population can then be targeted and approached through their physicians, and steps taken actively to prevent the onset of high risk diseases.

5.3 Preventive Care Guidelines

Preventive care is the process of identifying and providing healthcare services to patients so that they do not contract high-risk diseases. Preventive care includes services provided as a part of immunizations to more complex studies that look at risk of diseases like cancer and genetic diseases like Alzheimer's disease.

Summary
The USPSTF recommends annual screening for lung cancer with low-dose computed tomography (LDCT) in adults aged 55 to 80 years who have a 30 pack-year smoking history and currently smoke or have quit within the past 15 years.
…
Patient Population Under Consideration
The risk for lung cancer increases with age and cumulative exposure to tobacco smoke and decreases with time since quitting smoking.
…
Screening Tests
Low-dose computed tomography has shown high sensitivity and acceptable specificity for the detection of lung cancer in high-risk persons.
…

Fig. 5.10.: Abridged version of USPSTF's Lung Cancer screening recommendation statement [48].

The guidelines issued by the medical professionals and associations to generalize and facilitate healthy preventive care practices are called preventive care guidelines. These guidelines aim to answer the questions about the patient population that is at risk, diagnosing the risk, quantifying the risk, and providing appropriate preventive care services.

As a part of this research, a framework was developed to extract information from the guidelines issued by the United States Preventive Services Task Force (USPSTF) [74] [40]. USPSTF releases recommendations in the form of detailed preventive care guidelines that can be used to identify at-risk patient populations, and the rec-

Summary
The USPSTF recommends screening for abnormal blood glucose as part of cardiovascular risk assessment in adults aged 40 to 70 years who are overweight or obese.
…
Patient Population Under Consideration
This recommendation applies to adults aged 40 to 70 years seen in primary care settings who do not have symptoms of diabetes and are overweight or obese.
…
Persons who have a family history of diabetes, have a history of gestational diabetes or polycystic ovarian syndrome, or are members of certain racial/ethnic groups (that is, African Americans, American Indians or Alaskan Natives, Asian Americans, Hispanics or Latinos, or Native Hawaiians or Pacific Islanders) may be at increased risk for diabetes at a younger age or at a lower body mass index.
Screening Tests
Glucose abnormalities can be detected by measuring HbA1c or fasting plasma glucose or with an oral glucose tolerance test.
…
Screening Intervals
…
Cohort and modeling studies suggest that rescreening every 3 years may be a reasonable approach for adults with normal blood glucose levels.

Fig. 5.11.: Abridged version of USPSTF's Type 2 Diabetes screening recommendation statement [73].

ommended course to take to prevent the diseases. Some of the diseases USPSTF has issued guidelines for include atrial fibrillation, cardiovascular disease, cervical cancer, etc. Abridged versions of the USPSTF's Lung Cancer [48] and Type 2 Diabetes Mellitus [73] recommendation statements are shown in Figures 5.10 and 5.11 respectively.

Extracting information from these preventive guidelines that can be integrated into a module running in the background of an EHR can greatly assist in providing better preventive care to patients.

6. CONCEPT REPRESENTATION

Generating word embedding through using neural networks for biomedical concepts has drawn attention in the areas of natural language processing and machine learning [75] [76]. Based on the discussions about the various types of word embeddings presented in Chapter 2, Word2Vec (Chapter 2, Section 2.4) was chosen for further exploration during this work.

Word2Vec is a different approach from the other types of models, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). The continuous vector representation, as distributed representation of words are learned using a three-layer recurrent neural network with a skip-gram model.

The skip-gram architecture of Word2Vec (Chapter 2, Section 2.4.2) is a robust architecture that relies on the probabilities of a word to appear around its surrounding words to generate a vector representation for every word in the corpus.

The output vector representations of the words preserve the distances between words so that the words that have semantic and syntactic associations in the raw text corpus are located in close proximity to one another. The dimension of the vectors created depends on the number of neurons in the hidden layer of the neural network. In this research, this recurrent neural network learning model is used to create the distributed representation for biomedical concepts are extracted from the document corpus.

Each extracted biomedical concept is treated as a word w_t. At the end of the training process, each word vector is represented as shown in Equation 6.1.

$$WV = (wv_1, wv_2, \ldots, wv_m) \tag{6.1}$$

where m is the dimension of the vector, which corresponds to the number of neurons in the hidden layer of the recurrent neural network used in the training of the Word2Vec algorithm.

Relationships between word vectors are derived based on either their word vectors (Equation 6.1) or the similarity between word vectors. The similarity between word vectors is calculated by the cosine distance between two word vectors. For two word vectors WV_i and WV_j with m dimensions each, and θ angle between them, the cosine distance is calculated as shown in Equation 6.2.

$$\cos\theta = \frac{\mathbf{WV_i} \cdot \mathbf{WV_j}}{||\mathbf{WV_i}|| \cdot ||\mathbf{WV_j}||} \tag{6.2}$$

Cosine distance is also called the cosine similarity and Equation 6.2 is further simplified in terms of the vectors, to calculate the the cosine similarity, $s_{i,j}$ as shown in Equation 6.3.

$$\cos\theta = s_{i,j} = \frac{\sum_{k=1}^{m} \mathbf{WV_{ik}} \cdot \mathbf{WV_{jk}}}{\sqrt{\sum_{\mathbf{k=1}}^{\mathbf{m}} \mathbf{WV_{ik}^2}} \cdot \sqrt{\sum_{\mathbf{k=1}}^{\mathbf{m}} \mathbf{WV_{jk}^2}}} \tag{6.3}$$

The value of the cosine similarity can be in the range $[-1, +1]$. A cosine similarity of $+1$ represents words that are identical in the vector space, -1 represents concepts that are opposite and 0 represents words that are orthogonal (or unrelated) to one another. A higher cosine similarity between similar words, and lower cosine similarity between unrelated words is the desired behavior of good, reliable word embeddings.

Even though word vectors can be created by using word embeddings, the essence of this work relies on 'concept vectors'. This extension of word embeddings to 'phrase vectors' or 'concept vectors' that more accurately represent biomedical concepts.

The first step to representing biomedical concepts is by calculating vectors for *concepts*, instead of *words*. Mikolov et al. presented a method to generate phrase vectors using collocations [11]. The proposed method identifies phrases like publications (New York Times), airlines (Air China), countries (United States), cities (San

Francisco), etc. However, because the phrase generation algorithm is rooted in collocations, it works well only when the given set of words comprising a phrase, appear multiple times in the training corpus.

For biomedical concepts, especially within smaller corpora, appear fewer times. Another issue with biomedical concepts is the difference in their representation across authors, publications and vocabularies. For example, 'lung cancer', 'carcinoma of the lung', 'lung carcinoma' and 'malignant neoplasm of the lung' all mean the exact same thing. A similar problem exists because of abbreviations ('DM2' for 'Type 2 Diabetes Mellitus') and each doctor's individual representation in the Electronic Health Record.

In a typical vector space, all of these concepts discussed above, would be treated individually as separate vectors. Moreover, each word in the concept would be treated as its own vector, influencing the results obtained in the downstream tasks dependent on the word vectors.

Thus, biomedical concept vectors are generated for concepts using either an aggregation of word vectors to create concept vectors (Section 6.1) or by training a word embeddings model at the concept level (Section 6.2).

6.1 Word-Based Representation

Aggregated concept-based representations are calculated by summing individual vectors of each word of the concept. Word vectors contain information about the word in the form of numbers. Adding vectors together does not lead to a loss in information, but in the addition of properties of multiple words. Aggregation by summing and averaging are widely used methods of creating phrase, sentence, paragraph and document vectors. However, summing may lead to a loss in information if the number of word vectors added together are large.

In this work, because the length of concepts is not very long concept vectors are generated by adding word vectors. Thus, the vector for the concept 'lung cancer' is calculated as the sum of vectors of 'lung' and 'cancer'. If the concept is a single word

concept, its vector is as generated by the word embeddings model. For a concept with i words, the vector for the aggregated word-based representation is generated as shown in Equation 6.4.

$$CV = \sum_{k=1}^{l} (WV_{k1}, WV_{k2}, \ldots, WV_{km}) \tag{6.4}$$

The word embeddings model used for base word embeddings were created using a skip-gram model of Word2Vec. The training corpus for this model consisted of abstracts from PubMed, full text articles from PubMed Central, and an English Wikipedia dump. The word embeddings model was trained by Pyysalo, et. al [21], and is available for download [77].

Preliminary tests to examine the quality of word vectors can best be performed by measuring their cosine similarities against other word vectors. To do this, word vectors are first calculated for each of the disease concepts appearing in the PubMed (Chapter 5, Section 5.1.1) and Ohsumed (Chapter 5, Section 5.1.2) corpora.

A similarity matrix S is then computed that contains the similarity values for every concepts, with every other concept using the similarity scores (Equation 6.3). This matrix is extremely useful as it reduces the compute time during each iteration to calculate the similarity scores for every concept against the whole dictionary. For a concept dictionary of size N, Equation 6.5 shows an example similarity matrix.

$$S = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,N} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N,1} & s_{N,2} & \cdots & s_{N,N} \end{pmatrix} \tag{6.5}$$

After calculating these concept vectors and computing the similarity matrix S, some of the most frequent concepts are handpicked as primary concepts. The most similar disease concepts by cosine distance to these primary concepts are then exam-

Table 6.1.: Examples of concepts and the top 3 closest concepts based on the similarity scores from pre-trained word embeddings with concept vectors generated by aggregating word-based representations.

Primary Concept	**Most Similar Concepts**	**Similarity Score**
hypertension	essential hypertension	0.813
	hyperlipidaemia	0.692
	dyslipidemia	0.659
endothelial dysfunction	dysfunction	0.739
	renal dysfunction	0.660
	cortical dysfunction	0.639
carpal tunnel syndrome	bilateral carpal tunnel syndrome	0.970
	cts carpal tunnel syndrome	0.957
	carpal tunnel	0.941
diabetes	diabetes mellitus	0.918
	diabetes mellitus type ii	0.868
	dm diabetes mellitus	0.845
cardiovascular disease	cardiac diseases	0.8181
	metabolic diseases	0.8179
	heart diseases	0.787

ined. It can be seen from Table 6.1 that for most of the primary concepts, the most similar concepts are other either equivalent representations for the same biomedical concept or other concepts that are highly similar to the primary concept.

6.2 Concept-Based Representation

Concept based representations are calculated by considering the concept as a single entity. Thus, the concept 'diabetes mellitus type 2' is considered as a single entity instead of calculating the vectors of each word of the concept. The concept and the context of the context is used as an input to the recurrent neural network as a single unit and the vector of the concept is the same as the one output by the network.

The vector for the concept based representation thus generated is the same as Equation 6.1, but at the concept-level, and is shown in Equation 6.6.

$$CV = (cv_1, cv_2, \ldots, cv_m) \tag{6.6}$$

The input dataset for training includes PubMed Central's Non-Commercial Open Access (Chapter 5, Section 5.1.1), the subset of documents from MEDLINE that are a part of TREC 2005's Genomics Track (Chapter 5, Section 5.1.3) and the Ohsumed collection (Chapter 5, Section 5.1.2). All of these models were trained either on separate corpora or included other corpora. A variant of this model was also trained on the clinical notes from IU Health (Chapter 5, Section 5.2). All of these models were used individually to ensure equivalent comparison, and used based on application.

It is seen from the details about data sources (Chapter 5, Figures 5.6, 5.2, 5.4) that a lot of concepts appear just once in the corpora. Normally, top 5% of the most frequent words are dropped from the training, and so are words that do not appear in the corpus more than a few times. However, in training the custom models, it was ensured that none of the words from the vocabulary were truncated. This was done to ensure that even concepts that appear a handful of times have concept vectors. Training individual word embeddings model ensures that the whole vocabulary is a part of the embeddings vocabulary.

Words that match biomedical concepts are replaced with their preferred concept from MetaMap (Chapter 4 Section 4.1), before using the text as an input to train the word embeddings models. The training for these word embeddings model is performed using the gensim library [78]. The training parameters were set to 300 hidden layer nodes, and a variable window size.

Table 6.2.: Examples of concepts and the top 3 closest concepts based on the similarity scores from embeddings trained as concept-based embeddings on the TREC Genomics 2005 corpus.

Primary Concept	**Most Similar Concepts**	**Similarity Score**
alzheimer disease	alzheimer	0.829
	parkinson disease	0.813
	huntington disease	0.688
multiple sclerosis	multiple sclerosis relapsing remitting	0.661
	ms	0.633
	parkinson	0.600
cerebral amyloid angiopathy	caa	0.601
	cerebral	0.486
	hereditary cerebral hemorrhage with amyloidosis dutch type	0.471
colon cancer	colorectal cancer	0.799
	cancer of colon	0.755
	cancer of the colon	0.724

The trained embedding includes vectors for the biomedical concepts. Upon completion of the training, concept vectors are queried and a similarity matrix similar to Equation 6.5 is calculated with the concept-based representations. Table 6.2 shows

the top 3 closest concepts for some of the handpicked primary concepts, after a similarity matrix is generated using concept-based embeddings trained on the TREC Genomics 2005 corpus (Chapter 5, Section 5.1.3).

On comparing Table 6.2 with Table 6.1, it is evident that the similarity scores for concept-based representations are lower. At the same time, concept-based representations also capture similar disease concepts better, whereas word-based representations apture synonym concepts better. The lower similarity scores require algorithms to be modified in order for them to use the concept-based representations (discussed in detail in Chapter 7). One noteworthy similarity score from Table 6.2 is that of the concept 'cerebral amyloid angiopathy' with its own abbreviation, and other most similar concepts. This influences the document clustering results in a big way, resulting in none of the discussed TREC Genomics 2005 results being able to successfuly cluster 'cerebral amyloid angiopathy' documents into a separate cluster (Chapter 7, Section 7.7.3 and 7.4.3).

6.3 Concept Clustering

Exploratory work towards using symptoms concepts in conjunction with disease concepts was also performed [79]. However, there are limitations in how many symptom and disease concepts appear in conjunction in biomedical documents since most documents do not discuss trajectories, but focus on causes and treatment plan.

The corpus chosen for this exploratory work was the clinical notes data cohort for 500 patients from IU Health (Chapter 5, Section 5.2). The dataset contains 154,738 clinical encounter notes for 500 patients, spanning 15 years. From the clinical notes, disease and symptoms concepts are extracted and normalized using MetaMap (Chapter 4, Section 4.1).

The clinical notes are used as an input to the word embeddings algorithm to create word vectors for the concepts using concept-based representations (Chapter 6, Section 6.2). Using these word vectors, the similarity matrices of disease concepts against other disease concepts (S_D, and symptoms concepts against other symptoms (S_S) are calculated (Equation 6.5).

Clustering is performed at the concept level separately for the symptoms and diseases concepts. The similarity matrices (S_D and S_S) were used as inputs to the k-means clustering algorithm (Chapter 3, Section 3.1.2), results are presented for $k = 50$.

6.3.1 Disease Concept Clustering

Examining the most similarity disease concepts against other disease concepts, shows the relationships extracted from the clinical notes among the disease concepts. These results are obtained from the similarity matrix for disease concepts S_D.

The results of Table 6.3 show that the most similar concepts identified by training a word embeddings model on concept-based representations are most often, other disease concepts that are equivalent or closely related to the primary concept.

However, the similarity scores for these concepts are much lower than those seen in the case of word-based representations (Table 6.1), and slightly lower than concept-based representations trained on the TREC Genomics 2005 corpus of biomedical documents (Table 6.2). These scores are lower on account of the unstructured and free-form nature of clinical notes, which makes it harder to learn reliable word embeddings from this type of data.

After evaluating these similarity scores, an examination of the clustering performed on the similarity matrix, to cluster concepts is performed. The clustering results for some of the significant clusters are shown in a tabular form in Table 6.4.

Table 6.3.: Examples of disease concepts and the top 3 closest disease concepts based on the similarity scores from embeddings trained as concept-based embeddings on the IU Health EHR clinical notes data.

Primary Disease	**Most Similar Diseases**	**Similarity Score**
chronic obstructive pulmonary disease	severe chronic obstructive pulmonary disease	0.518
	pulmonary disease obstructive	0.496
	chronic obstructive lung disease	0.471
diabetes mellitus type 2	diabetes type 2	0.629
	diabetes mellitus type ii	0.582
	diabetes type 2 on insulin	0.570
breast cancer	breast ductal carcinoma	0.642
	breast cancer female	0.626
	invasive ductal carcinoma breast	0.622
coronary artery disease	coronary disease	0.662
	peripheral arterial disease	0.538
	coronary artery disease with myocardial infarction	0.534

Investigating the disease concepts within the clustering results, it was found that some of the clusters contained diseases that were highly related or the same disease at different stages, such as different stages of chronic kidney disease, or different representations of the same disease.

Table 6.4.: Cluster-wise results of clustering disease concepts based on their similarity scores.

Cluster 5	**Cluster 6**	**Cluster 4**
diabetes type 2	cardiomyopathy	congestive heart failure
diabetes type ii	stroke	failure heart
diabetes mellitus type ii	ischemic cardiomyopathy	chronic systolic heart failure
type ii diabetes	nonischemic dilated cardiomyopathy	diastolic heart failure
hyperglycemia	chronic atrial fibrillation	acute heart failure
diabetes type 2 on insulin	aortic stenosis	biventricular failure
hypertension	sinus tachycardia	left ventricular failure
ESRD	atrial fibrillation	chronic heart failure
	nonischemic dilated cardiomyopathy	hypoxemic respiratory failure
	rapid atrial fibrillation	chronic diastolic heart failure

Cluster 5 contains different representations of 'type 2 diabetes mellitus', but 'hypertension' is also included in that cluster. This occurs because a lot of encounter notes contain 'diabetes' in conjunction with 'hypertension'.

Cluster 6 is most representative of different types of 'cardiomyopathy' and diseases related to heart muscles. The literature shows that 'ischemic cardiomyopathy', 'atrial fibrillation' and 'aortic stenosis' are causes or conditions associated with 'dilated cardiomyopathy' [80].

Cluster 4 contains different types of heart failures. This shows that the concept-based representations built with word embeddings can distinguish sufficiently between diseases of heart muscles and heart failures. However, 'hypoxemic respiratory failure'

is also included in this cluster. After investigating the clinical notes of the patients, it is found that the 'hypoxemic respiratory failure' co-occurred with heart failure in some patients' clinical notes.

Cluster 15 (not shown in Table 6.4) contains only one disease, erythema, which means this disease did not co-occur with other diseases in this study cohort.

6.3.2 Symptom Concept Clustering

Table 6.5 shows the most similar symptom concepts to other symptom concepts as extracted from the clinical notes of the IU EHR data.

Table 6.5.: Examples of symptom concepts and the top 3 closest symptoms concepts based on the similarity scores from embeddings trained as concept-based embeddings on the IU Health EHR clinical notes data.

Primary Symptom	**Most Similar Symptoms**	**Similarity Score**
vertigo	dizziness	0.461
	lightheadedness	0.415
	headaches	0.370
chronic pain	chronic back pain	0.627
	back abdominal pain	0.517
	intractable pain	0.487
swollen legs	cramps in legs	0.312
	swelling of legs	0.209
	swollen feet	0.302
breast pain	groin pain	0.635
	rib pain	0.627
	flank pain	0.604

As seen from Table 6.5 symptom concepts that are similar, occur with similar diseases or are occur in similar areas of the body.

Table 6.6.: Cluster-wise results of clustering symptoms concepts based on their similarity scores.

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
pitting edema	headache	chest tightness	joint stiffness	seizure
massive edema	dizziness	chest pain	joint swelling	spasm
pedal edema	headaches	chest discomfort	knees stiffness	tremor
hand edema	vertigo	chest pressure	costovertebral angle tenderness	tremors
edema knees	generalized headache	pain in chest	joint crepitus	dystonia
postpartum hemorrhage	global headache	chronic chest pain	decreased grip strength	cramp
extremity edema	chronic vertigo	acute chest pain	stiffness of wrist	ataxia
bilateral pedal edema	headache throbbing	chest wall pain	painful joints	clonus
penile edema	morning headache	chest pain angina	stiffness fingers	asterixis
	intermittent dizziness	chest burn	facet arthropathy	shakes

Table 6.6 shows the results of clustering symptoms performed with using the symptom similarity matrix (S_S) as an input. The clustering is performed by setting $k = 50$ using the k-means clustering algorithm.

The results show that the symptoms clustered together are those that occur alongside each other or in the progression of similar diseases. Each cluster demonstrated here is associated with one category of symptom. For example, cluster 0 is about different types of edema, cluster 1 is about headache and dizziness, cluster 2 is for symptoms of the chest, and cluster 3 is about joints related symptoms. The cluster 4 contains a lot of single word symptoms. These single word symptoms were mostly related to movement disorders in one or more parts of the body.

7. DOCUMENT CLUSTERING

Document clustering is a text mining technique used to provide better document search and browsing in digital libraries or online corpora. The large repositories of unlabeled biomedical data and articles available online has led to a continuing need for development of techniques to discover and search these documents and articles.

Biomedical document clustering based on the concepts of diseases can provide an overview of the literature repository based on the diseases and relationships between the diseases, so that researchers can further explore or review the articles in certain clusters that are related to their research interests. Biomedical document clustering is different from the general text document clustering task because in the latter, semantic similarities between words or phrases are not usually considered.

A medical concept of disease might be represented in different forms, and some medical concepts of diseases might be highly correlated. For example, 'Type 2 Diabetes' is the same concept of disease as 'Diabetes Mellitus Type 2'. 'Hypertension' often co-occurs with 'Stroke'. In order to capture the semantic similarities between words or phrases, previous research on document representation reforming relies on using existing ontology such as MeSH or WordNet to identify the semantic relationships. However, this increases the dependence onto these ontologies and requires is difficult without normalization of the disease concepts.

These limitations of previously used document clustering techniques in the biomedical domain, make word embeddings an appropriate choice for further exploration. In this chapter, a framework for biomedical text clustering and visualization based on the concept embedding of diseases is proposed and evaluated. Concept representations were presented as a part of Chapter 6. Various clustering and visualization techniques were discussed in Chapter 3. The document clustering framework can be summarized as Figure 7.1.

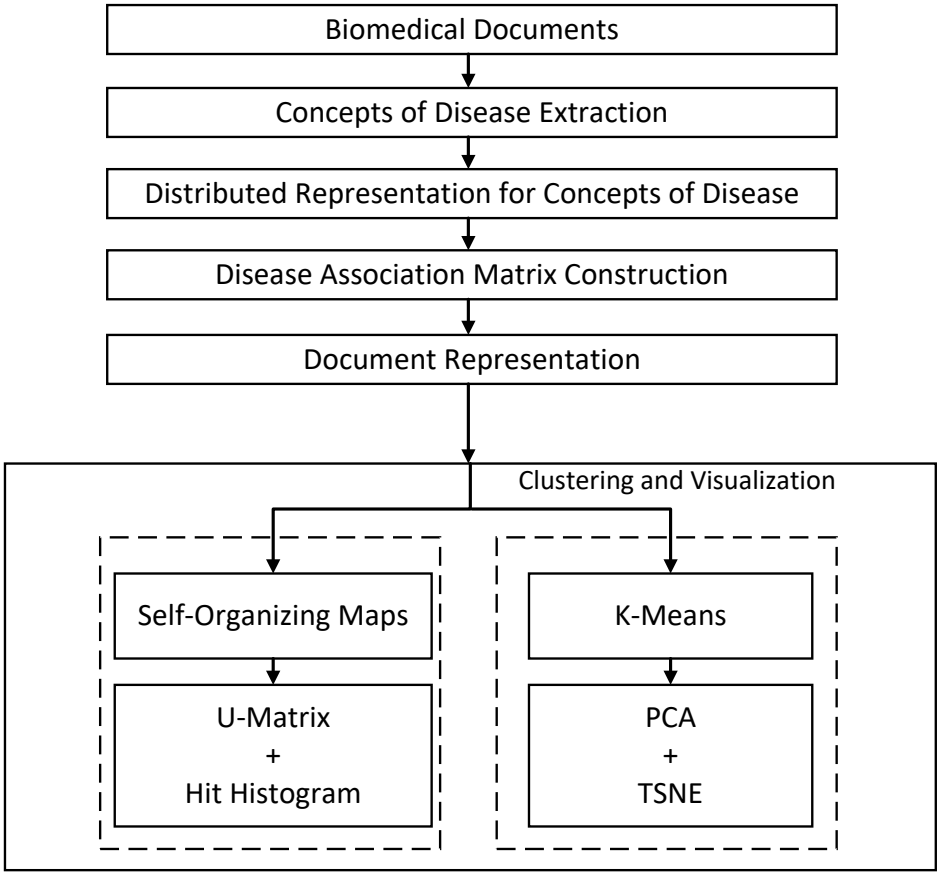

Fig. 7.1.: A summary of the document clustering framework.

Further details about the document representation in the vector space are discussed in Section 7.2, whereas the document vector generation is described in Sections 7.3 and 7.5. The results of each of the proposed techniques are discussed in Sections 7.4 and 7.7 respectively.

7.1 Literature Review

A lot of research has been done in biomedical document clustering in the past decades. Some of it focused on document presentation reforming with methods based on medical ontology, or on using weighting schemes other than TF-IDF, while some others focused on investigating various clustering algorithms.

Yoo et al. [81] used a graphical representation method to represent a set of documents based on the MeSH ontology, and proposed the document clustering and summarization with this graphical representation. They gained comparable results on clustering performance and also provided some visualization on the document's cluster model based on the relationships of the terms.

Similarly, Zhu et al. used a combination of semantic similarity described by the MeSH thesaurus and similarities of documents to generate a similarity matrix between documents [65]. The similarity matrix was calculated by comparing the distances of the MeSH headings with the MeSH headings of other documents. The distance measurement relies on the hierarchy of the MeSH thesaurus. Thereafter, the authors used four spectral clustering algorithms to cluster the documents.

Both these two research provided visualizations of clusters of documents. However, the visualization relied on the MeSH ontology and independent of the contents the documents.

Zhang et al. reviewed three different ontology based term similarity measurements: path-based [82], information content-based [83] and feature-based [84], and then proposed their own similarity measurement and term re-weighting scheme [34]. The authors used k-means algorithm for document clustering. Based on the results, some of them are slightly worse than the results from proposed word-based weighting scheme [34]. The authors mentioned that the poor performance in certain cases could be attributed to the limitation of the domain ontology, term extraction and sense disambiguation [34]. Visualization of document clustering was not included in this research.

Logeswari et al. proposed a concept weighting scheme based on the MeSH ontology and tri-gram extraction to extract concepts from the text corpus [33]. The semantic relationship between tri-grams are weighted through a heuristic weight assignment of four predefined semantic relationships. The authors proposed a concept weight

calculation framework based on identity and synonym relationships in MeSH. The k-means clustering algorithm was used, but the visualization of the clustering results was not investigated.

Gu et al. proposed a concept similarity measurement by using a linear combination of multiple similarity measurements based on the MeSH ontology and the local content which includes TF-IDF weighting and co-efficient calculation between related document sets [85]. A semi-supervised clustering algorithm was employed at the stage of document clustering. Clustering visualization was not discussed.

Some research has been done about the visualization process to support biomedical literature search. Gorg et al. developed a visual analytics system (Bio-Jigsaw) by using the MeSH ontology [86]. This research demonstrated how visual analytics can be used to analyze a search query on a gene related to breast cancer. Neither document representation nor document clustering were discussed.

7.2 Document Representation

The first step for document clustering is to convert the textual documents to a vector space representation that can be used as an input to the clustering algorithms. Because the experiments conducted are based on the concepts of diseases, these concepts are first extracted from the documents by using UMLS MetaMap (Chapter 4, Section 4.1). All of the disease concepts that appear in these documents are used as the vocabulary for the corpus.

A vector space model derived from the concept embeddings vector space, is used to represent a biomedical document in a given corpus. Each document vector contains the same number of units as the size of the corpus' vocabulary. A generic document vector is shown in Equation 7.1.

$$DV = (dv_1, dv_2, \ldots, dv_N) \tag{7.1}$$

A corpus of M documents can be represented in terms of these individual document vectors as Equation 7.2.

$$CM = (DV_1^T, DV_2^T, DV_3^T, \ldots, DV_M^T)$$
$$CM = \begin{pmatrix} dv_{1,1} & dv_{1,2} & \cdots & dv_{1,M} \\ dv_{2,1} & dv_{2,2} & \cdots & dv_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ dv_{N,1} & dv_{N,2} & \cdots & dv_{N,M} \end{pmatrix} \tag{7.2}$$

A similarity matrix for the vocabulary of extracted concepts (S) is also calculated (Chapter 6, Equation 6.5). The TF-IDF scores for the corpus, with the vocabulary of extracted concepts are also calculated (Chapter 2, Section 2.3). Sections 7.3 and 7.5 discuss different weighing schemes used to combine the TF-IDF and vector similarity scores to compute the values of the individual units of the corpus matrix CM (Equation 7.2).

7.3 Document Weighting Scheme

Concepts are first distinguished as appearing or not appearing in the document. All the concepts C that appear in the document d of corpus D have their TF-IDF score non-zero, which implies Equation 7.3 [87].

$$\begin{aligned} tf_{c_i,d} \cdot \log \frac{|D|}{df_{c_i,D}} \neq 0 \quad \forall c_i \in C, \forall c_i \in d, d \in D \\ tfidf_{c_i,d,D} \neq 0 \quad \forall c_i \in C, \forall c_i \in d, d \in D \end{aligned} \tag{7.3}$$

where $tf_{c_i,d}$ is the term-frequency of concept c_i in document d (Chapter 2, Equation 2.2), $df_{c_i,D}$ is the document frequency of the concept c_i in corpus D, and $|D|$ is the size of the corpus (Chapter 2, Equation 2.4).

For a concept c_i that appears in the document, the value for the term $dv_{i,d}$ is calculated by summing the similarity scores of all of the concepts ($C = \{c_1, c_2, \ldots, c_K\}$) that co-occur in the document and the $tfidf$ score for that concept. On the other hand, for concepts that do not appear in the document, the value for the term $dv_{i,d}$

is calculated by taking a weighted average of the most similar J terms that appear in the document. For experiments conducted with this weighing scheme, J was set to 3. The document weighing scheme can thus be summarized by Equation 7.4.

$$dv_{i,d} = \begin{cases} tfidf_{c_i,d,D} + \sum_{k=1}^{K} s_{i,k} & tfidf_{c_i,d,D} \neq 0 \\ \sum_{j=1}^{J} \frac{J-(j-1)}{J} s_{i,j} & tfidf_{c_i,d,D} = 0 \end{cases} \tag{7.4}$$

where $tfidf_{c_i,d,D}$ is the $tfidf$ value for concept c_i in document d in corpus D (Equation 7.3, $s_{i,j}$ is the similarity between concept c_i and concept c_j, c_k is any concept that is in document d and $i \neq k$, and c_j is the j^{th} most similar concept to c_i such that c_j is in document d.

By using this weighting scheme, the representation measures the occurrences of different representations of the same or similar concepts. For example, if 'diabetes' occurs in one document, but 'diabetes mellitus' occurs in another document, by using TF-IDF weighting scheme, their values would be 0 for documents in which the concept does not appear. However, by using the proposed weighting scheme, they are weighted based on the similarity between the concept and its closest concepts. Thus, for the document that does not contain the concept 'diabetes mellitus', instead of using 0, the similarity score between 'diabetes mellitus' and other concepts that appear in the document is used.

7.4 Preliminary Results

The document weighting scheme proposed (Section 7.3) is tested using the PubMed Central – Open Access (Chapter 5, Section 5.1.1), Ohsumed (Chapter 5, Section 5.1.2) and TREC 2005 Genomics track (Chapter 5, Section 5.1.3). After extracting concepts and generating the document vectors, clustering is performed using self-organizing maps.

Self-organizing maps (Chapter 3, Section 3.1.1) have been used for document clustering after concepts extraction and document representation using the proposed weighting scheme. The size of the map is 10 by 10 which contains 100 neurons. The training iterations are set to be 50,000. Upon the completion of the training and clustering, DB index (Chapter 3, Section 3.2.1) is used to evaluate the best clustering. Self-organizing maps are best visualized with the U-matrix and his histogram (Chapter 3, Section 3.3.1).

7.4.1 PubMed Central – Open Access

Through the U-matrix and hit histogram, 11 clusters (Clusters 2, 3, 4, 6, 7, 8, 9, 10, 11, 12 and 13 in Figure 7.2) are identified initially. A neuron of each cluster is selected as centroid, then DB index based on the partitions is calculated to decide whether the partition is optimal. The lower the DB index value is, the better the partition is. It is observed that lower DB index value is returned when the neurons with highest number of hits are chosen as the centroids. Figure 7.2 shows the major clusters visualized on the SOM map. The centroids which are selected through the calculation of the DB indexes are marked with a red dotted circle on the hit histogram. The clusters are marked with a black boundary.

The visualized major clusters do not include all the input data. Some of the data hit the neurons that are far away from the 14 centroids as visualized. In order to fully evaluate the best number of clustering partitions to cover all the input data instances, the number of clusters were increased by adding the neurons that are not covered by the 14 clusters as centroids.

Figure 7.3 shows that the DB index value decreases as the number of clusters are increased. That is because adding clusters to separate the data that is far away from the existing centroids creates better cluster partitions.

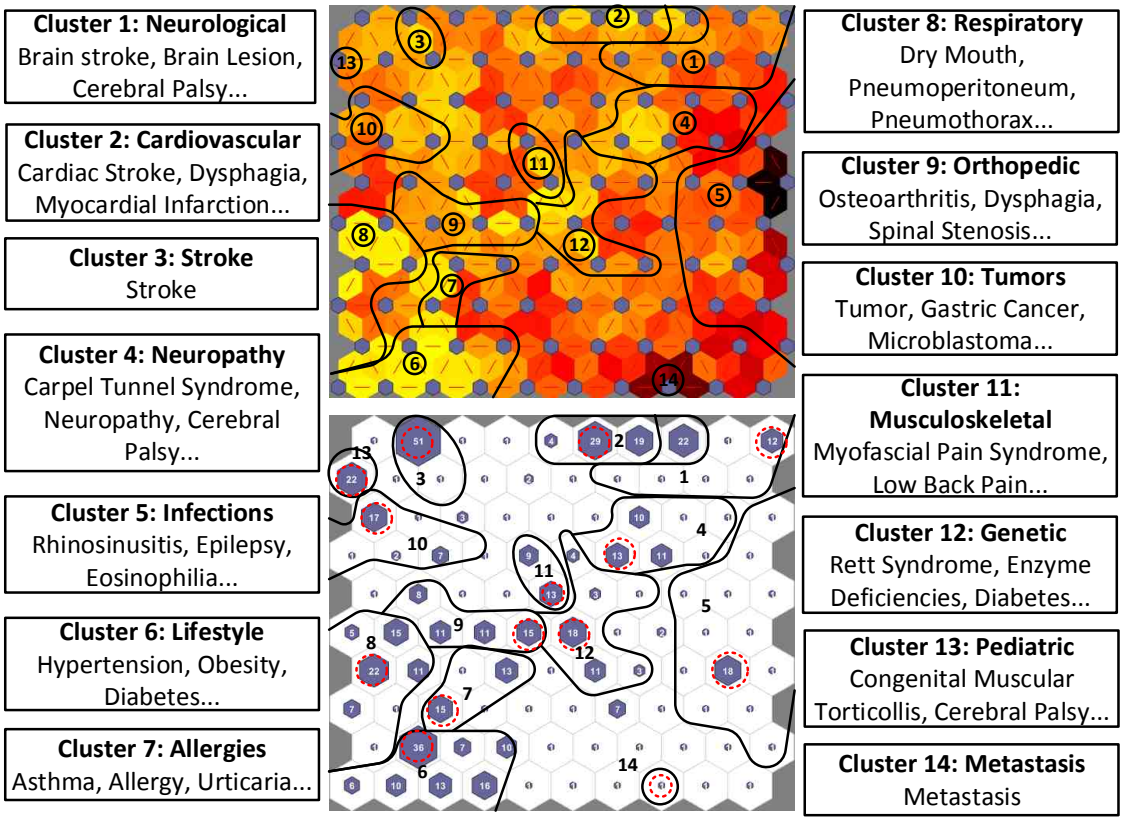

Fig. 7.2.: Clustering visualization for PubMed Central – Open Access using the document weighting scheme and clustered with self-organizing maps.

Through further analysis on the major clusters, it is discovered the concepts of diseases of each cluster as shown in Figure 7.2. A majority of documents in cluster 1 are articles that discuss neurological diseases like brain strokes, brain lesions, cerebral palsy and diseases that lead to speech disorders; most of the documents in cluster 2 are related to cardiovascular diseases such as 'hypertension', 'coronary artery disease', 'ischemic strokes' and so on. Cluster 1 and cluster 2 have one overlapped neuron on the top right of the map, it is because over half of the documents that hit this neuron discuss both neurological and cardiological 'strokes' concepts. The distances between the neurons within cluster 5 are larger than that of the other neurons. The diseases discussed in the documents in this cluster include infections like 'rhinosinusitis',

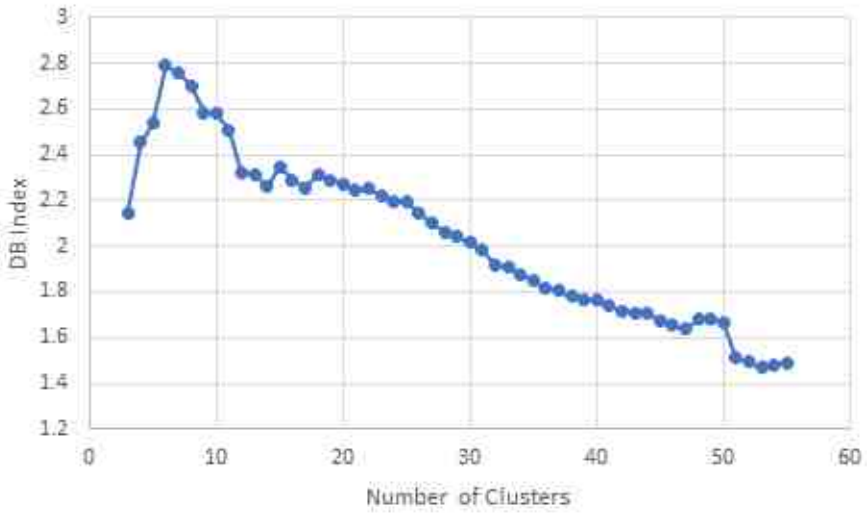

Fig. 7.3.: DB index evaluation for PubMed Central – Open Access using the document weighting scheme and clustered with self-organizing maps.

'epilepsy', 'eosinophilia' and so on. Other concepts that are found in this cluster are 'seizures' and obstructions of intestines and throat. Although these concepts are not very closely related, they are more closely related to each other than to the concepts in other clusters. Cluster 6 has documents related to 'obesity', 'diabetes', 'hypertension' and 'hyperglycemia'. The concept 'coronary artery disease' is also discussed in some documents of this cluster. It was found that this occurred because some articles discuss 'coronary artery disease' as a possible outcome of 'hypertension', 'hyperglycemia' or their combination. Cluster 11 has neurons within short distances of cluster 9. This proximity is also seen in the form of the diseases discussed by the documents of these clusters, since muscle pain and orthopedic concepts are highly related to each other. Cluster 12 is very closely located to cluster 4, 9 and 11. Upon analyzing the documents in this cluster, it was found that genetic disorder related diseases that are discussed in cluster 12 are related to neurological, paralytic and orthopedic concepts which are discussed in cluster 4, 9 and 11 respectively. Cluster 14 has only 1 document. It is identified that this document is related to 'metastasis'.

It is worth mentioning that it was found that cluster 3 contains all the documents in which the concepts of diseases tagged by UMLS MetaMap is 'stroke'. However, further analysis shows that most of the documents are not related to cardiological or neurological 'stroke'. This is also reflected on the U-matrix that cluster 3 is not close to cluster 1 and 2. It confirmed that the proposed document presentation and weight scheme based on the concept similarity measure can effectively differentiate documents based on the concepts of diseases. On the other side, it also shows that UMLS MetaMap might not accurately map all concepts to the corresponding phrases through the lexion.

7.4.2 Ohsumed Collection

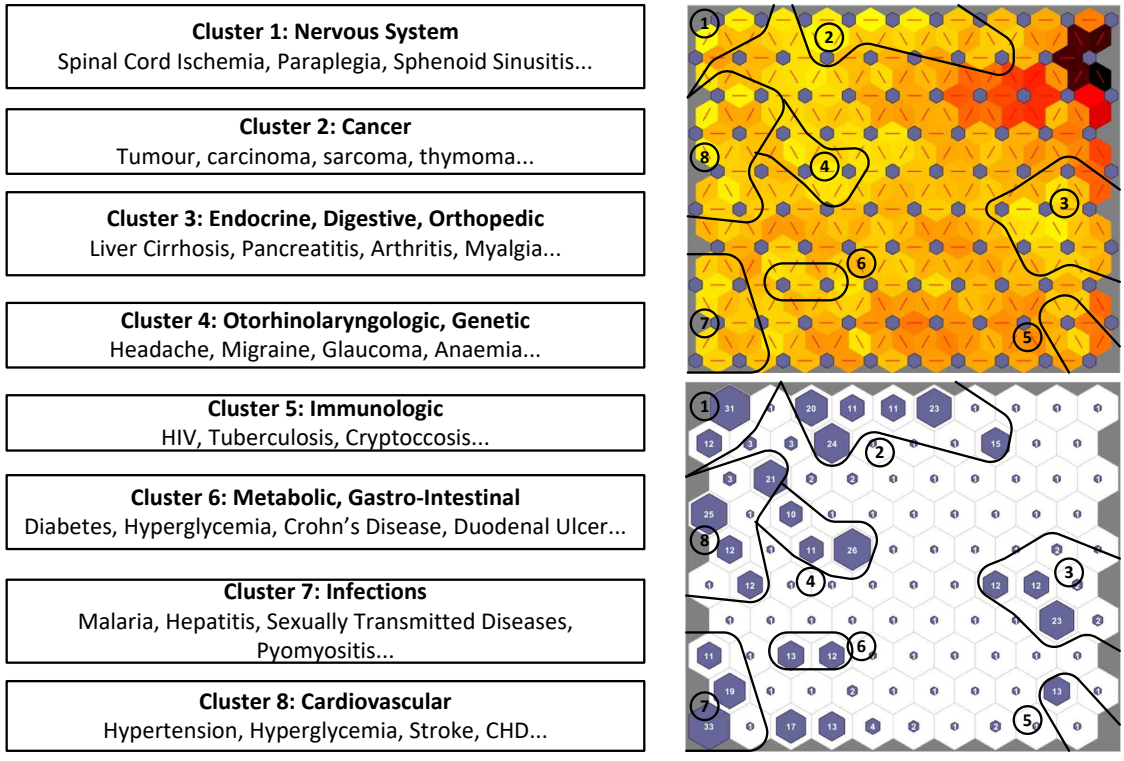

Fig. 7.4.: Clustering visualization for Ohsumed Collection using the document weighting scheme and clustered with self-organizing maps.

From the Ohsumed collection, a subset of ∼600 documents was used to evaluate the proposed document clustering weighting scheme. Based on the original data set description, all documents are related to cardiovascular diseases. This led to the shorter distances between neurons which is reflected by the color of the U-matrix. By analyzing the U-matrix and hit histogram of the trained map, 8 clusters are identified, as seen in Figure 7.4. These clusters and their contents are described in Figure 7.4.

A majority of the documents in cluster 7 are about infections and infectious diseases, with half of them from the categories of bacterial infections and mucoses, virus diseases and parasitic diseases. The rest of the documents from this cluster discuss other infections from categories like respiratory tract diseases and digestive system diseases. There are also a few documents from immunologic diseases in this cluster. Notably, all of the documents talk about infections of different types. Cluster 1 has documents that discuss diseases about the nervous system. Whereas, the documents in cluster 2 discuss neoplasms which include different types of cancers of the brain, prostate, neck and so on. The documents in this cluster are from all the categories except virus diseases and diseases of environmental origin. Cluster 3 includes documents about diseases related to hormone secretion and distribution. This cluster also includes diseases of the bones and blood, since these concepts are closely related. Cluster 4 contains documents with diseases about the ear, nose, throat, head and surrounding areas of the face. Cluster 5 is the smallest cluster and the documents concentrate on different types of tuberculosis and sexually transmitted diseases like AIDS, HPV, etc. Documents about 'cryptococcosis', which is often seen in patients with HIV whose immunity has been lowered, also fall in this cluster. Cluster 6 consists of documents with concepts relating to diabetes. Documents containing concepts like 'nephropathy', 'impaired glucose tolerance','non-insulin dependent diabetes' are in the left half of the cluster. Whereas, the right half of the cluster is dominated by documents with concepts such as 'Crohn's disease', 'renal ulceration', and 'kidney stone'. Cluster 8 has documents about diseases related to different heart conditions

and obstruction in the flow of blood. Since the theme of documents in Ohsumed collection is cardiovascular concepts, this cluster has documents from all of the categories except parasitic diseases, neoplasms and digestive system diseases.

7.4.3 TREC Genomics 2005

The self-organizing maps algorithm does not provide the number of clusters and cluster centroids, but rather identifies the distribution of the data. So, a centroid identification process is employed based on the distribution of the data. First, major visualizable clusters are identified initially through analyzing the U-matrix and hit histogram. Second, the evaluation results from the DB index and purity values are observed (Figures 7.5 and 7.6).

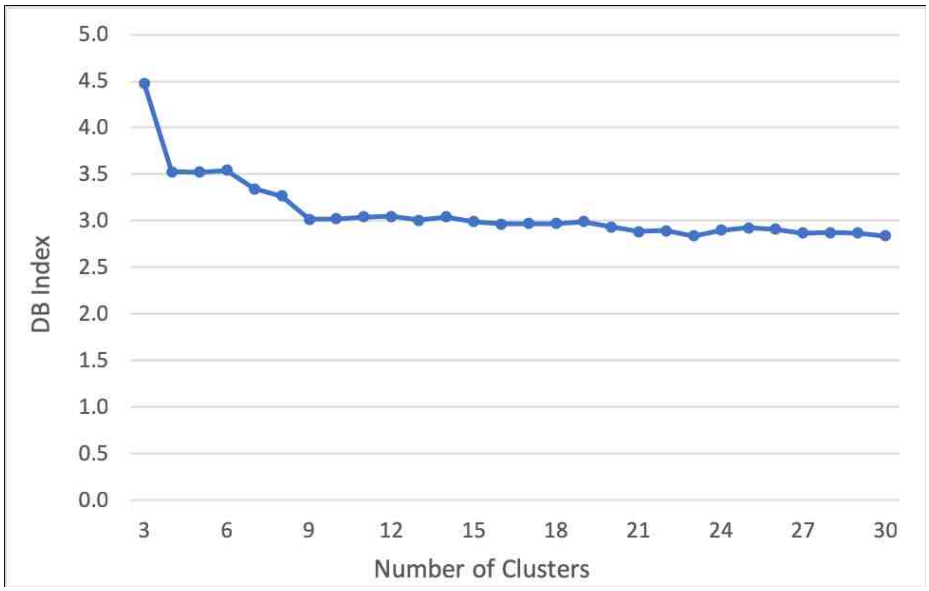

Fig. 7.5.: DB index evaluation for TREC Genomics 2005 corpus using the document weighting scheme and clustered with self-organizing maps.

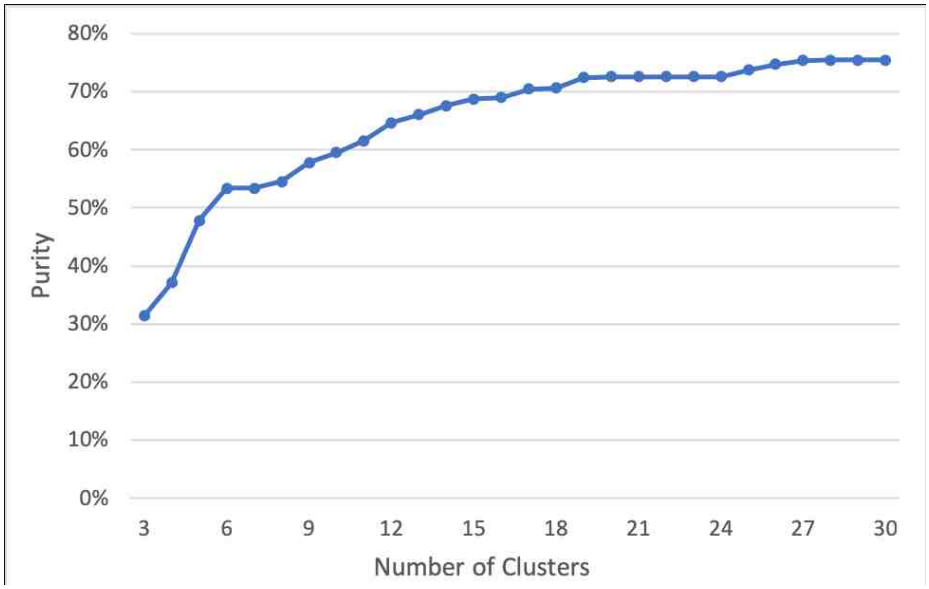

Fig. 7.6.: Purity evaluation for TREC Genomics 2005 corpus using the document weighting scheme and clustered with self-organizing maps.

The neuron of each cluster that has the highest hits is selected as the centroid. However, there are data hits to the neurons that are not included in the major visualizable clusters. In order to fully evaluate the best number of clustering partitions to cover all the input data instances, additional clusters are added one by one on top of the major visualized clusters to identify the optimum number of clusters.

Clustering with 29 clusters provides the best clustering performance based on the observations from DB index and purity, for visualization demonstration purposes, Figure 7.7 shows the large clusters from the 29 clusters in the SOM. The clusters and the corresponding centroids are marked. Further analysis has been done to understand the contents of the documents within those clusters.

Most of the documents in cluster 1 are of the disease concept 'multiple sclerosis'. Cluster 2 contains documents that discuss the concept 'multiple sclerosis relapsing remitting'. The proximity of this cluster to cluster 1 reflects from the proximity of this concept to 'multiple sclerosis'. Cluster 3 is a large cluster with most documents that relate to different types of cancer, with majority of them belonging to 'breast cancer',

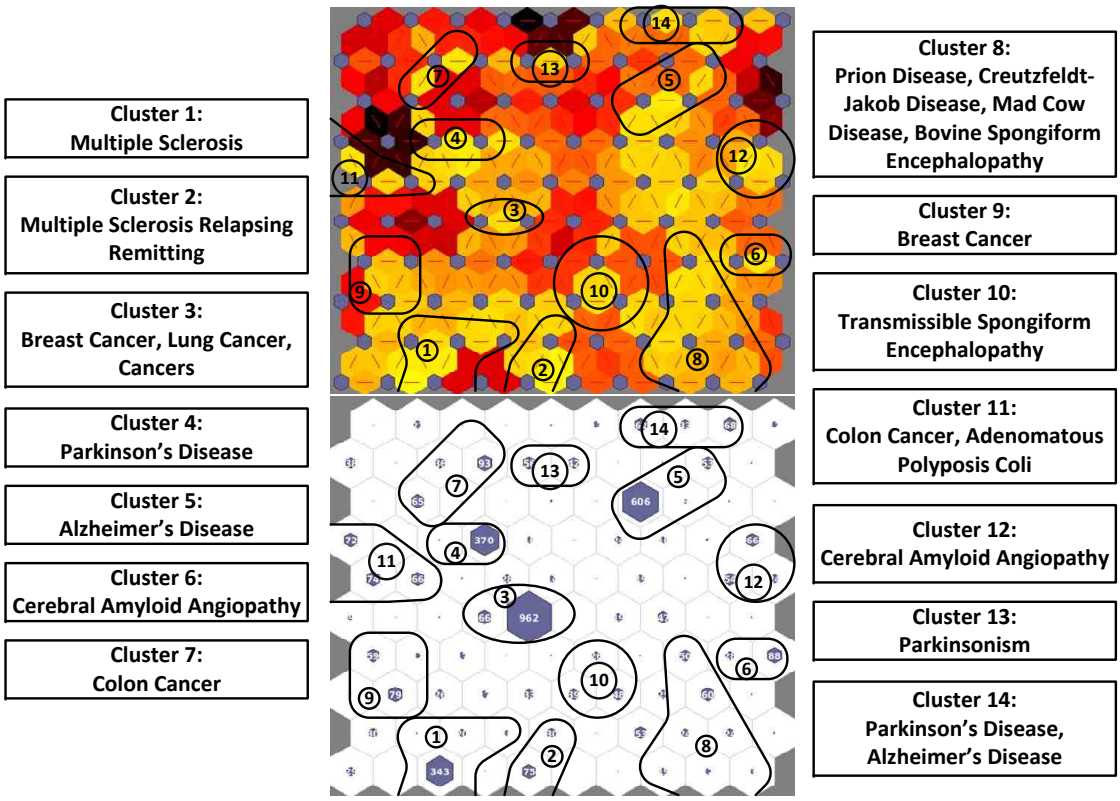

Fig. 7.7.: Clustering visualization for TREC Genomics 2005 corpus using the document weighting scheme and clustered with self-organizing maps.

'lung cancer' and other types of 'cancer'. The documents of this cluster are spread across all the categories. Cluster 9 contains documents about 'breast cancer' only. Documents which discuss 'Parkinson's disease' only are a part of cluster 4. Documents discuss 'parkinsonism' related disease belong to cluster 13, which is close to cluster 14. Documents in cluster 14 talk about both 'Parkinson's disease' and 'Alzheimer's disease'. Cluster 5 contains documents that relate to 'Alzheimer's disease'. This cluster is placed near cluster 14, showing the closeness of the concepts. Documents that hit cluster 12 talk about 'cerebral amyloid angiopathy'. Cluster 7 contains documents that belong to 'colon cancer' and 'colorectal cancer'. Documents discussing 'Adenomatous polyposis coli', (APC gene) as a cause of 'colorectal cancer' are a part of cluster 11. Cluster 8 contains various documents of the concept 'prion diseases'

and the various examples of this category of diseases like 'Creutzfeldt-Jakob disease', 'bovine spongiform encephalopathy' and 'mad cow disease'. This cluster is made up mostly of documents from the 'mad cow disease' category of the dataset. Documents in cluster 10 discuss 'transmissible spongiform encephalopathy', also known as 'prion diseases'. The proximity of these clusters is indicative of the similarities of the concepts discussed.

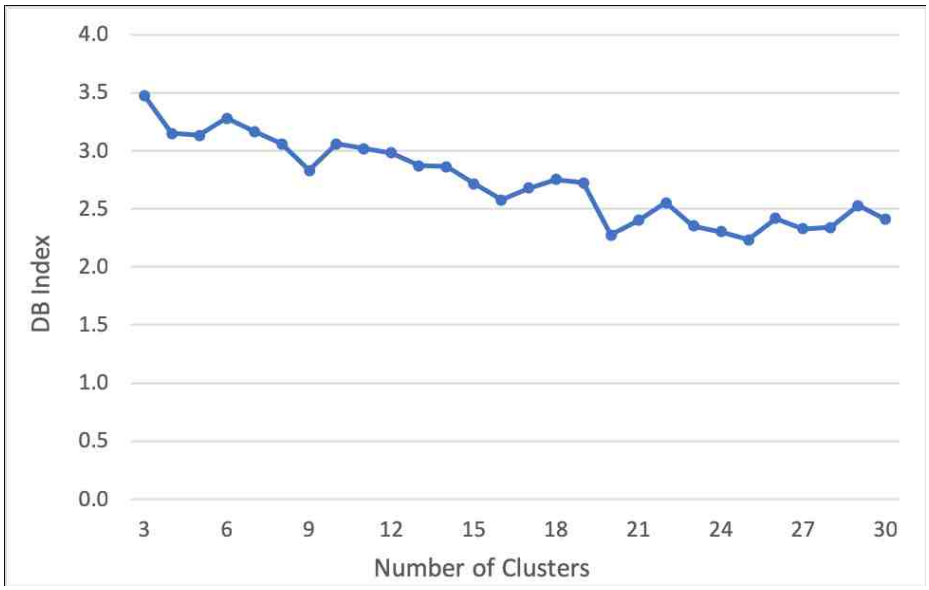

Fig. 7.8.: DB index evaluation for TREC Genomics 2005 corpus using the document weighting scheme and clustered with k-means clustering.

The k-means clustering algorithm provides the clusters and centroids of clusters by value k. There is no extra steps involved to identify them. However, selecting the best clustering results involves evaluating the DB index and purity values over the various cluster sizes (Figures 7.8 and 7.9.

Since $k = 25$ provides the best clustering results, Figure 7.10 shows the distributions of the 25 clusters given by k-means clustering algorithm and visualized through a scatterplot after applying t-SNE (Chapter 3, Section 3.3.3). The original dimensions

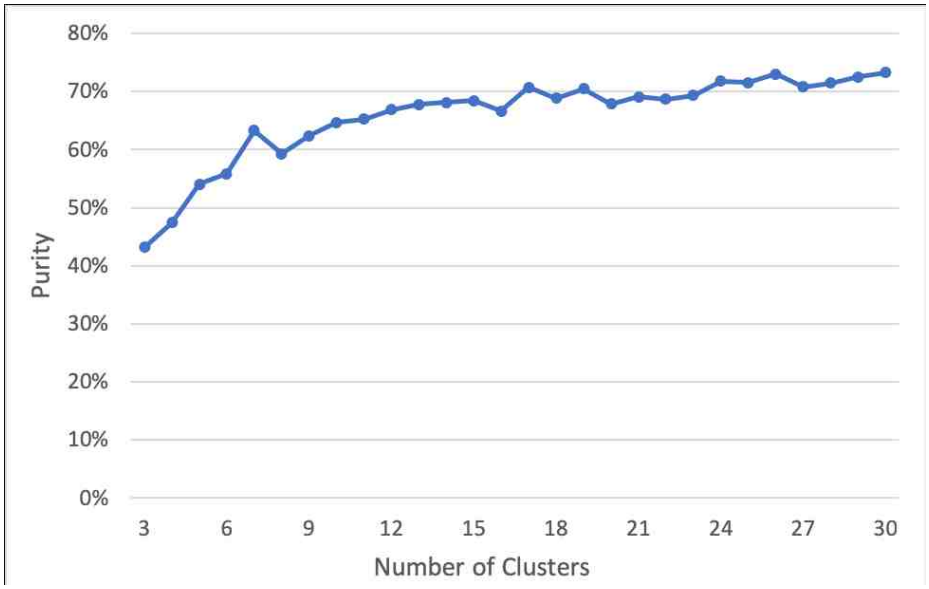

Fig. 7.9.: Purity evaluation for TREC Genomics 2005 corpus using the document weighting scheme and clustered with k-means clustering.

are reduced to 900 using PCA (Chapter 3, Section 3.3.2) before t-SNE is applied. Although some of the smaller clusters are overlapped by bigger clusters, most clusters that can be clearly visualized are shown in the figure.

Cluster 2 has about 800 documents and majority of them have 'Alzheimer's disease' discussed in the content. The documents in this cluster are mostly in cluster 5 of the SOM. Cluster 19 and 21 contain documents that discuss the 'Creutzfeldt-Jakob disease', which is a rare form of dementia related to 'Alzheimer's disease'. Documents having both 'Alzheimer's disease' and 'Parkinson's disease' as concepts are within cluster 17. Cluster 17 is very similar to cluster 14 on the SOM. Cluster 23 has very few overlapping with other clusters. It has documents about 'cerebral amyloid angiopathy' which is like the cluster 12 on the SOM. The documents in cluster 15 contain the concepts of 'Alzheimer's disease' and 'Amyloid'. Thus cluster 15 is located between cluster 17 and 23, as the concepts of cluster 15 are an intersection of the concepts of clusters 17 and 23. Cluster 22 contains documents that have 'parkinsonism' as the primary identified concept, whereas cluster 5 contains docu-

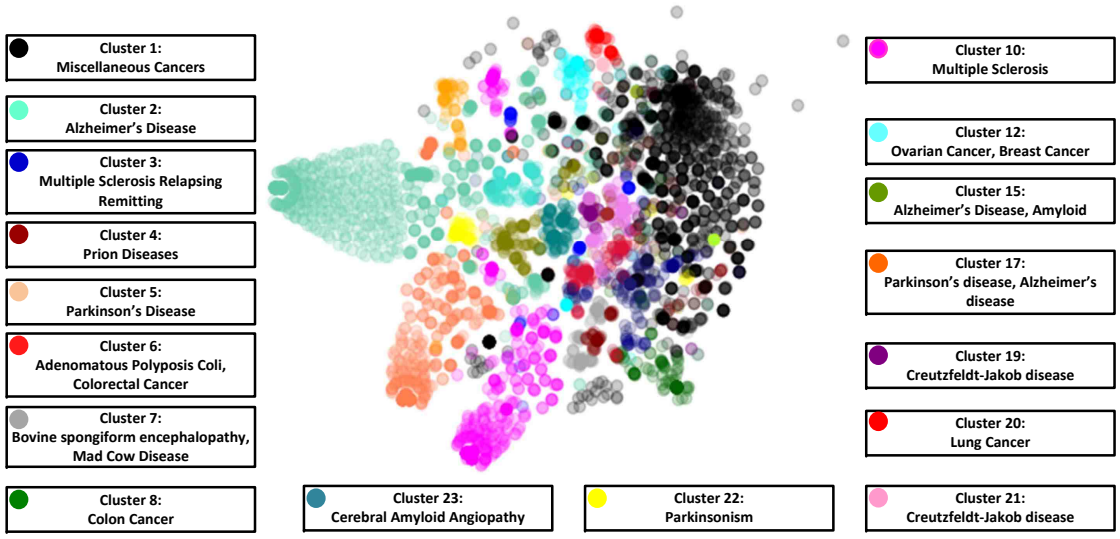

Fig. 7.10.: Clustering visualization for TREC Genomics 2005 corpus using the document weighting scheme and clustered with k-means clustering.

ments that have 'Parkinson's disease' discussed in the documents. There is a small overlap between these clusters which reflect the proximity of the concepts the documents describe. Cluster 10 contains documents that talk about 'multiple sclerosis', while cluster 3 consists of documents that discuss 'multiple sclerosis relapsing remitting'. Cluster 10 is similar to cluster 1 on the SOM, whereas cluster 3 is similar to cluster 2 on the SOM. Cluster 7 consists of documents that are related to 'Mad cow disease' and 'Bovine spongiform encephalopathy'. Since these concepts are synonyms of each other, this cluster is well formed. The concept 'prion diseases' belong to the category of disease that affect the brain and nervous system, and are discussed in cluster 4. The other concepts in this cluster include 'Creutzfeldt-Jakob disease' and 'Bovine spongiform encephalopathy'. This cluster includes documents that belong to the categories 'Alzheimer's disease' and 'Mad Cow Disease' of the original document set. This cluster is very similar to cluster 8 on the map of SOM. 'Colon cancer' documents are a part of cluster 8 which is similar to cluster 7 on the SOM. Documents describing 'Colorectal cancer' resulting from the mutations in the 'ade-

nomatous polyposis coli' (APC gene) form cluster 6 which is similar as cluster 11. Cluster 12 contains documents that discuss both 'breast cancer' and 'ovarian cancer'. Documents from the set that discuss 'lung cancer' form cluster 20. Various other documents that describe other forms of cancers and neoplasms form clusters 1 which is similar as cluster 3 of SOM.

In summary, the clustering and visualization of both clustering algorithms show similar results on clearly defined clusters. The documents that discuss the similar concepts are clustered in the same cluster. It can be said that the proposed document clustering and visualization framework works well on the representative corpora used in this research.

However, there is a limitation with this document weighting scheme. Equation 7.4 considers all of the concepts that are present in the document. Some documents may discuss unrelated concepts to draw comparisons, review work in other fields, and to provide an insight into other applications of work. This adds to the noise in the document vector, which further propagates down to the clustering results. It was seen that such unrelated concepts appeared within a lot of the incorrectly clustered documents, and to prevent the same the modified document weighting scheme is proposed.

7.5 Modified Document Weighting Scheme

In this weighting scheme, Equation 7.4 remains the same except one significant change: the similarity scores between concepts $s_{i,j}$ are now subject to a threshold τ [88] [89]. The modification implies that document vectors thus created now include the sum of similarity scores that are greater than a set threshold, instead of the sum of similarity scores of all of the concepts that appear in the document.

With this modification, the document weighting scheme changes to Equation 7.5.

$$dv_{i,d} = \begin{cases} tfidf_{c_i,d,D} + \sum_{k=1}^{K} s_{i,k} & tfidf_{c_i,d,D} \neq 0, s_{i,k} >= \tau \\ \sum_{k=1}^{K} \frac{K-(k-1)}{K} s_{i,k} & tfidf_{c_i,d,D} = 0, s_{i,k} >= \tau \end{cases} \tag{7.5}$$

Thus, if a concept c_i occurs in a document, the sum of the similarity scores between the concept c_i and other concepts (c_k, $k = 1, \ldots, K$) that also occur within the document, and are greater than the threshold τ is calculated. This sum is added to the value of $tfidf_{c_i,d,D}$ to compute the value of the term $dv_{i,d}$. By using this weighting scheme, the vector representation measures the co-occurrences and associations of concepts within each document, while ignoring concepts that do not appear in the document or are not similar to the concept c_i.

On the other hand, if a concept does not occur in a document, the weighting scheme calculates the weighted average of the association scores between the concept and all the concepts that appear in the document (c_k, $k = 1, \ldots, K$), that have a similarity value greater than the threshold τ. This measurement ensures that concepts that do not appear in the document, but are closely related to other concepts appearing in each document have a non-zero weight. By using such a weighted measurement for document representation, the importance of similar concepts is emphasized. The process of selection of threshold is discussed in Section 7.6.

7.6 Threshold Selection

The threshold (τ) plays an important role in determining which concepts should be considered as similar concepts, based on their similarity score. Using such a threshold ensures that only relevant and similar concepts are considered while generating the vector for each document. For selecting the threshold, experiment with different values of the threshold τ in Equation 7.5, for both the concept-based and word-based representations, are performed. The following subsections describe the

differences between the representations that cause the variances in the performance of the threshold across them, whereas an in-depth comparison of their clustering and evaluation performance is discussed in Section 7.7.

7.6.1 Word-based Representations

In the case of word-based representations, it is observed that the vector of the concept is dominated by frequent words in the corpus. Words like 'cancer' and 'disease' that appear very frequently across documents in the corpus. Thus, when concept vectors are generated from word-based representations (Chapter 6, Equation 6.4), these higher frequency words have a larger impact on the vector.

For this reason, concepts like 'colon cancer' and 'breast cancer' are similar to each other in terms of their similarity measurements, despite not being similar concepts. At the same time, because the word-based representation is generated by adding vectors of the individual words, their similarity values are higher than those for concept-based representations.

For tackling both of these problems, results produced by using various thresholds to generate the document vectors were tested. It was observed that using a higher threshold gave better results for clustering when using word-based representations, in all of the corpora, when compared to the concept-based representations.

7.6.2 Concept-based Representations

Concept-based representations have multiple representations of the same concept, and each of them are treated as individual concepts. For instance, 'lung cancer', 'cancer of the lung', 'malignant neoplasm of the lung' are the same biomedical concept, but are different concepts for the concept-based representation.

Because of this, during the training as well, each of these concepts are less frequent leading to lower individual vector values, and thus relatively lower similarity values among semantically similar concepts. Thus, a lower threshold value better accounts for concepts that are more closely related when using concept-based representations.

7.7 Results

Clustering of the documents is performed by using k-means clustering (Chapter 3, Section 3.1.2). The clustering performance of both the concept representations (Word-based ans Concept-based) is evaluated against the TF-IDF baseline. As discussed above in section 7.6, the threshold values play an important role in the generation of document vectors. Thus, both the representation methods for all values between 0.75 and 0.95 spaced at 0.05 are evaluated. For each corpus (PubMed Central – Open Access, Ohsumed Collection, TREC Genomics 2005), each configuration of concept vector calculation (word-based and concept-based) and threshold selection, clustering is run for k values ranging from 2 to 30.

Results produced are compared against a baseline where TF-IDF is the only algorithm used to generate document vectors. The baseline, too, is clustered for k values from 2 to 30. Since there is no change in the document vectors for TF-IDF baseline with changing threshold τ, changes in threshold value are not considered.

The internal evaluation metric DB index validates clustering by measuring the similarity between all of the clustering results (Chapter 3, Section 3.2.1). Lower values of DB index imply lower similarities between clusters, and thus better results of clustering. This helps us select the optimal values of clusters for the k-means clustering used. On the other hand, F-measure relies on the true positive, true negative, false positive and false negative results of clustering to give a score to the clustering (Chapter 3, Section 3.2.3). Higher the F-measure value, better the result of the clustering is.

7.7.1 PubMed

TF-IDF Baseline

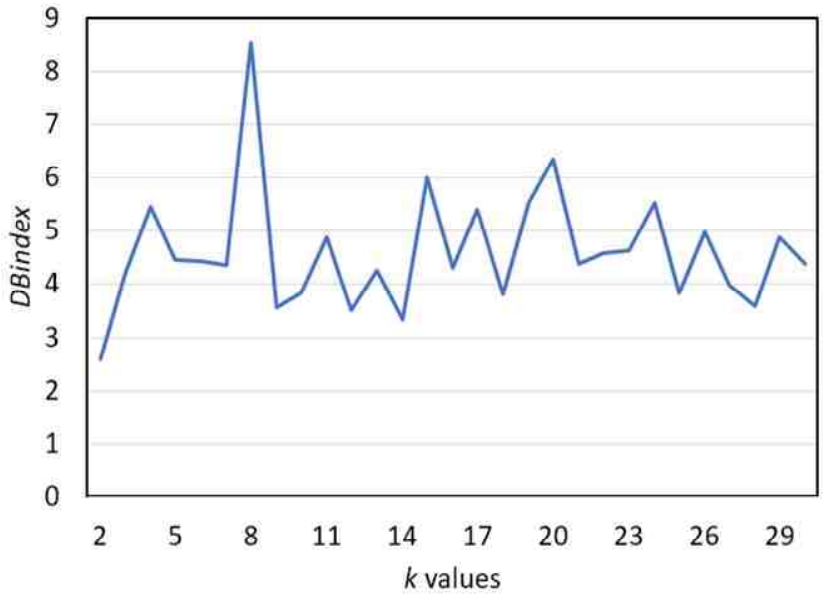

Fig. 7.11.: DB index values for TF-IDF baseline for the PubMed Central – Open Access corpus using the modified document weighting scheme and clustered with k-means clustering.

From Figure 7.11 it can be seen that the DB index value is lowest for 2 clusters. Low initial values do not indicate good clustering. The clustering for total clusters from 9-14 have much lower values than other cluster sizes. The lowest value is seen for cluster size 14. Based on these observations, evaluation of visualization of cluster size 14 is presented below.

The visualization of clustering with 14 clusters (Figure 7.12) shows that there is one major central cluster (black) with more than 3/4th of the documents. Most of the documents in this central cluster contain the concept 'stroke' and the other concepts within these documents are cardiac and neurological concepts. A lot of

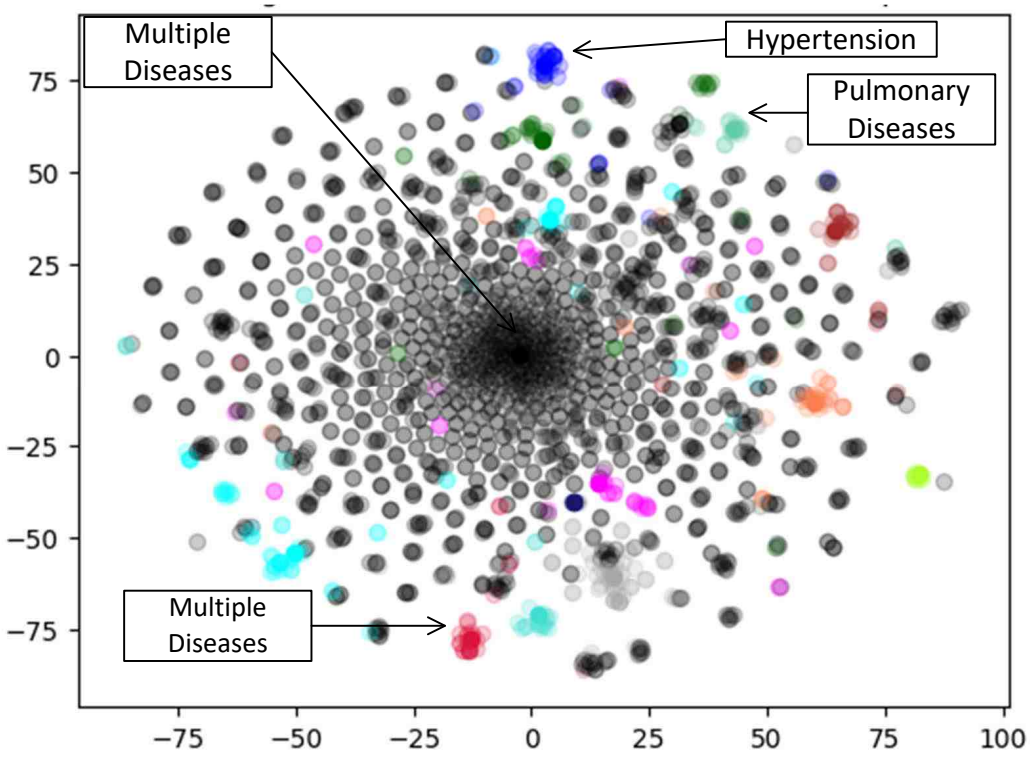

Fig. 7.12.: Clustering visualization for TF-IDF baseline for PubMed Central – Open Access corpus using the modified document weighting scheme and clustered with k-means clustering, with $k = 14$.

other documents with concepts that do not appear more than once in the dataset also belong to this cluster. This shows the need for increasing the number of clusters which can better distribute the large variety of concepts described within the dataset. Except the above described cluster, other clusters have documents covering specific diseases with the blue green cluster on the top right focused on pulmonary diseases and the blue cluster on top containing documents describing cardiovascular diseases and hypertension.

Leaving aside a few clusters, most other clusters are randomly filled with concepts that are not similar to each other or dependent on one another in any significant way. This visualization, which is the best based on DB index values, show that a TF-IDF approach to clustering documents is not very successful. The lack of inclusion of contextual information in clustering is a challenge produced by the large number of concepts, and exactly the kind of situation that the proposed weighting scheme proposed algorithm aims to tackle.

Word-Based Representations

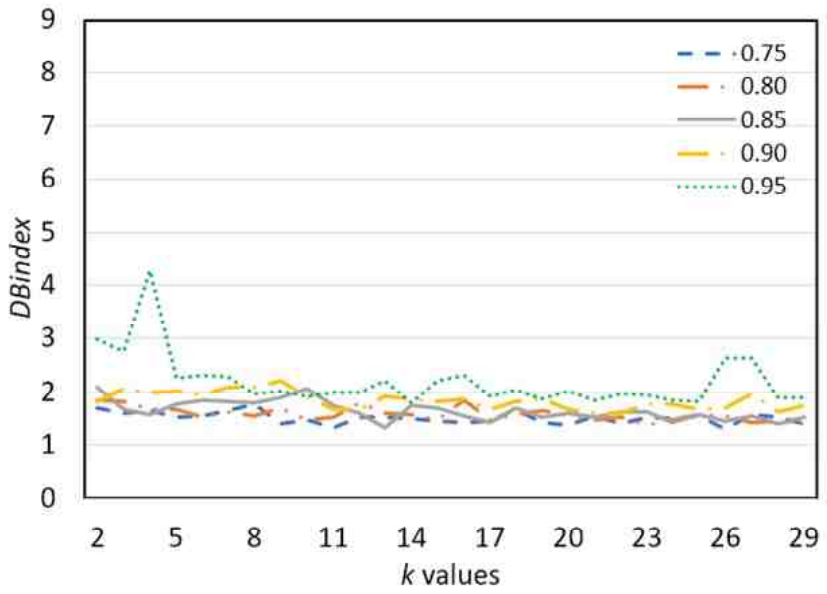

Fig. 7.13.: DB index values for the PubMed Central – Open Access corpus using the modified document weighting scheme that relies on word-based representations and clustered with k-means clustering.

Figure 7.13 shows the variation of DB index as total clusters are changed from 2 to 30, over various threshold values. It is clear from the figure that the threshold 0.95 performs poorly and does not produce very good clusters. The performance of threshold value 0.90 is slightly better than 0.95. At the same time, the performance of the thresholds 0.75, 0.80 and 0.85 are very similar, especially for the higher number of clusters. Based on the DB index graph, visualizations of 14 clusters for threshold 0.85 is chosen for further evaluation.

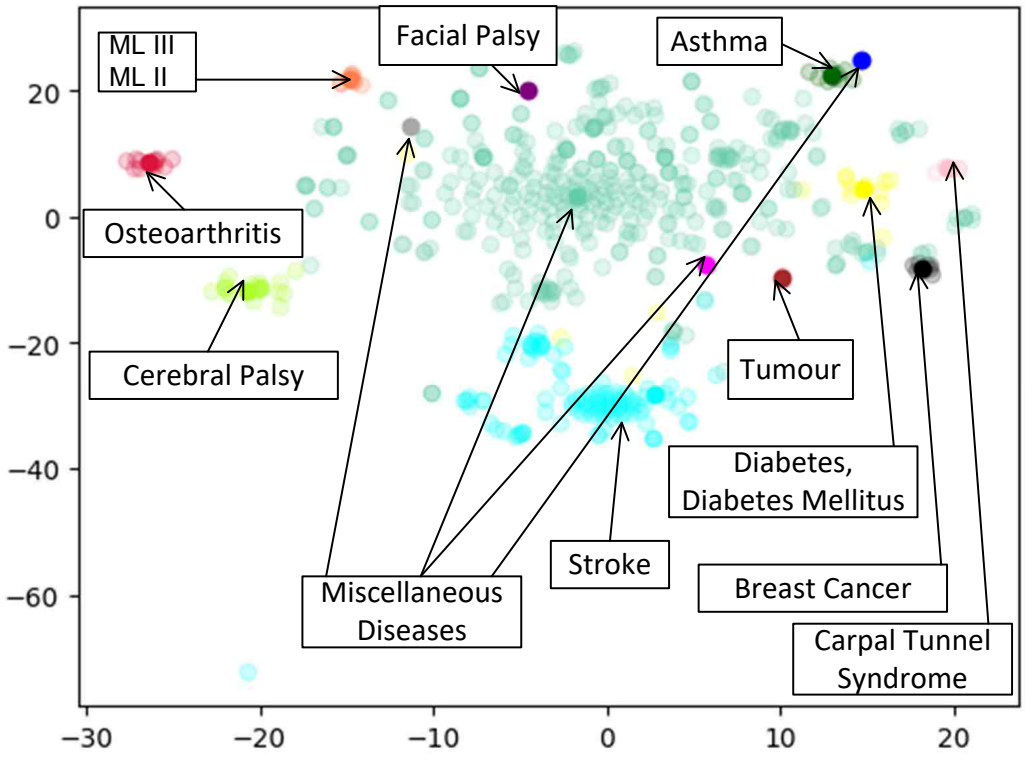

Fig. 7.14.: Clustering visualization for the PubMed Central – Open Access corpus using the modified document weighting scheme that relies on word-based representations and clustered with k-means clustering, with $k = 14$, $\tau = 0.85$.

It is clear from the visualizations shown in Figure 7.14 and its comparison with Figure 7.16 that the central cluster is much smaller for a similar number of total clusters. At the same time, individual clusters all around are more prominent and densely packed together. Further, the Figure 7.14 shows that all of the documents describing brain strokes, especially those with supplementary terms regarding the nature of the stroke (ischemic stroke, chronic stroke, stroke with hemiparesis), all belong to a single cluster (aqua). Similarly, all of the breast cancer-related documents belong to the black cluster, while the osteoarthritis documents belong to the maroon cluster on the top-left. Similar distinct clustering is observed for cerebral palsy (lime green), diabetes (yellow), asthma (dark green), etc. All these clusters also include some documents from other concepts that are either semantically equivalent or similar to the concepts mentioned above.

These observations establish the performance of the algorithm and its ability to identify relevant documents that describe a similar set of diseases and group them together into the same cluster, despite not having the same keywords or concept.

Concept-Based Representations

Figure 7.15 shows the plot of DB index for varying total clusters for the concept-based representation of words. It is visible from the figure that threshold 0.80 has consistently low values of DB index, while 0.75 and 0.90 have higher values and spikes. For evaluating the visualization, 14 clusters with threshold 0.85 is chosen.

One of the things that immediately stands out when looking at Figure 7.16, as compared to Figure 7.14, is the tightly bound clusters. One major observation Figure 7.16 is the number of small, well-defined clusters. Upon evaluation, it is found that most of these clusters contain documents with a very specific subset of concepts, all very closely related to one another. For instance, the orange cluster on the top right contains documents containing concepts like obesity, morbid obesity, obese and coronary artery disease, which is a major risk for people with obesity. Similarly, the

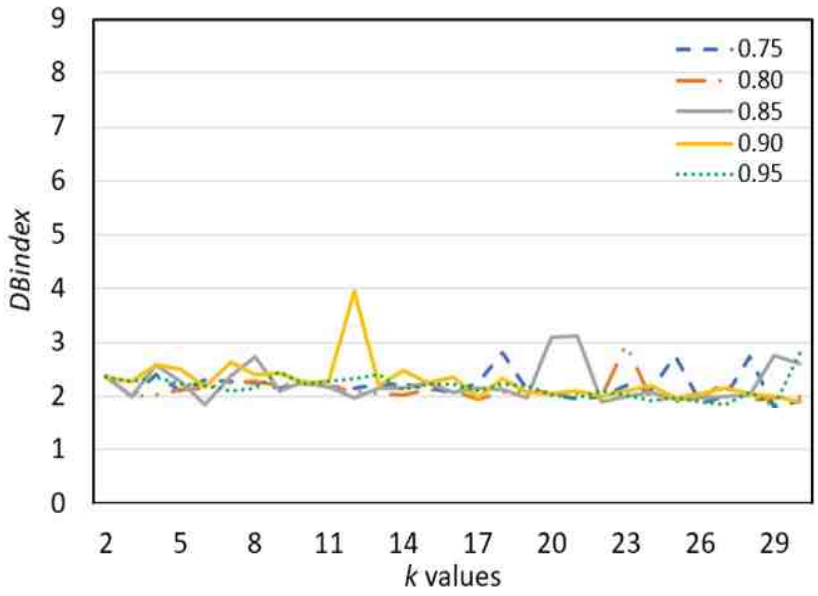

Fig. 7.15.: DB index values for the PubMed Central – Open Access corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering.

dark blue cluster on the top left contains documents describing hypertension, diabetes, hypoglycemia, heart blockage, etc. most of which are very closely related to one another. Pneumoperitoneum documents are in the green cluster on top left, chronic low back pain in the dark green cluster in the center, and myofascial pain syndrome documents in the lime green cluster on the top right. Stroke related documents form the aqua blue cluster in the center, whereas the fuchsia pink cluster close to it contains documents pertaining to ischemic stroke. This trend is seen across all the other small clusters. This proves that a threshold value on the lower side produces better results for the concept-based representations.

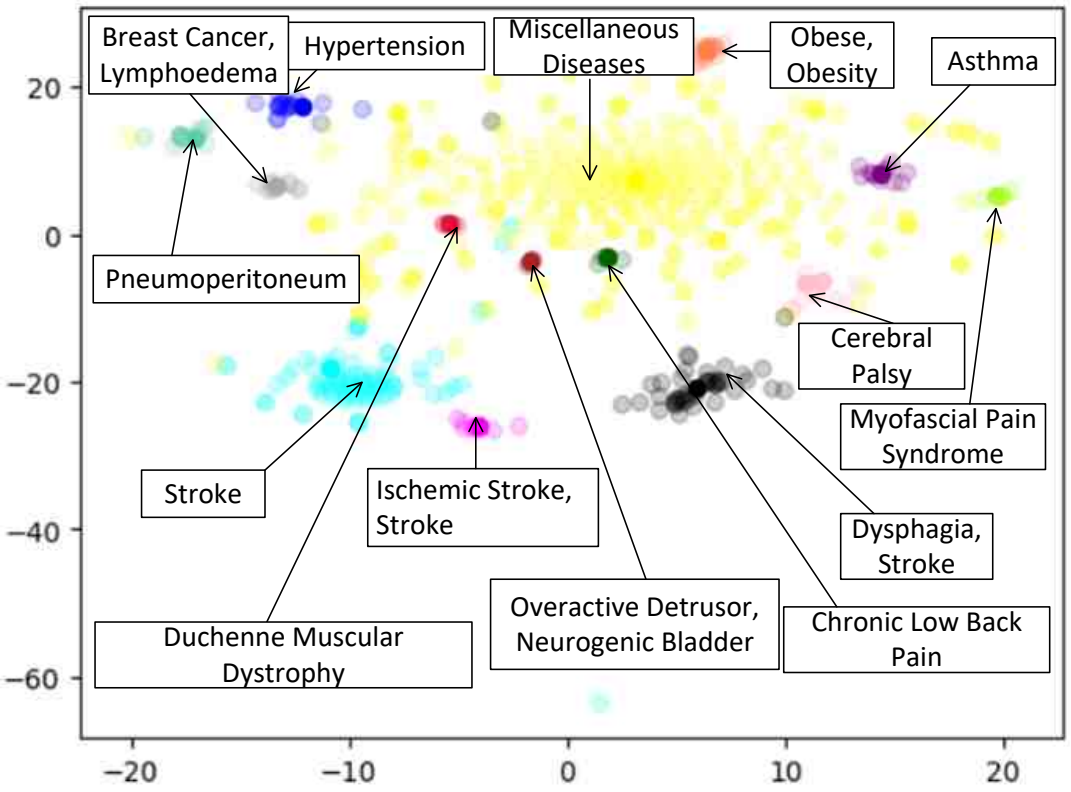

Fig. 7.16.: Clustering visualization for the PubMed Central – Open Access corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering, with $k = 14$, $\tau = 0.85$.

7.7.2 Ohsumed

TF-IDF Baseline

Figure 7.17 shows the variation of DB index as the number of clusters increases, when the document vector generated consists of only the TF-IDF values. Despite the consistently high values of DB index, indicative of poor clustering, this result shows that the cluster size of 9 has a lower DB index values as compared to all of the other clusters tested.

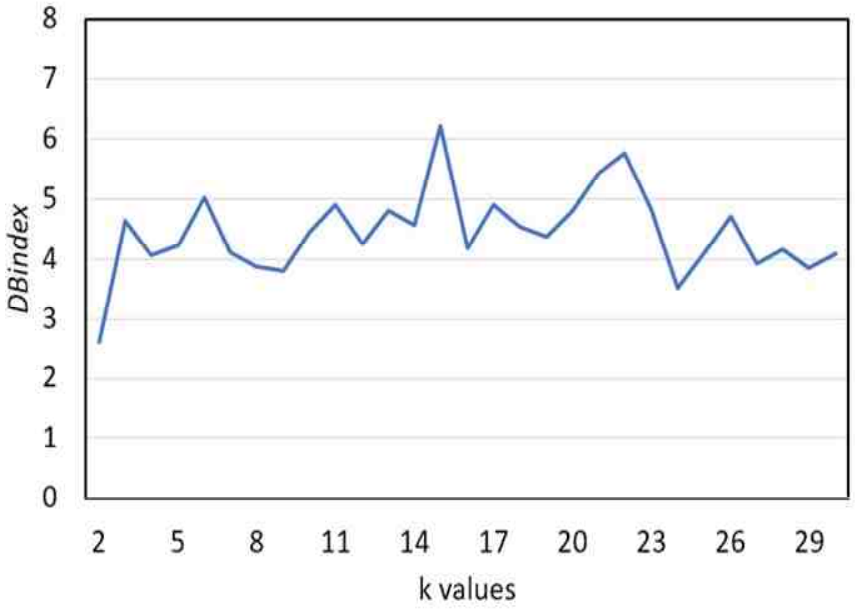

Fig. 7.17.: DB index values for TF-IDF baseline for the Ohsumed Collection corpus using the modified document weighting scheme and clustered with k-means clustering.

Figure 7.18 shows the visualizations for the cluster size of 9. The presence of a large central cluster with a majority of documents shows that neither of them are optimal cluster sizes with most of these documents being unrelated and grouped together because they do not fit in any other cluster. Among the other clusters for Figure 7.18, malignancy (black), recurrence (blue green on the bottom left), infection (maroon), etc. are the most frequently occurring concepts, with very few similar or semantically equivalent concepts in each cluster. Most clusters show the same lack of disparateness of concepts within clusters. At the same time, these clusters only represent a subset of documents describing the same / similar concepts. Other documents from these clusters are in other mixed clusters.

This shows that TF-IDF alone is not a good method for generating document vectors, especially for large, unlabeled datasets which have a variety of concepts.

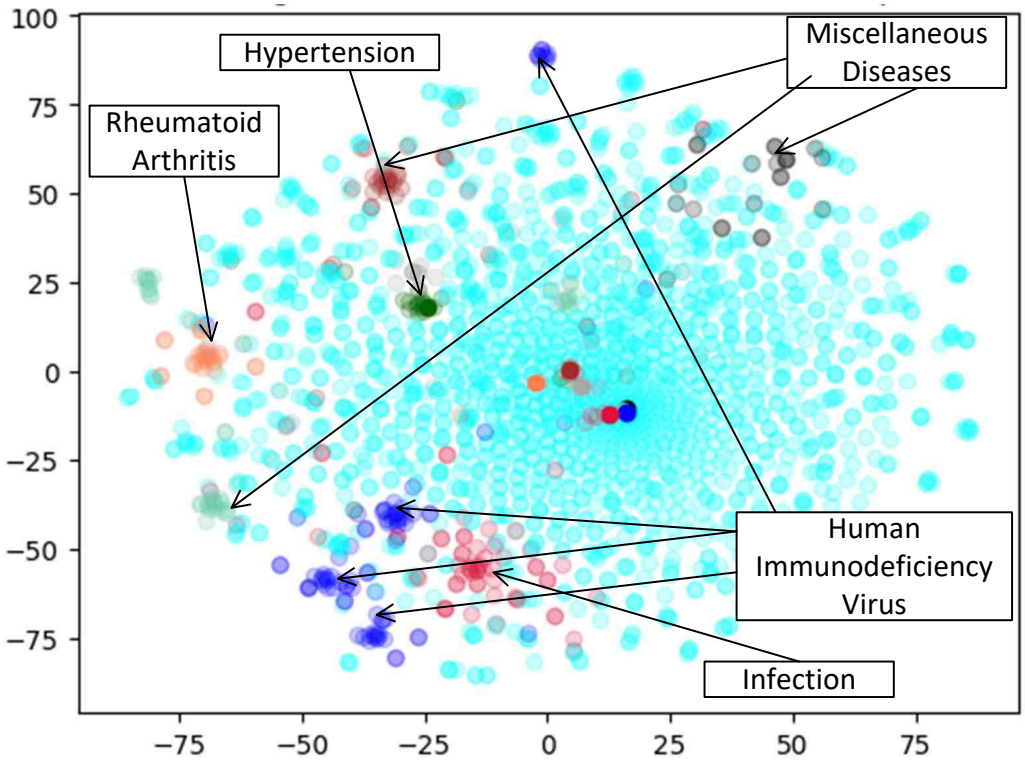

Fig. 7.18.: Clustering visualization for TF-IDF baseline for Ohsumed Collection corpus using the modified document weighting scheme and clustered with k-means clustering, with $k = 9$.

Word-Based Representations

From Figure 7.19 it is evident that threshold 0.95 is not a very good threshold. In fact, the range in which the DB index of 0.95 varies is very close to that of TF-IDF (Figure 7.19). While all the other thresholds have DB index in the same range, 0.85 outperforms all of them as the cluster number increases, while 0.80 produces better results with fewer clusters. Based on these observations, an evaluation of 10 clusters for threshold 0.80 is presented.

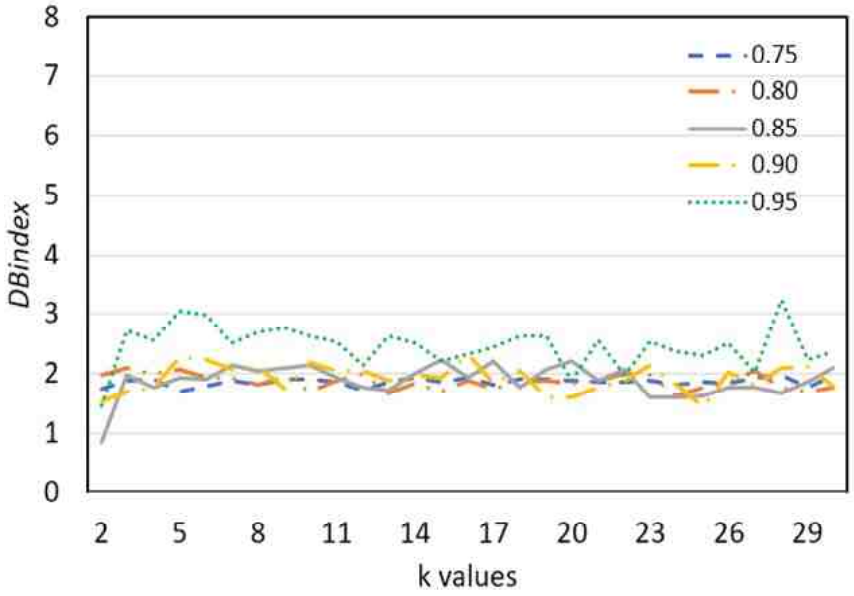

Fig. 7.19.: DB index values for the Ohsumed Collection corpus using the modified document weighting scheme that relies on word-based representations and clustered with k-means clustering.

The visualizations shown in Figure 7.20 show the trend seen with word-based representation earlier with the tightly bound clusters. It is observed that most of the documents in the dark blue cluster describe infection concepts like bacterial and sexually transmitted diseases. On the other hand, most documents about 'myocardial infarction' (blue green). There is a distribution of the cancer concepts and an overlap is seen across clusters, which can also be seen through their visualization without a clear boundary in the clusters in the bottom right. The grey cluster on the top left contains documents with concepts related to heart failure. Many clusters including those colored in brown, black, light green are labeled as 'Miscellaneous Diseases', since the document frequencies of the concepts for those clusters show that none of them contains documents related to a specific category disease. While these clusters aren't ideally separated, they are better than the TF-IDF approach and a lot of similar concepts are closely placed to one another.

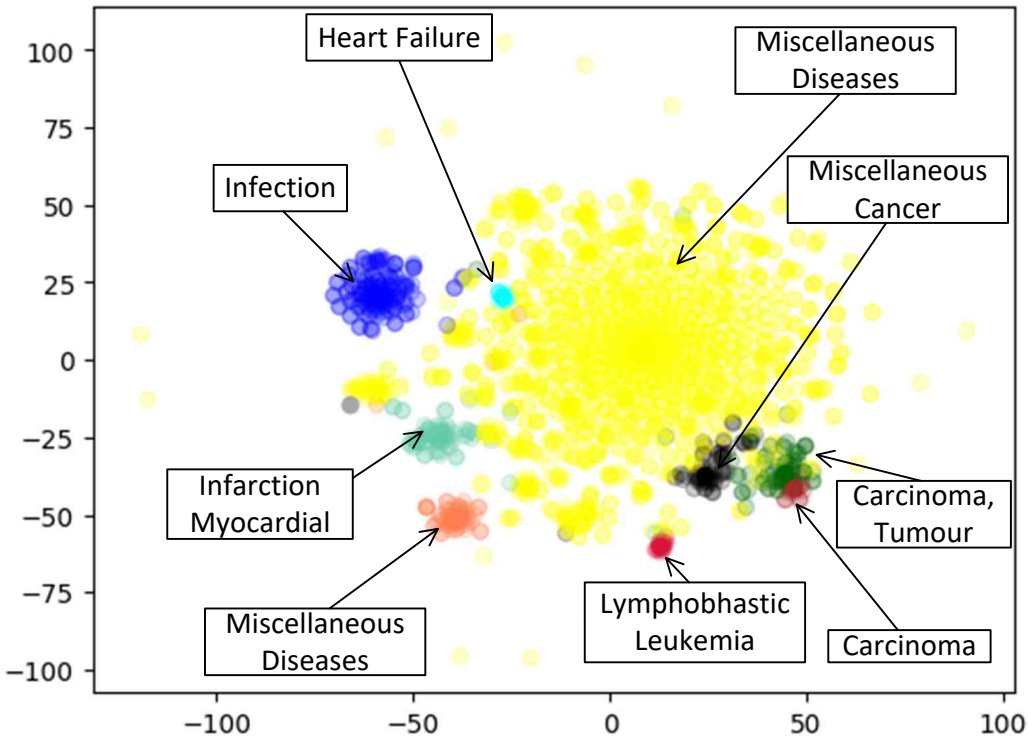

Fig. 7.20.: Clustering visualization for the Ohsumed Collection corpus using the modified document weighting scheme that relies on word-based representations and clustered with k-means clustering, with $k = 10$, $\tau = 0.80$.

Concept-Based Representations

For concept-based representations, the DB index chart shows that all threshold values perform fairly well with results very close to each other. However, at the elbows around 11 clusters, threshold 0.75 performs significantly better than any of the other threshold values. One noteworthy observation about the visualizations for concept-based representations (Figure 7.22) is about the density and tight packing of all of the clusters within the same region. This is in stark contrast to that of word-based representations (Figure 7.20), but follows the observations from Section 7.7.1.

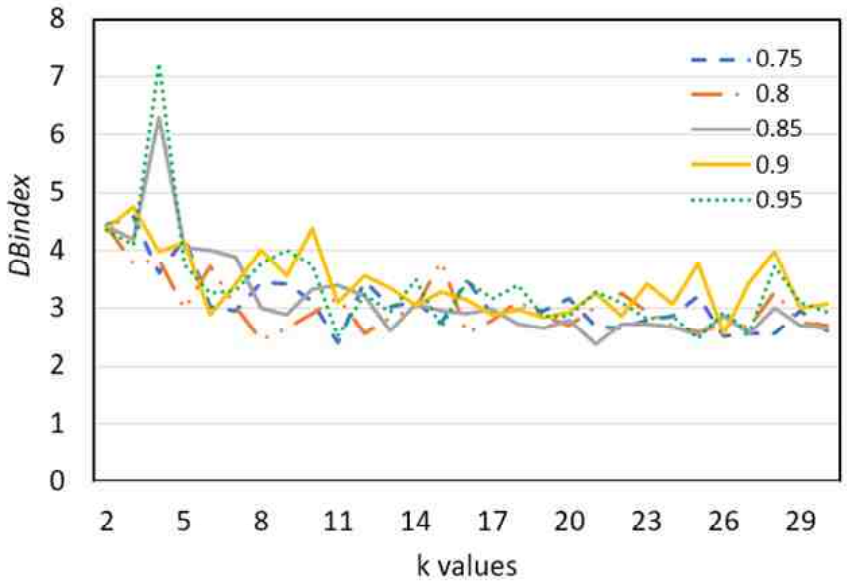

Fig. 7.21.: DB index values for the Ohsumed Collection corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering.

In the visualization of 11 clusters (Figure 7.22), the aqua blue cluster on the top right consists of documents talking about cardiac diseases, while the black cluster consists of a variety of unrelated concepts. The blue green cluster on the top right contains documents describing different types of diabetes, obesity and hypertension. These concepts often co-occur and thus are well related. At the same time, the proximity of this cluster with that of the cardiac diseases cluster (aqua blue on the top right) shows how there is some connection between these concepts as well. Lung cancer (dark blue), infections (brown), HIV (red and grey) form other major clusters. But these clusters also have a few other unrelated documents and concepts.

The similarity of concepts within the same cluster also shows that a lower threshold does not deteriorate the performance of clustering for concept-based representation, but in fact improves it.

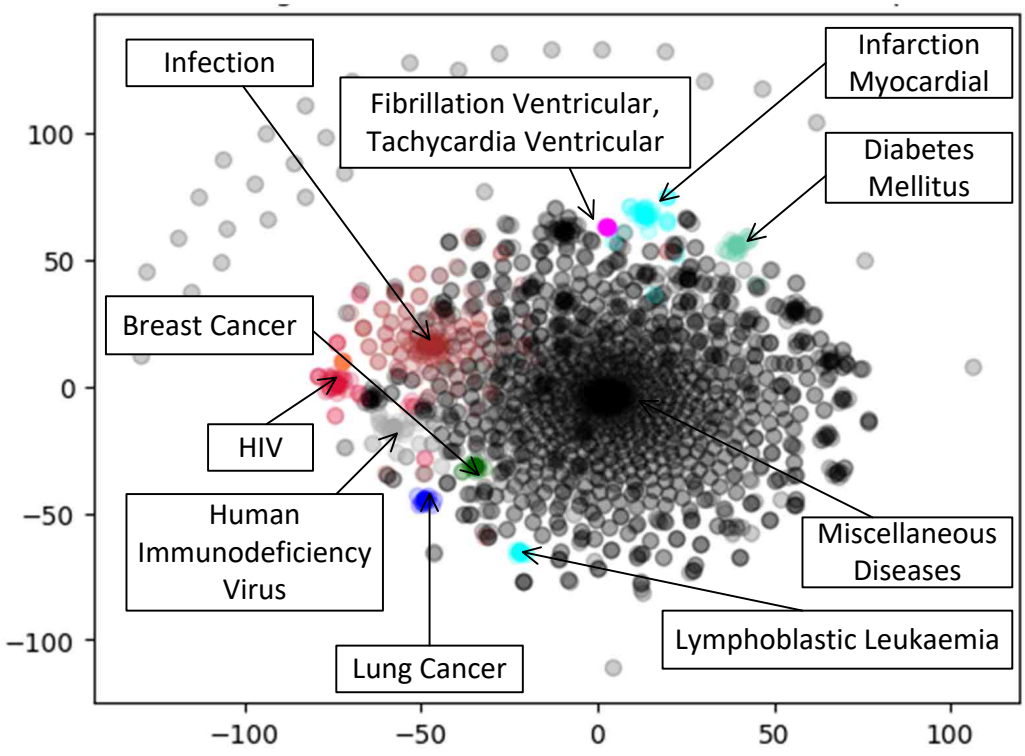

Fig. 7.22.: Clustering visualization for the Ohsumed Collection corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering, with $k = 11$, $\tau = 0.75$.

7.7.3 TREC Genomics 2005

TF-IDF Baseline

The results obtained for this dataset help us understand the correlation between the internal and external evaluation metrics, since this is the only labeled dataset. Looking at the TF-IDF results obtained for various cluster sizes, it is seen (Figures 7.23 and 7.24) that clusters 7, 17 and 30 have better values of DB index and F-measure than the rest of the cluster sizes. Clustering with 7 clusters is particularly

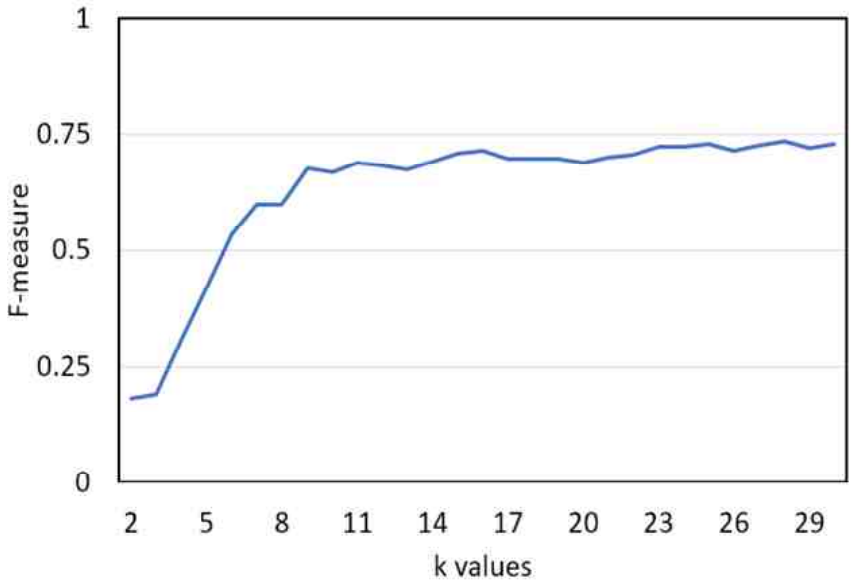

Fig. 7.23.: F-measure values for TF-IDF baseline for the TREC Genomics 2005 corpus using the modified document weighting scheme and clustered with k-means clustering.

important because that is the number of categories in our dataset. There is an elbow in the graph at 7 clusters, which shows the trend of lower values of DB index and higher values of F-Measure, starts after 7 clusters.

Visualizing the clustering of these values it is observed that for 7 clusters (Figure 7.25), Alzheimer's disease (teal), colon cancer (black), mad cow disease (blue), Parkinson's disease (maroon), multiple sclerosis (orange) and breast cancer (red) are the primary concepts for each of the clusters. The light blue cluster at the center contains documents with multiple concepts, most of which appear few times in the corpus. It is also seen that all of the documents belonging to the cerebral amyloid angiopathy category are in the light blue cluster. This shows that while 7 clusters separate out almost all of the primary categories into their individual clusters, but the clustering can be improved further.

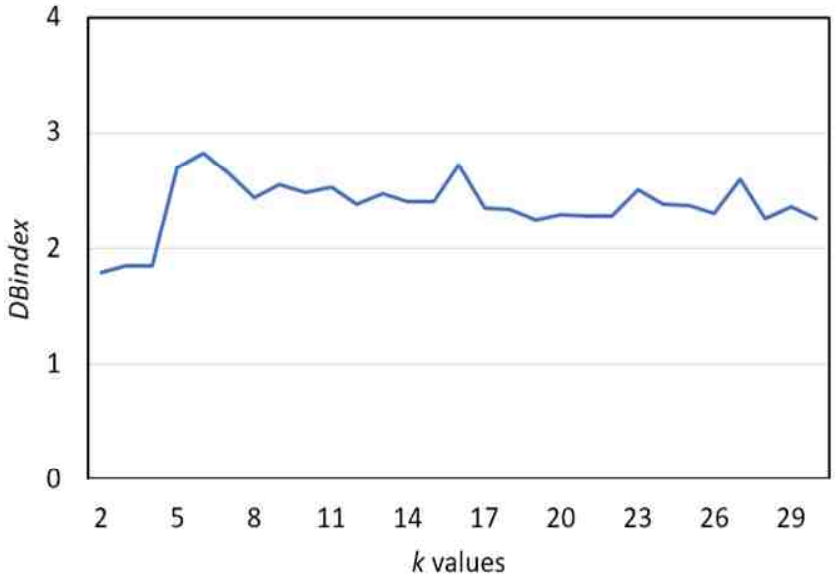

Fig. 7.24.: DB index values for TF-IDF baseline for the TREC Genomics 2005 corpus using the modified document weighting scheme and clustered with k-means clustering.

Word-Based Representations

From Figures 7.26 and 7.27, it can be seen that thresholds 0.80, 0.85 and 0.95 have the highest F measure values. At the same time, threshold 0.80 and 0.85 have consistently lower DB index values than 0.95 (Figure 7.27). At the same time, for lower values of clusters and at the elbow of 7 clusters, 0.80 has better DB index value for a similar F-measure score. This shows that 0.85 is a better threshold for generating the vectors if using higher cluster sizes, and 0.80 is better for lower cluster sizes.

It can also be seen that 0.75 has the best results if only DB index is looked at. However, F measure shows that the clustering results for that threshold are very poor. On further analysis, it is seen that lower thresholds increases the number of similar concepts calculated for each concept and thus even concepts that are not very closely semantically equivalent are added together. Similarly, if only F-measure is looked at,

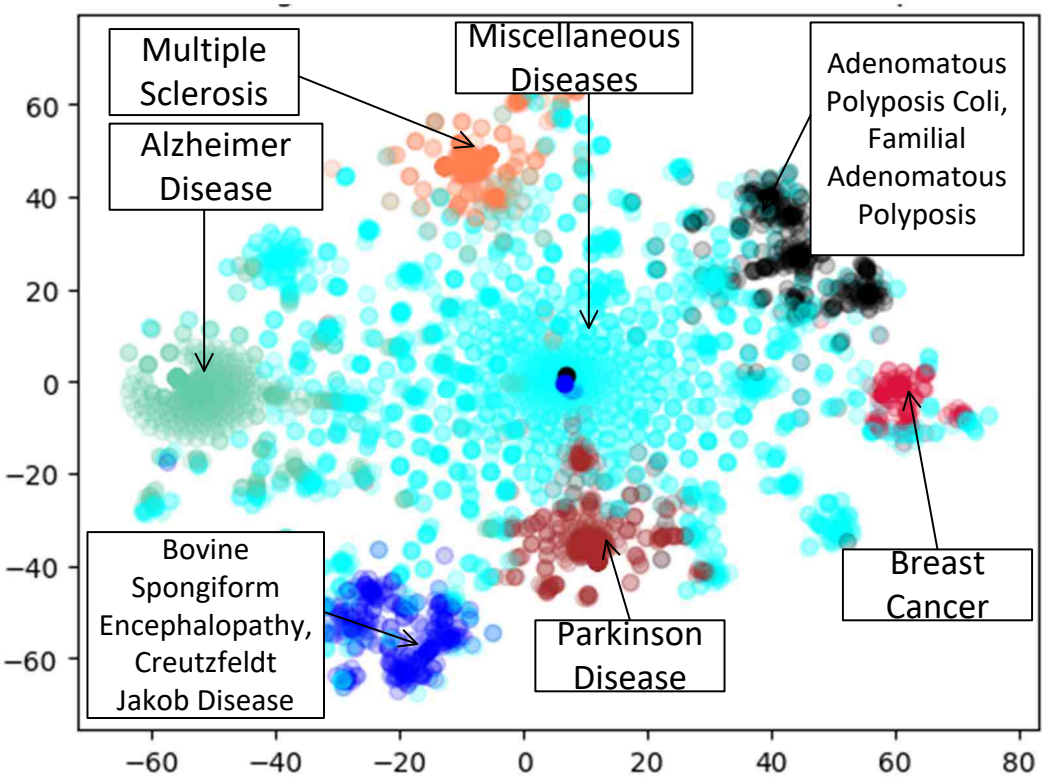

Fig. 7.25.: Clustering visualization for TF-IDF baseline for TREC Genomics 2005 corpus using the modified document weighting scheme and clustered with k-means clustering, with $k = 7$.

it is easy to assume 0.95 performs the best, when it actually has a much higher DB index value. Based on this observation, the visualization of clustering with 7 clusters, for a threshold value of 0.80 is explored further (Figure 7.28).

On comparing the differences between the word-based representation visualization of 7 clusters (Figure 7.28), with the TF-IDF clustering visualization for 7 clusters (Figure 7.25), it is apparent that the clusters for the word-based representation are much more defined and well separated. That is a trend seen in all of the word-based representations that the clusters are clear and distinct, with few overlaps. As the

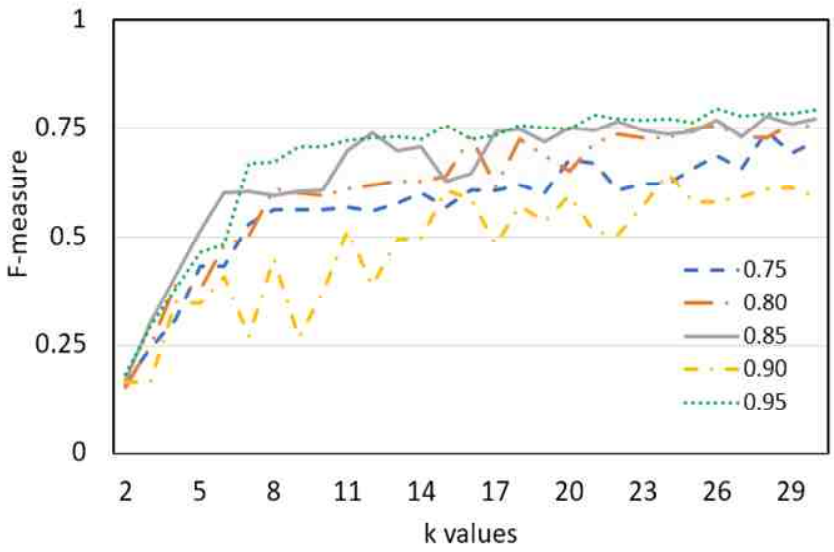

Fig. 7.26.: F-measure values for the TREC Genomics 2005 corpus using the modified document weighting scheme that relies on word-based representations and clustered with k-means clustering.

cluster number is increased, it is observed that a few clusters that do have overlapping boundaries share the similar or same concepts and are sub-clusters of the same concept, thus proving that the algorithm clusters and identifies similar concepts.

In Figure 7.28, each of the 7 clusters is a cluster of a different disease. Alzheimer's disease (cyan), Parkinson's disease (black), multiple sclerosis (blue green), breast cancer (blue), Mad Cow disease (orange), colon cancer (brown and red) describe all of the clusters. Among the categories covered by the dataset (Table 5.3), cerebral amyloid angiopathy (CAA) is the only category that does not have its own cluster. Documents covering CAA are distributed across the clusters of Alzheimer's disease, Parkinson's disease and Mad Cow disease. All these other clusters represent disorders of the brain, just like CAA. This emphasizes the performance of the algorithm as it clusters documents with similar concepts together.

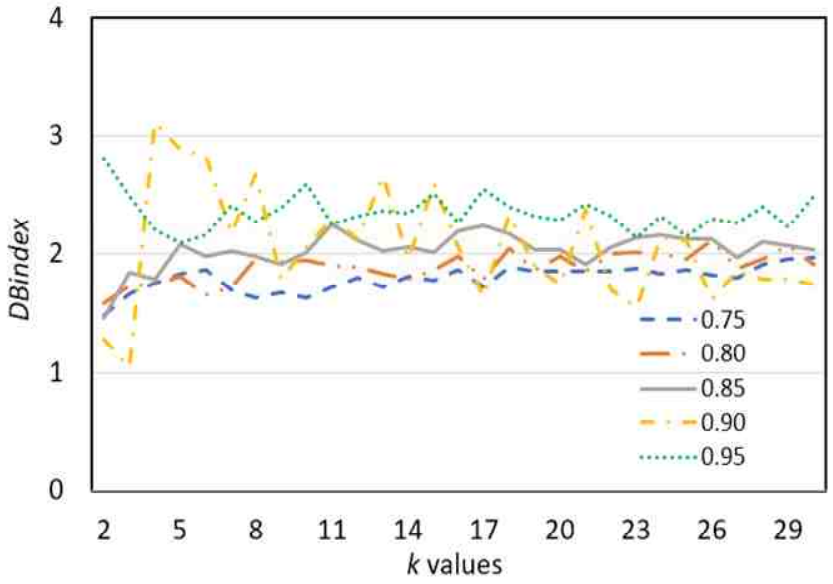

Fig. 7.27.: DB index values for the TREC Genomics 2005 corpus using the modified document weighting scheme that relies on word-based representations and clustered with k-means clustering.

Concept-Based Representations

With concept-based representations, it is seen from Figures 7.29 and 7.30 that as the number of clusters increases, the F-measure for all of the thresholds is very similar and plateaus. A similar trend is seen for DB-index, but the DB-index values for 0.75 are generally lower than that of the rest of the thresholds. However, at the elbow point of 7 clusters, the F-measure of threshold 0.90 is much higher than others, and its DB index is similar to 0.75. Based on these observations, the evaluation of visualizations of 7 clusters with 0.90 threshold is presented below.

While 7 clusters showed the elbow for the word-based and TF-IDF baseline, in the case of concept-based representations, 8 clusters show an elbow for thresholds ≤ 0.85, while those > 0.85 have the elbow at 7 clusters. This is because for a higher threshold for concept-based representations, there are very few concepts with a low

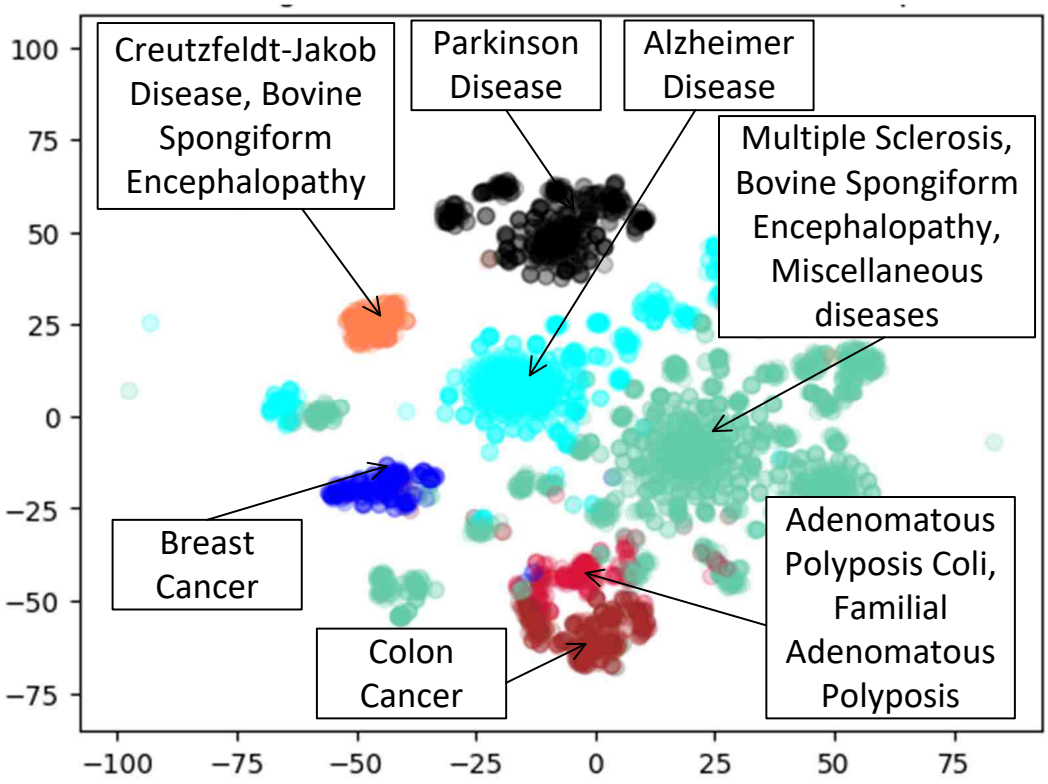

Fig. 7.28.: Clustering visualization for the TREC Genomics 2005 corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering, with $k = 7$, $\tau = 0.80$.

similarity value to other concepts and the document matrix generated looks largely similar to the one generated by the TF-IDF baseline. There also aren't many concepts very similar to each other (similarity score ≥ 0.85), which is also why these values have very similar DB-index and F-measure values to each other. As the threshold is lowered, the semantically equivalent concepts and similar concepts are identified and they appear in the document vectors.

From Figure 7.31, it can be seen that the concept-based representations are much more densely packed to one another than Figure 7.28, while also having lesser overlaps than Figure 7.25. This is characteristic of the visualization of concept-based

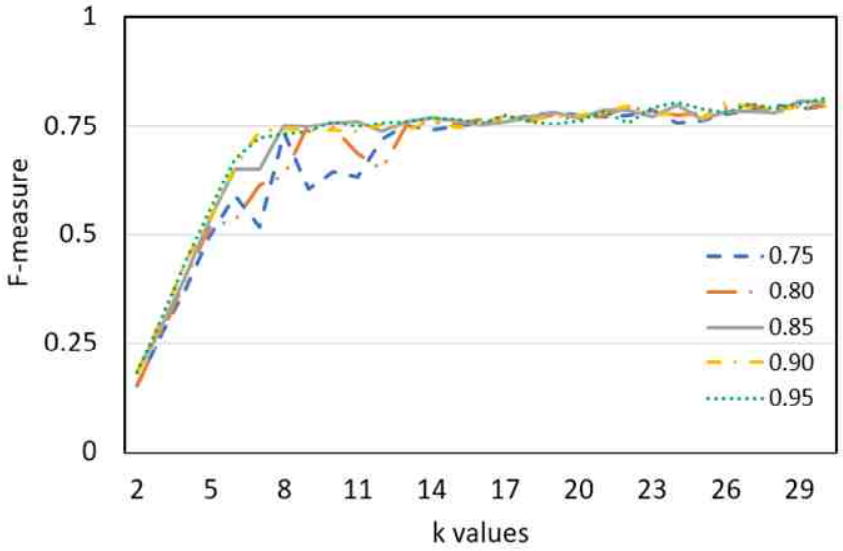

Fig. 7.29.: F-measure values for the TREC Genomics 2005 corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering.

representations across various threshold values. All of the major concepts except CAA have their own clusters, but the black cluster in the center is a collection of a lot of documents across categories. Multiple sclerosis is in the maroon cluster and breast cancer in the red cluster. The orange cluster contains documents from colon cancer, colorectal cancer and the adenomatous polyposis coli gene that causes colon cancer. The proximity between the clusters of Alzheimer's disease (dark blue) and Parkinson's disease (blue green) is also because of their high similarity in the vector space. There are also documents that contain both the diseases, and such documents are a part of the large black clusters close to them.

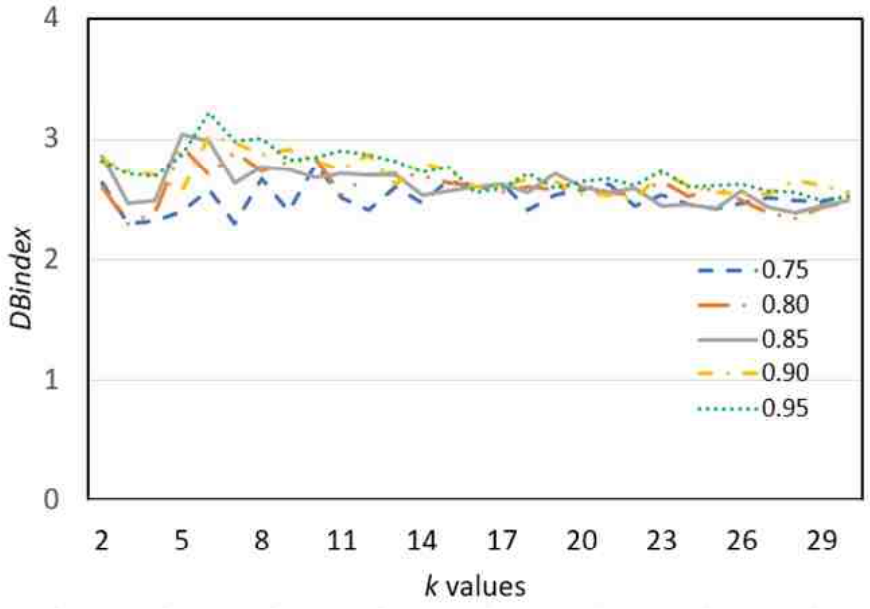

Fig. 7.30.: DB index values for the TREC Genomics 2005 corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering.

As the cluster size increases, all of the categories get divided into further sub-categories. Each of the sub-categories concentrate on a specific concept or a combination of concepts, and the cluster is close to other clusters that contain documents with similar concepts.

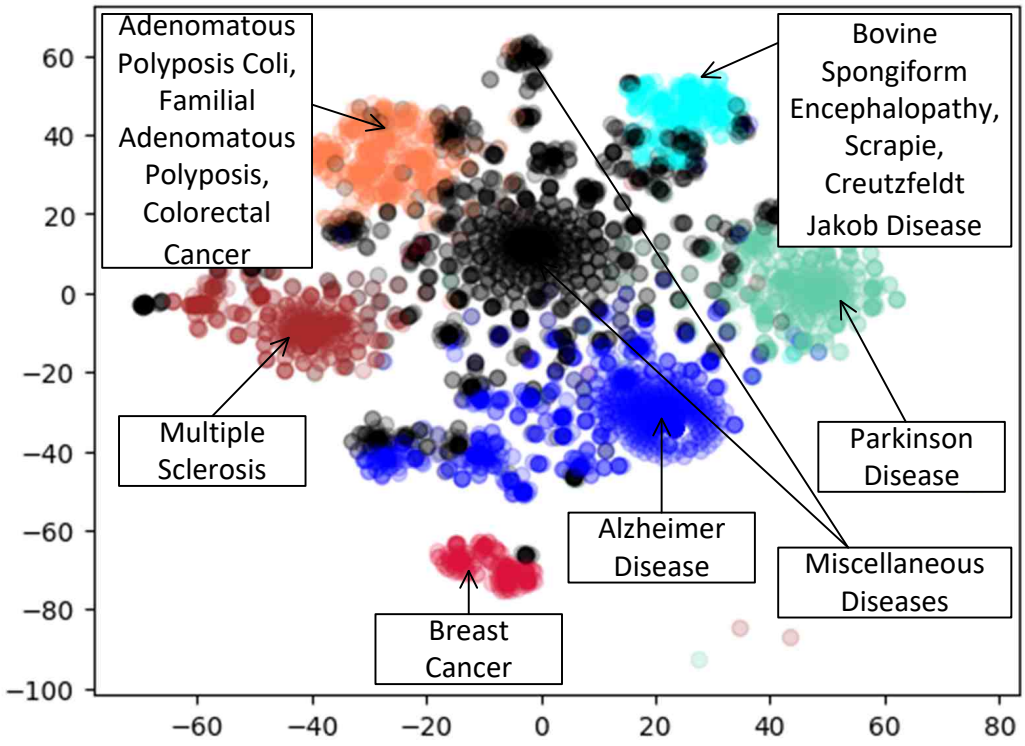

Fig. 7.31.: Clustering visualization for the TREC Genomics 2005 corpus using the modified document weighting scheme that relies on concept-based representations and clustered with k-means clustering, with $k = 7$, $\tau = 0.90$.

8. IMPROVING PATIENT CARE

Most health systems were designed to wait for an individual to become sick before they kick into reactive action. Hence, for the most part, health systems are designed to diagnose and treat illnesses or injuries instead of preventing the onset of a disease. Research also shows that this reactive model in health care is both expensive and, to some degree, ineffective in meeting the needs of today's population [90].

Preventing the onset of diseases is the key to improving people's health and keeping rising health care costs under control. General preventive health care guidelines based on age and gender have been integrated with Electronic Health Record systems through basic decision support systems and have led to improved performance in health care delivery.

Advanced integration which considers factors such as ethnicity, social history, medical history, family history need to be investigated. Integrating the preventive health care guidelines with the EHR based on above factors requires the extraction of the relevant information from these guidelines using text mining techniques.

Towards promoting the provision of preventive health care, the U.S. Preventive Services Task Force (USPSTF) has been established to identify scientific, evidence-based recommendations on dozens of clinical preventive health care services that are intended to reduce the risk of heart diseases, cancer, infectious diseases as well as improve the health of children, adults, and pregnant women [74].

However, despite the development and publication of these national preventive care guidelines, the actual rates of delivery of preventive health care services remain low [91]. Several studies have identified lack of time during the office visit as the most common barrier to the implementation of preventive care [92] [93].

The adoption of Electronic Health Record (EHR) systems has been encouraged by governments in many countries during the past decade [1]. The purpose of EHR is to help physicians better manage patient record and provide better health care. A considerable added benefit can be achieved by extracting information from the guidelines of clinical preventive health care services and automating their integration into the EHR.

Clinicians receive reminders when patients are due for taking the screening tests or exams. However, the criteria in these clinical decision support modules need to be manually updated by an expert, and the criteria is limited to age, gender, screening laboratory tests and / or diagnostic imaging exams and screening intervals.

None of the modifiable risk factors are extracted from the preventive guidelines and the modifiable risk factors are not considered by the CDS modules. In recent literature, researchers have identified that social history, including behavioral and environmental factors, are increasingly recognized as key modifiable factors for many causes of disease, disability, and mortality in the United States [90].

8.1 Literature Review

Clinical guidelines are "systematically developed statements to assist practitioner and patient decision making about appropriate health care for specific clinical circumstances" [94]. Their aim is to improve the quality of health care and reduce cost. Preventive health care guidelines guide physicians in helping patients prevent diseases before they happen.

Several previous researchers have focused on extracting some patterns to model medical guidelines. In 2007, Serban et al. developed a pattern extraction approach based on an ontology-driven linguistic pattern identification in order to automatically reconstruct the knowledge in the guidelines [95]. Other research work focused on associating the guidelines with a health care plan based on a domain-specific ontology [94]. It is not about extracting information from the content of the medical guidelines.

Researchers have also analyzed the design pattern of the guidelines and proposed better design patterns using computer-interpretable templates. For example, Peleg et al. compared five different design patterns of clinical guidelines and proposed new design patterns for two types of preventive care guidelines, namely, screening guidelines and immunization guidelines [96]. However, they focused on standardizing the guidelines to support the screening process instead of extracting the modifiable risk factors and integrating them with the EHR to provide more personalized preventive care.

There are a few examples of research work that have made use of natural language processing techniques to extract specific information from medical documents. Meystre et al. developed a system that makes use of UMLS MetaMap and a negation detection algorithm that used NegEx to extract different medical problems [97]. Rosales et al. investigated extracting medical concepts or events from electronic medical notes according to a pre-determined compound dictionary [98]. Li et al. proposed a tool to recognize medical concept by first extracting nouns, then constructing noun phrases from medical documents [99]. None of these researches investigated the extraction of modifiable risk factors, such as social history or family history related factors.

Previous research in creating associations between ontologies relied on using existing ontology that is a part of MeSH or WordNet to identify relationships between different medical terms [33] [81] [34]. However, the ontologies that present the entity and attribute relationships in a hierarchy without emphasizing the co-occurrences of the terms based on the content of the text.

On the other hand, the ontologies often do not include the different representations of the same medical term. For example, 'Type II Diabetes' and 'DM2' both represent 'Type 2 Diabetes Mellitus'. Different representations for same terminology happen quite often in the clinical notes within the Electronic Health Records (EHR), because physicians have their own preferences of recording notes.

In recent years, the distributed representation of word also called word embedding has gained a lot of interest in the research areas of text mining, natural language processing and health informatics (Chapter 2) [11] [25] [28]. Word embeddings have also been used in biomedical text processing [21] [75]. Word embedding emphasize the co-occurrences of the words based on the content of a given text document collection.

There are different ways to generate the distributed vector representations, which include probabilistic models [100] and dimensionality reduction on the word co-occurrence matrix [101]. Neural network is a new technique to generate the word embedding. It has been recently studied for biomedical text classification and clustering, where word is the basic unit for the text documents and word embedding is learned through neural networks [75] [102].

However, in the biomedical domain, clinical or medical concepts often contain more than one word. Especially, it is very hard to describe the symptoms using one word. Hence, it is necessary to analyze the associations between disease and symptom concepts based on their representations as more than one word. A system for generating 'concept' vectors from word embeddings is presented in Chapter 6.

In general, there is limited previous research that focuses on extraction of information from the guidelines. There is also no previous research on extraction of modifiable risk factors from the healthcare guidelines, for the purpose of automated integration with EHR. The integration with EHR is performed by using word embeddings.

Bridging the gap between preventive health care guidelines and EHR systems has a great potential to improve health care delivery.

8.2 Information Extraction from Preventive Care Guidelines

A framework that focuses on extracting the modifiable risk factors from the published clinical preventive health care services guidelines that includes demographic information, social history, family history, non-acute disease history, medical history, family history, preventive screening tests, etc. related information from the published

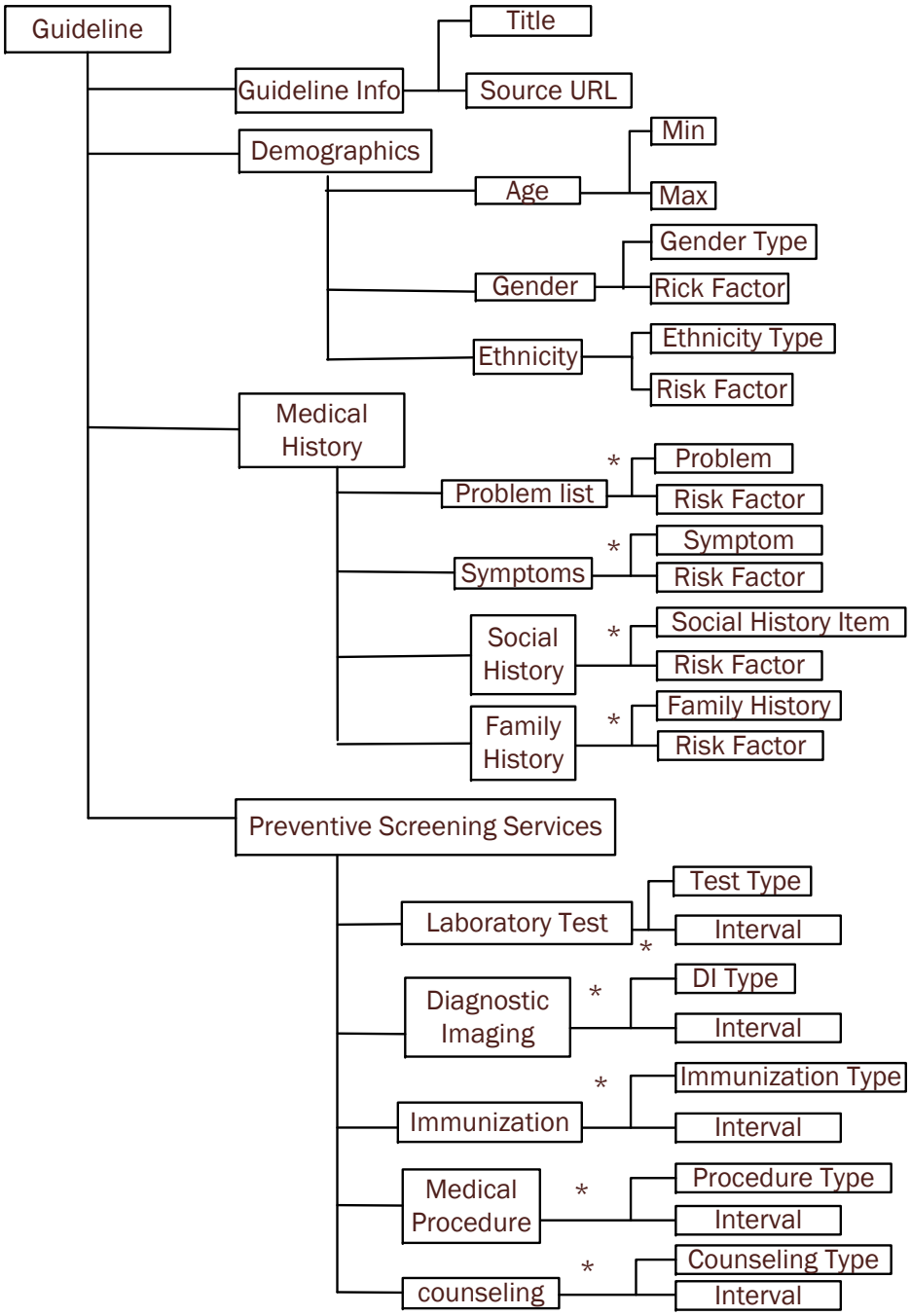

Fig. 8.1.: Interchange structure for mapping between preventive care guidelines and Electronic Health Records.

USPSTF preventive healthcare guidelines, and organize this information into sections of the proposed JSON data interchange structure according to standard EHR modules [103]. The interchange structure is shown in Figure 8.1, which is inspired from the structure of Electronic Health Record (Chapter 5, Figure 5.7).

This information can then be integrated with the EHR to support personalized preventive healthcare recommendation where modifiable factors are extracted from patients medical records in EHR and compared with the ones that are extracted from the preventive guidelines. The proposed framework consists of data pre-processing steps and a rule engine that makes use of different natural language processing and biomedical informatics modules. Through the proposed framework, modifiable risk factors are extracted and mapped into the sections of the proposed data interchange structure.

Preventive healthcare guidelines focus on disease prevention and provide details about categorized preventive health care services such as screenings, counseling services, or preventive medications. There are several challenges in information extraction from these guidelines such as distinguishing between screening tests and intervals for different age groups of population. For example, the guideline for cervical cancer prevention recommends different screening tests and intervals for women who are younger or older than 30 years.

Moreover, some guidelines describe preventive health care services that vary according to the ethnicity, family history or medical history of an individual. Hence, screening services also need to be adjusted based on an individual's family history and past medical history. Social history, family history and past medical history are usually stored in different modules or sub-modules of the EHR. The objective of this research is to extract information from the guidelines and prepare them for EHR integration.

In Figure 8.1, the information into three categories: demographics, medical history and preventive screening services. The second level of the JSON file structure corresponds to the modules of the EHR which include age, gender, ethnicity, labora-

tory tests, diagnostic imaging tests and so on. The third level includes the detailed items and the associated risk factor. The risk factor reflects how critical it is for an individual to receive the preventive services. For example, if an individual has a family history of diabetes, the risk factor of 'type 2 diabetes' increases. The original USPSTF recommendation title and URL in the JSON file structure. The asterisk ('*') represents the possibility of multiple entries.

The medical history section has all the modifiable risk factors. The Ethnicity element in the demographics is also one of the modifiable risk factor, since individuals of certain ethnicity have higher risk of getting the some preventable disease than the others. Each modifiable risk factor has an associated element risk degree that reflects how critical the risk factor is for an individual to receive the preventive services.

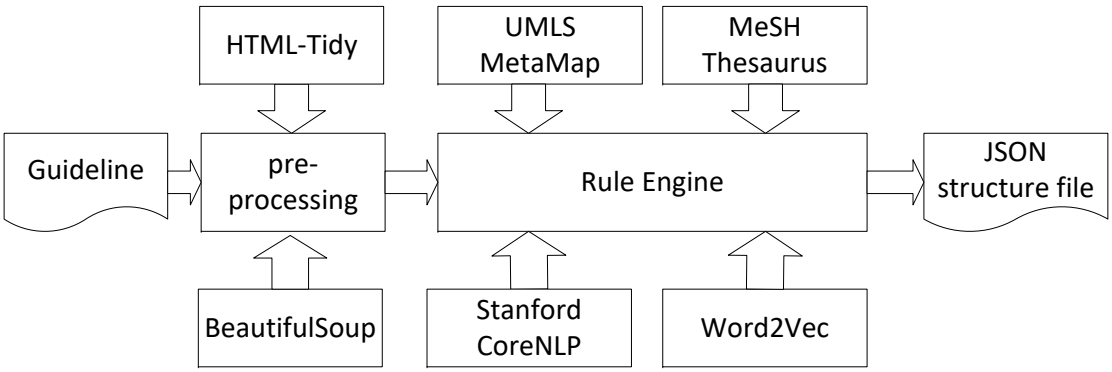

Fig. 8.2.: Framework of information extraction architecture from preventive care guidelines.

Figure 8.2 shows the proposed framework for information extraction and mapping to the JSON file structure. It is worth mentioning that the rule engine applies different information extraction and mapping algorithms for different EHR modules. In this research, algorithms have been implemented for all of the EHR modules except

'Counseling'. The absence of existing tools' scope to identify clinical 'Counseling' services led us to reserve this module as a potential future implementation. Some of these algorithms are presented in the following sections.

Pre-processing steps are applied to the guidelines before they are fed into the rule engine. Once the entire guideline is downloaded as an HTML document from the USPSTF website, the syntactical errors in the HTML document are corrected by using the HTML-Tidy library [104]. Among all the sections of a full guideline published by USPSTF, the sections 'Summary Statement' and 'Clinical Considerations' contain the clinical details of preventive health care services, the associated time intervals and the description of the patient population at risk [105]. These two sections are extracted by using the Python library BeautifulSoup [106].

Table 8.1.: Semantic types of MetaMap used for information extraction from preventive care guidelines.

EHR Modules	**Semantic Types**
Age Group	Age Group (aggp), Temporal Concept (tmco), Organism Attribute (orga)
Gender	Organism Attribute (orga), Population Group (popg)
Ethnicity	Population Group (popg)
Social History	Clinical Attribute (clna), Finding (fndg), Individual Behavior (inbe)
Problem History	Disease or Syndrome (dsyn), Neoplastic Process (neop)
Symptoms	Sign or Symptom (sosy)
Laboratory Tests	Laboratory Procedure (lbpr), Laboratory or Test Result (lbtr)
Imaging	Diagnostic Procedure (diap)
Procedure	Diagnostic Procedure (diap)

In order to extract information from the text-rich guidelines and map them to the EHR modules, we develop a rule engine which makes use of natural language processing or text mining modules: UMLS MetaMap (Chapter 4, Section 4.1), Stanford CoreNLP (Chapter 4, Section 4.2) and Medical Subject Headings thesaurus [62].

Table 8.1 shows the semantic types available from MetaMap output used to identify terms related to the EHR modules. In addition to those listed, the semantic type 'Idea or Concept(idcn)' is used to extract the possible risk factor associated with the extracted information. MetaMap makes use of various biomedical sources to map the phrases or terms in the input text to different semantic types. MetaMap is highly configurable with many options.

From Stanford CoreNLP, NLP tools such as the named entity recognizer and the dependency parser and used to analyze the syntactical relationships between words in a sentence. The dependency parser extracts relationships between words are dependencies between a governor (also known as a head) and a dependent, and displayed as a relationship's abbreviation. For instance, the abbreviation 'amod' represents a dependency relationship between an adjective and a noun. The named entity recognition tool is used to identify name entities within given phrases or terms like durations and numbers. In addition, the relationships between words are labeled using abbreviations. For example, word 'years' is in a 'tmod' relationship with word 'aged', which shows that 'years' is a temporal modifier of word 'aged'.

A Word2Vec model is also employed in this section using the trained model from [21]. The training was learned from two public corpora: PubMed and PubMed Central (PMC) [107]. These two corpora contain a large number of medical and biomedical words. Through the learned Word2Vec model, the most related words of a given word can be identified by calculating the distances between the source word and other targeted words. For example, given the word 'family', the most related words are: 'parent', 'father', 'mother', 'guardian', 'spouse', and 'grandparent'. The similarity scores for the most similar words for 'father' are presented in Table 8.2.

Table 8.2.: Top closest words of 'father' and the similarity scores from a word embeddings model [21].

Concept	**Most Similar Concepts**	**Similarity Score**
father	grandmother	0.839
	grandfather	0.836
	uncle	0.826
	aunt	0.817
	brother	0.808
	niece	0.802
	nephew	0.795

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) [62]. It is used for indexing, cataloging, and searching biomedical and health-related information in documents. Many synonyms, near-synonyms, and closely related concepts are included as entry terms. These entry terms are organized in an ontology. This ontology is very helpful for seeking the most relevant medical subject terms or the upper level subject category for a given concept. For example, given a term 'Computed Tomography', related medical terms can be identified through MeSH. The MeSH ontology shows that 'Computed Tomography' is a concept under the category 'Diagnostic Imaging'. This tool is used to map concepts into categories associated with EHR modules.

The rule engine applies different algorithms for information extraction and mapping to the corresponding EHR module. Six algorithms from those implemented in the rule engine are described next. These algorithms are for the EHR modules: 'Age Group', 'Social History', 'Problem History', 'Family History', 'Risk Degree Extraction' and 'Diagnostic Imaging'. Along with the 'Diagnostic Imaging' screening tests, the associated 'Time Interval' is also identified.

8.2.1 Age Group

Preventive health care services usually apply to a population of certain age group and the age group is specified in the 'Recommendation Summary' section of the guideline. All sentences of this section are passed through UMLS MetaMap to identify the phrases that are mapped to the semantic types corresponding to 'Age Group' listed in table 8.1.

Based on all the phrases mapped to the semantic types of 'Age Group', extended phrases (the mapped phrase, one phrase before the mapped phrase and one phrase after the mapped phrase) are constructed and passed to Stanford CoreNLP's NER tool in order to extract the range identified as 'Age Group'. If any of the words in the extended phrases are tagged as 'DURATION' by the NER tool, it is processed to extract the age range defined by minimum and maximum ages.

8.2.2 Social History

In the medical record, a patient's social history addresses aspects of the patient's personal life that have the potential to be clinically significant [108]. It includes the patient's alcohol and tobacco consumption status, sexual preference, diet and exercise. Social history also includes preconditions for some preventive services according to the guidelines.

Identifying social history related information is more challenging than information related to other EHR modules, especially when screening services vary based on different social history statuses. For example, in the lung cancer screening guidelines [48], identifying 'smoking history' as a social history without additional qualifications describing the duration or frequency of the smoking habit will not be accurate. Algorithm 8.1 presents the process of extracting social history information with these necessary qualifications. The input ($InputPhrases$) to this algorithm is the set of

Algorithm 8.1 Function to extract 'Social History' information from the preventive care guidelines, 'Social-History-Extraction($InputPhrases$)'.

```
1: OutputPhrases ← {}
2: for Phrase in InputPhrases do
3:     if (len(Phrase) > 1 word) and (Semantic Type of the Phrase is not 'fndg')
   then
4:         OutputPhrases ← OutputPhrases ∪ {Phrase}
5:     else
6:         DependentWords ← CoreNLP_DP(Phrase)
7:         p ← location of first DependentWord
8:         for word in Phrase do
9:             if  p >= location of word in Phrase  then
10:                Phrase ← Phrase + word
11:            end if
12:        end for
13:        OutputPhrases ← OutputPhrases ∪ {Phrase}
14:    end if
15: end for
16: return OutputPhrases
```

phrases mapped to the semantic types that represent 'Social History' in Table 8.1. The DP tool of Stanford CoreNLP ($CoreNLP\_DP$) is used to analyze the dependencies between words for each input.

8.2.3 Diagnostic Imaging

Different diagnostic imaging procedures are specified in the preventive guidelines as preventive health care services. For example, the breast cancer preventive guideline mentions that a screening mammography should be done for women between ages of 50 and 74 every two years [74]. In this guideline, screening test mammography is a diagnostic imaging test whereas 'every two years' is the time interval.

To extract the diagnostic imaging service mentioned in the guideline, the sentences are fed into the UMLS MetaMap to extract the phrases or terms that are mapped to the semantic type 'Diagnostic Procedure' as specified in Table 8.1. Based on the experimental tests, it was found that all of the diagnostic imaging tests, such as 'X-Ray', are mapped to the semantic type - 'Diagnostic Procedure'. However, other procedural tests, such as 'Pap smear', are also mapped to 'Diagnostic Procedure'.

To differentiate between the phrases that belong to the diagnostic imaging category and those belonging to procedures, the MeSH thesaurus is used. The disambiguation process works as follows: if any phrase or term is mapped to the semantic type 'Diagnostic Procedure', it is then used as query phrase to the MeSH thesaurus. If query phrase or a synonym of the phrase is located under the 'Diagnostic Imaging' branch in the MeSH ontology, the phrase is classified as diagnostic imaging, and not as medical procedure. Otherwise, the query phrase is classified as a medical procedure.

subsubsectionTime Interval

In order to extract the time interval associated with a diagnostic imaging service, the named entity recognition (NER) tool of Stanford CoreNLP is used to identify the name entities in the sentence that contains the extracted diagnostic imaging services. The terms that are tagged by NER as 'DURATION', 'SET' or 'NUMBER' are extracted as time interval.

For some guidelines, the screening interval is not mentioned along with the diagnostic imaging services within the same sentence, but included in a different subsection titled 'Screening Interval'. In these cases, the sentences of this subsection are used to

extract the time interval, and the extracted time interval is then applied to all the preventive services extracted from the guidelines including the diagnostic imaging services.

8.2.4 Problem History and Family History

An important aspect mentioned in healthcare guidelines is the presence of pre-existing conditions or previous diseases in patients or family members that put them at higher risk of a particular disease. The identification of such pre-existing conditions or previous diseases is accomplished by extracting all the disease related concepts from the guidelines that are tagged as two semantic types of UMLS MetaMap: 'Disease or Syndrome' and 'Neoplastic Process'.

After removing duplicates and the concepts that are the same as the preventive disease within the recommendation statement (e.g. 'lung cancer' and 'neoplasm of the lung' within the lung cancer recommendation statement [48]), the rest of the disease concepts are either related to patient which should be identified and extracted as problem history or related to a family member of the patient which should be identified and extracted as family history.

Algorithm 8.2 presents the process of extracting problem and family history based on the presence of family member related entity in the context. As discussed in Section 8.2, a list of family member related words are identified through Word2Vec. In this research, words 'parent', 'mother', 'father', 'brother' and 'sister' are used to identify all other family member related words. If any word(s) within a sentence are identified as family member related entity, the extracted disease concepts within the same sentence are extracted as risk factors of family history. Otherwise, the extracted disease concepts are extracted as risk factors of problem history.

Algorithm 8.2 Function to extract 'Problem History' and 'Family History' information from the preventive care guidelines, Problem-History-Extraction($Phrases - In - a - Sentences$)

$FamilyWords \leftarrow Word2Vec(`parent') \cup Word2Vec(`mother') \cup Word2Vec(`father') \cup Word2Vec(`brother') \cup Word2Vec(`sister')$
$ProblemHistory \leftarrow \{\}$
$FamilyHistory \leftarrow \{\}$
for $Phrase$ in $Phrases - In - a - Sentences$ **do**
  $flag = False$
  for $word$ in $FamilyWords$ **do**
    if $word$ in $Phrase$ **then**
      $FamilyHistory \leftarrow Phrase$
      $flag = True$
    end if
  end for
  if $flag = False$ **then**
    $ProblemHistory \leftarrow Phrase$
  end if
end for
return $ProblemHistory$, $FamilyHistory$

8.2.5 Risk Factor

The risk degree associated with each risk factor such as social history or family history is an attribute that determines the risk of a disease to an individual. For example, individuals with a prior history of polycystic ovarian syndrome or gestational diabetes are at a higher risk of diabetes [74]. Similarly, the risk degree of getting lung

Algorithm 8.3 Function to extract 'Risk Factor' information from the preventive care guidelines, Risk-Factor-Identification($Identified - risk - factors$)

$PhrasesToEval \leftarrow \{Phrase\ Before\ Identified - risk - factors\} \cup \{Identified - risk - factors\} \cup \{Phrase\ After\ Identified - risk - factors\}$
if word 'risk' in $PhrasesToEval$ **then**
  for $word$ in $PhrasesToEval$ **do**
    if $word$ is tagged as 'qnco' **then**
      $RiskQuantifier \leftarrow word$
      return $RiskQuantifier$
    end if
  end for
end if
return $None$

cancer decreases after an individual quits smoking. Thus, risk degree is an important indicator to determine if specific preventive care for certain diseases should be offered to a patient at a different interval than the standard time interval.

The algorithm to extract the associated risk degree for risk factors are detailed in Algorithm 8.3. The algorithm first locates all the identified risk factors, such as social history, ethnicity, family history and problem history. Then, a phrase before and after the identified risk factors are evaluated to see whether risk degree related words can be located. If the word 'risk' is found in the any of the phrases, a word that is tagged as the 'Quantitative Concept' by UMLS MetaMap is extracted as risk degree for the corresponding risk factor.

8.3 Extraction Results

Using the algorithms shown in Section 8.2, proof-of-concept extraction is performed on the lung cancer [48] and diabetes [73] screening recommendation guidelines. The results are shown in the format of the interchange structure proposed in Figure 8.1.

Only a segment of the guidelines are shown in Figures **??** and **??** from Chapter 5, Section 5.3. However, the input to the framework consists of all the paragraphs in the guideline.

```
{   { Guideline Info:        { Title: Lung Cancer Screening,
                               URL: https://www.uspreventiveservicestaskforce.org/
                              Page/Document/RecommendationStatementFinal/lung-
                              cancer-screening } }
    { Demographics:          { Age:  { Min: 55,
                                       Max: 80 } }
    { Medical History:       { Social History:
                               [{name: cumulative exposure to tobacco smoke },
                                { name: smoking },
                                { name: 30 pack-year smoking history },
                                { name: current smokers },
                                { name: 15 years of smoking cessation } ] } }
    { Preventive Screening Services: { Diagnostic Imaging:
                                       [{ Test: Computed Tomography,
                                        interval: 3 annual },
                                      { Test: Chest Radiography } ] },
                                     { Laboratory Test:
                                       [{ Test: Sputum Cytology } ] }
}
```

Fig. 8.3.: Populated output in interchange structure of Lung Cancer Screening Guideline.

Figure 8.3 shows the populated data interchange structure of the lung cancer preventive guideline. The age range for lung cancer screening is different from that of type 2 diabetes. The age range of 55 to 80 which is extracted from the guideline and stored in the age group of JSON output interchange structure.

The preconditions for lung cancer screening are straight forward. They relate to the smoking history of an individual. The detailed description of the smoking history is provided in the guideline. This description will be mapped to the social history of the EHR module. By using the social history extraction algorithm described in Section 8.2.2, five different descriptions of smoking history are extracted: 'cumulative exposure to tobacco smoke', 'smoking', 'current smoker', '30 pack-year smoking history' and '15 years of smoking cessation'. These descriptions correspond to different facts about the smoking behavior of the patient. So, they are included in the JSON files as five entries under the social history.

Although these smoking history descriptions are similar, they reflect different facts about the smoking behavior of the patient. 'Current smoker' details the current status of social behavior of the patient, whereas '30 pack-year smoking history' details their history over a long period of time. The smoking history of a patient recorded in the EHR might be in different forms. Same information extraction process can be applied to medical records to extract the modifiable risk factors and word embedding models can be used to measure the similarities between the concepts.

After applying the algorithms of the rule engine, two kinds of diagnostics imaging tests are identified: computed tomography and chest radiography. They are included in the diagnostic imaging section of the JSON interchange file. Only computed tomography is associated with a time interval which was extracted from the original guideline. The laboratory test 'sputum cytology' is also extracted. 'Chest radiography' and 'sputum cytology' are not assigned a time interval since time intervals for these tests are not mentioned in the guideline.

{ { Guideline Info: { Title: **Abnormal Blood Glucose and Type 2 Diabetes Mellitus: Screening** ,
URL: **https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/screening-for-abnormal-blood-glucose-and-type-2-diabetes** } },
{ Demographics: { Age: { Min: **40**,
Max: **70** } },
{ Ethnicity: [{ ethnicity: **African Americans**,
risk factor: **Increase** },
{ ethnicity: **American Indians**,
risk factor: **Increase** },
{ ethnicity: **Alaskan Natives**,
risk factor: **Increase** },
{ ethnicity: **Asian Americans**,
risk factor: **Increase** },
{ ethnicity: **Hispanics**,
risk factor: **Increase** },
{ ethnicity: **Latinos,**
risk factor: **Increase** },
{ ethnicity: **Native Hawaiians**,
risk factor: **Increase** },
{ ethnicity: **Pacific Islanders**,
risk factor: **Increase** }] } },
{ Medical History: { Problem list:
[{ problem: **Obese** },
{ problem: **Polycystic Ovarian Syndrome**,
risk factor: **Increase** },
{ problem: **Cardiovascular Disesase (CVD)**,
risk factor: **Increase** }] }
{ Family History: [{ name: **Diabetes**,
risk factor: **Increase** }] },
{ Social History: [{ name: **Gestational Diabetes**,
risk factor: **Increase** }] },
{ Symptoms: [{ symptom: **Overweight** }] } },
{ Preventive Screening Services: { Medical Procedure:
[{ Procedure: **Body Mass Index**,
interval: **every 3 years** },
{ Procedure: **Oral Glucose Tolerance Test**,
interval: **every 3 years** }] },
{ Laboratory Test: [{ test: **Plasma Glucose**,
interval: **every 3 years** },
{ test: **Hemoglobin A1C**,
interval: **every 3 years** },
{ test: **Glucose Levels**,
interval: **every 3 years** }] } }
}

Fig. 8.4.: Populated output in interchange structure of Type 2 Diabetes Mellitus Screening Guideline.

As for type 2 diabetes mellitus screening guideline, the age group (minimum and maximum age) mentioned in the guideline is extracted using the algorithm described in Section 8.2, and results shown in Figure 8.4.

Ethnic groups are also extracted. There are eight different ethnicities mentioned in the guideline. People belonging to these ethnic groups are at increased risk of contracting type 2 diabetes. Hence, all these ethnicity groups are extracted along with the risk factor 'increase'.

The development of type 2 diabetes is affected by many factors. As described in the guideline, if an individual has a medical problem, such as obesity, cardiovascular disease (CVD) or polycystic ovarian syndrome, the risk of contracting type 2 diabetes increases. These risk problems are extracted and used to populate the 'Problem History', which is the non-acute diseases section of the interchange structure.

In addition, a family history of diabetes and a history of gestational diabetes for women indicates an increased risk of contracting type 2 diabetes. Hence, this information is also extracted and used to populate the corresponding elements of the JSON output.

The preventive healthcare services mentioned in the guideline include blood tests such as Plasma Glucose, Hemoglobin A1C and Glucose levels. These tests should be performed every three years. The BMI measurement is extracted and stored as a medical procedure, whereas the other blood tests are stored under laboratory tests.

To fully evaluate the correctness of the proposed framework, physicians or nurse practitioners need to be recruited to manually evaluate the extracted information against the information in the original guideline, to evaluate integration of the proposed data interchange structure with the EHR, further research needs to be done to extract the patient's medical record from the EHR and compare the information against that extracted from the preventive guidelines. However, the proposed framework initiates the first step towards personalized preventive care by extracting modifiable risk factors from the preventive healthcare guidelines. Furthermore, the

proposed extraction process and algorithms can be applied to other narrative medical files or guidelines where modifiable risk factors are critical to diseases' prevention and treatments.

By extracting all this information, the populated data interchange structure can be compared against the patient's medical records in the EHR to generate a personalized preventive care plans for physicians to make final preventive care decision.

8.4 Note Concept Extraction

Using UMLS MetaMap (Chapter 4, Section 4.1) data for different EHR modules is extracted from the clinical notes [79]. Doing this ensures that no data is lost in the clinical notes, and helps create a better, organized picture of the patient's medical history.

Disease and symptoms concepts were extracted from the clinical notes as a part of Chapter 6, Section 6.3. These extracted concepts were clustered separately, and the results were presented in Chapter 6, Section 6.3 using the k-means clustering algorithm (Chapter 3, Section 3.1.2) with $k = 50$ clusters for each subset of concepts.

The clustering results provide a picture of the relationships between disease and symptom concepts, and their clusters. Table 8.3 also shows the most similar symptoms concepts to every disease concept. It is clear from Table 8.3 that there isn't as strong a relationship between the diseases and symptoms concepts, and the similarity scores are low.

In order to validate the associations between the diseases and symptoms in the vector space, the clinical notes were investigated to validate whether the diseases occur in the same clinical notes with the associated symptoms. It was found that indeed the diseases co-occur with the symptoms in the clinical notes.

However, often a few diseases were described in the notes as existing problems, and are mentioned together in the clinical notes with the symptoms. Without other interpretation, it is hard for the algorithm to determine which symptoms correspond

Table 8.3.: Examples of disease concepts and the top 3 closest symptoms concepts based on the similarity scores from embeddings trained as concept-based embeddings on the IU Health EHR clinical notes data.

Diseases	**Most Similar Symptoms**	**Similarity Score**
chronic obstructive pulmonary disease (COPD)	peptic ulcer symptoms	0.442
	chronic pain	0.392
	chronic cough	0.359
	chronic back pain	0.339
	chronic abdominal pain	0.338
	chronic chest pain	0.306
	gastroesophageal reflux disease symptoms	0.302
alzheimer disease	sleep disorders	0.467
	groin tenderness	0.327
breast cancer	breast pain	0.458
	breast discomfort	0.455
	breast tenderness	0.424
coronary artery disease (CAD)	coronary chest pain	0.471
	peptic ulcer symptoms	0.426
	coronary symptoms	0.334
diabetes mellitus type 2	symptom nausea	0.344
	weakness of lower limb	0.310

to what diseases exactly. For example, if a clinical note describes that the patient has peptic ulcer symptoms and a history of chronic obstructive pulmonary disease (COPD) and coronary artery disease (CAD), it is hard to determine whether the peptic ulcer symptoms were more associated with COPD or CAD without it being interpreted by a physician in the presence of other information.

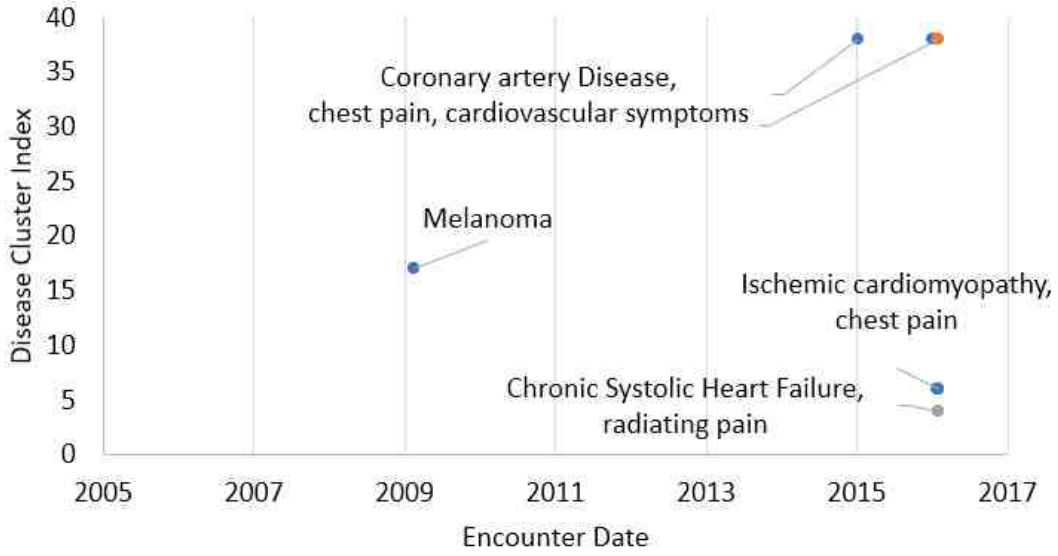

Fig. 8.5.: The disease and symptom progression timeline of a patient diagnosed with Coronary Artery Disease.

If patients in the dataset had COPD and CAD with symptoms of peptic ulcer, the peptic ulcer symptoms were found to be associated with both diseases, as shown in Table 8.3. Literature was used to validate the diseases and symptoms associations identified through our proposed methods. For example, we have searched literature about type 2 diabetes and weakness of lower limb, and found that previous research in diabetes has shown a decrease in lower-limb muscle strength in diabetic patients [109] [110].

Through the concept extraction and concept association mining, we have generated clusters of diseases and symptoms, and also identified the diseases and associated symptoms from the clinical notes. To support clinical decisions by efficiently making use of the clinical notes, a two-dimensional visualization tool to visualize the development of diseases and associated symptoms over time. The X-axis of the visualization represents the time of the encounters, while the Y-axis represents the cluster index of the disease(s).

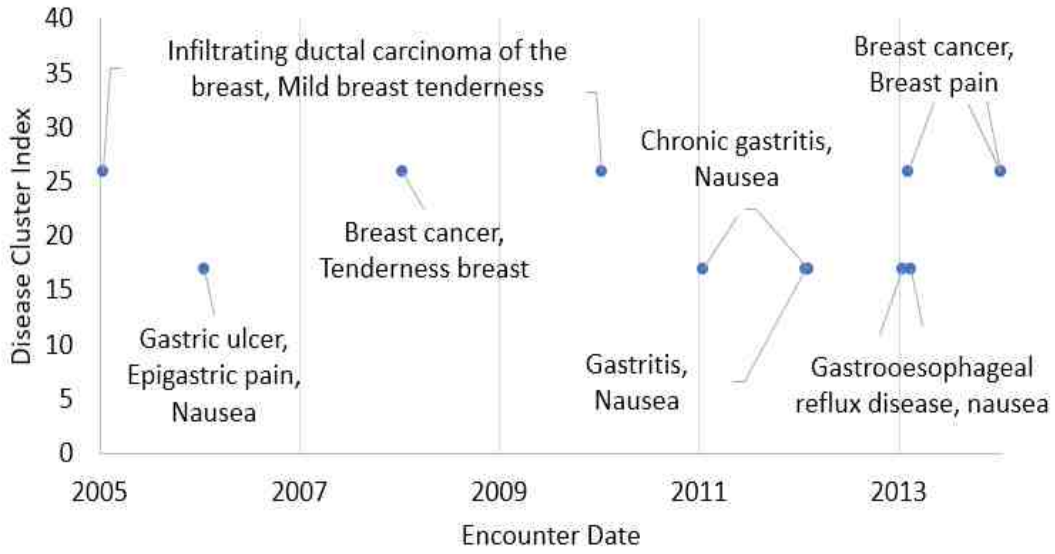

Fig. 8.6.: The disease and symptom progression timeline of a patient diagnosed with Breast Cancer.

For example, if the clinical note mentions 'ischemic cardiomyopathy', it belongs to cluster 6 according to Chapter 6, Table 6.4. So, 6 is the value for Y-axis.

Two patients to explore the visualization of diseases and symptoms extracted from the clinical notes over time as demonstrated in Figures 8.5 and 8.6. These two patients had a diagnosis of 'ischemic cardiomyopathy' and 'breast cancer' respectively in the diagnosis module of the EHR.

Based on Figure 8.5, it is visualized that the patient had 'Melanoma' mentioned in the clinical note only around 2009. However, there are no associated symptoms mentioned during that period. 'Melanoma' is not mentioned in the clinical notes after 2009. Starting from 2015, this patient had CAD and related symptoms, such as 'chest pain' and 'cardiovascular symptoms' mentioned in notes. Around 2016, the patient had been diagnosed with 'ischemic cardiomyopathy'. Following that, this disease and

'Chronic Systolic Heart Failure' were both mentioned in the clinical notes along with related symptoms. From this visualization, the development of the diseases along with the symptoms can be clearly demonstrated.

Figure 8.6 shows the diseases and most related symptoms extracted from a patient who had been diagnosed with 'breast cancer' in 2005. It shows that 'breast cancer' and the related symptoms have been mentioned in the clinical notes periodically from 2005 till late 2013. Other than 'breast cancer', gastric diseases and the most related symptoms, such as 'nausea' are mentioned in the clinical notes periodically over the years from 2006 to 2013.

This disease and symptom visualization can help physicians' review the medical problems of a patient. It is envisioned that if related medications and laboratory tests can be added to the visualization, it will serve as a good decision support tool for physicians.

9. CONCLUSION

As a part of this thesis, different methods to generate biomedical concept word embeddings, and clustering of biomedical concepts are evaluated. The generated concept vectors are evaluated by clustering concepts, and for clustering biomedical documents based on concepts of diseases. Exploratory work is performed to extract information from preventive healthcare guidelines. In the end, an extraction and visualizating algorithm for biomedical concepts from EHR clinical notes was examined.

Word embeddings are among the primary technical concepts used as a part of this thesis. Chapter 2 discusses the development of word embeddings from simple bag-of-words representations to complex representations that have roots in neural networks like recurrent neural networks and long short-term memory networks.

Part of the evaluation for concept vectors is generated by using unsupervised clustering algorithms. The clustering algorithms used as a part of this research, self-organizing maps and k-means clustering, are detailed in Chapter 3, Section 3.1. Details about internal clustering evaluation metrics like Davies-Bouldin index, and external evaluation metrics like purity and F-measure are provided in Chapter 3, Section 3.2. The visualization procedure of these generated clusters is detailed in Chapter 3, Section 3.3. These visualization techniques were used to visually evaluate the clustering results.

Chapter 4 describes the third-party tools used in this work, like UMLS MetaMap (Section 4.1) and Stanford CoreNLP (Section 4.2). UMLS MetaMap is a natural language processing tool developed by the U. S. National Library of Medicine to extract biomedical concepts from any text and map it to the entries in Metathesaurus. Stanford CoreNLP is a suite of linguistic tools to extract relationships like part-of-speech tags, dependencies, entities, etc. from text structure.

The different data sources used in this thesis are detailed as a part of Chapter 5. This chapter discusses the different biomedical document sources derived from PubMed like PubMed Central – Open Access (Chapter 5, Section 5.1.1), Ohsumed text collection (Chapter 5, Section 5.1.2) and the TREC Genomics 2005 corpus (Chapter 5, Section 5.1.3). A corpus of data from clinical EHR was also used in Chapters 6 and 8. This dataset included data from all of the different EHR modules, like diagnosis, medications, clinical notes, etc. Preventive care guidelines issued by the USPSTF were also used in Chapter 8, and a short description of them is provided in Chapter 5, Section 5.3.

Word embeddings are generated by using Word2Vec's skip-gram architecture and used to generate concept representations of biomedical concepts (Chapter 6. The two proposed methods to generate concept vectors are by aggregating word vectors to concept vectors (Chapter 6, Section 6.1), and generating concept vectors by training a Word2Vec model after pre-processing concepts into a single entity. The generated word embeddings are also compared based on their similarity scores. Both concept embedding generation approaches can capture the association of the concepts based on the content of the training text collection. Section 6.3 of Chapter 6 shows the results of clustering disease and symptom concepts separately. The results are promising and show that similar concepts tend to be clustered together, because of the similarity scores. The concept clustering also shows that concepts in similar parts of the body, or ones that co-occur for related diseases or symptoms, are also clustered together. The word embedding model successfully captures the associations between words based on the co-occurrences of the word within the clinical notes.

In Chapter 7, a framework for biomedical document clustering and visualization based on vectorizing concepts of diseases is proposed and evaluated. The concept vectors generation was described in Chapter 6, but the document representation was discussed in Section 7.2. Section 7.3 contains details about the first proof-of-concept weighting scheme, and Section 7.4 shows the results of clustering generated by using the proposed weighting scheme.

The proposed representation of documents (Chapter 7, Section 7.3) considers the local content and semantic similarity between the concepts within the documents is used. The results show that the clustering occurs based on the concepts of similar nature, similar area and organs of the body, and concepts which are synonymous to one another. Nearby clusters are related in most cases, as well. This kind of visualization will help researchers explore related articles based on concepts of diseases.

Modifications to the document weighting scheme, proposed in Chapter 7, Section 7.5, improve the clustering document vectors generated and the clustering results. The modified weighting scheme is compared against baseline TF-IDF for text clustering and visualization. Results from three different biomedical text document collections demonstrate that the proposed weighting scheme using the concept embedding achieve better much clustering performance than the baseline TF-IDF.

In Chapter 8, methods to extend the proposed concept representations towards improving the quality of healthcare services provided to patients are discussed. Integration of the preventive care guidelines into the EHR, by generating a directly mappable structure from narrative guidelines. The process to convert text guidelines into an interchange structure is proposed in Chapter 8, Section 8.2. In order to promote the personalized preventive healthcare services, we propose a modifiable risk factor extraction framework that can be applied to the preventive healthcare guidelines to facilitate the integration of the guidelines with electronic health record (EHR) systems. This framework extracts many factors, such as ethnicity, social history, family history, medical history from the guidelines and populates the data interchange structure for EHR integration.

The discussion in Chapter 8 also includes various methods that can help improve patient outcome based on improved representations of the data stored within the EHR (Chapter 8, Section 8.4). The proposed concept extraction method uses disease and symptom concepts to represent the progression of diseases over the time period. The Y-axis contains information about the clusters in which the given concept was identified. The temporal visualization tool assists in visualizing the history of diseases

along with the symptoms that are recorded in the narrative clinical notes. This visualization tool can provide physicians an overview of the medical history of a patient and support decision making.

In conclusion, the work presented as a part of this thesis has potential to improve representation of biomedical concepts in word embeddings. The applications that can benefit from such a vector representation of biomedical concepts like document clustering, improving preventive healthcare, and in providing a holistic patient history visualization technique.

10. FUTURE WORK

The most profound limitations of this work is the requirement to first capture biomedical concepts by pre-processing and then re-train a model based on the concepts extracted for a concept-based representation model. This increases the computation time exponentially as the corpus size grows. This limitation is also accentuated by the fact that there are multiple normalized representations for the same concept in UMLS' Metathesaurus, and provided after processing with UMLS MetaMap. A different pre-processing tool could help with simplifying the pre-processing required.

Another limitation is imposed if the corpus size is small. The model can not accurately capture the associations between concepts when the number of instances is very small, because the training samples are not enough to train the neural networks to identify patterns. This problem is also known as the 'curse of dimensionality'.

The word-based representations, on the other hand, do not require the model to be retrained to ensure coverage across all of the concepts. At the same time, for concept vectors generated from word-based representations, if two concepts have a common word, these two concepts tend to be very close to each other, although they might represent two different concepts. This is especially common for high frequency words like 'disease', 'cancer', 'pain', among others.

Even for word-based representations, UMLS MetaMap still needs to run over this model to extract the biomedical concepts. At the same time, word embeddings that are representative of the corpus are difficult to find, since the input corpus to word embeddings can change the behavior and relationships found among word vectors. Preliminary tests with word embeddings trained on non-biomedical data (news, Wikipedia dumps) showed issues with coverage of concept terms, and low similarity scores.

Future work in improving the quality of word embeddings may focus on using newer word embeddings creation algorithms like ELMo, which relies on semantics and synatical structure of each sentence to generate the word embeddings. ELMo has shown promising results in tests that are often used to measure word embedding performances such as association tasks, coreferencing, entity recognition and question-answering.

With regards to the document clustering section, potential future work may include extending this framework to biomedical document clustering by including concept embedding of other types of biomedical concepts, such as medications, diagnostic and laboratory tests, treatments, etc.

Evaluating the visualization aid for the task of biomedical document search is also another possible step in the direction of improving biomedical document clustering.

On the other hand, more complex clustering algorithms like hierarchical clustering architecture can be explored for clustering and visualization of larger text collections.

The integration of the data extracted to the interchange structure from preventive care guidelines with the EHR should be straight-forward with if-else conditional rules that scan a patient's history and previous medical records. This integration with the clinical decision support system of the EHR is bound to improve the quality of healthcare services provided and aid the prevention of onset of non-acute diseases.

The information extraction algorithms can also be customized and improved to pull information from other preventive care guidelines like those from the Center for Disease Control and Prevention, Health Resources & Services Administration, and so on. Comparative evaluation towards other related methodologies on risk factor extraction, developing concept similarities and association mining algorithms are also other directions in which work in this domain can be concentrated.

On the whole, integrating the extracted information from healthcare guidelines, with the stored information in the EHR systems facilitates building a personalized preventive care recommendation engine for each patient.

With respect to concept clustering and patient history representations from previously stored information in the clinical notes, future work may include working with physicians or clinical annotators to evaluate the effectiveness of the concept association mining model and visualization tool for decision support by including a large number of clinical notes from the EHR system. The model can be expanded further to analyze the associations of other clinical concepts, such as social history, family history, medications, diagnostic tests and so on.

The field of biomedical natural language processing is at a nascent, exploratory stage and in the long-run has the potential to transform the quality of healthcare provided. Various directions for this include changes that can improve information is consumed using document clustering techniques, or personalized care delivery with automated integration with preventive care guidelines. Physician and doctors can also be assisted by improving how they interact with, analyze and consume information stored in the EHRs, and thus simplifying their work.

REFERENCES

REFERENCES

[1] S. L. Meigs and M. Solomon, "Electronic health record use a bitter pill for many physicians," *Perspectives in health information management*, vol. 13, no. Winter, 2016.

[2] C. for Disease Control, Prevention *et al.*, "Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services," *MMWR: Morbidity and mortality weekly report*, vol. 52, no. Suppl. 1, pp. 1–17, 2003.

[3] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.

[4] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[5] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1988, pp. 281–285.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. January, pp. 993–1022, 2003.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013, (Accessed: November 14, 2018). [Online]. Available: https://arxiv.org/abs/1301.3781

[8] Google, "Vector Representation of Words - TensorFlow," 2018, (Accessed: September 23, 2018). [Online]. Available: https://www.tensorflow.org/tutorials/representation/word2vec

[9] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[10] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[12] Google, "Google code archive - long term storage for google code project hosting," 2018, (Accessed: November 9, 2018). [Online]. Available: https://code.google.com/archive/p/word2vec/

[13] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. August, pp. 2493–2537, 2011.

[15] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," *arXiv preprint arXiv:1406.3676*, 2014, (Accessed: November 14, 2018). [Online]. Available: https://arxiv.org/abs/1406.3676

[16] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 1555–1565.

[17] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 384–394.

[18] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013, (Accessed: November 14, 2018). [Online]. Available: https://arxiv.org/abs/1309.4168

[19] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, 2014, pp. 595–603.

[20] A. Roy, Y. Park, and S. Pan, "Learning domain-specific word embeddings from sparse cybersecurity texts," *arXiv preprint arXiv:1709.07470*, 2017, (Accessed: November 14, 2018). [Online]. Available: https://arxiv.org/abs/1709.07470

[21] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing," in *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, 2013, pp. 39–43, (Accessed: October 3, 2018). [Online]. Available: http://bio.nlplab.org/pdf/pyysalo13literature.pdf

[22] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.

[23] D. McClosky, E. Charniak, and M. Johnson, "Automatic domain adaptation for parsing," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 28–36.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," 2018, (Accessed: November 9, 2018). [Online]. Available: https://nlp.stanford.edu/projects/glove/

[25] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[26] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[27] Facebook Open Source, "fastText," 2018, (Accessed: November 9, 2018). [Online]. Available: https://fasttext.cc

[28] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018, (Accessed: November 14, 2018). [Online]. Available: https://arxiv.org/abs/1802.05365

[29] Allen Institute of Artificial Intelligence, "ELMo: Deep contextualized word representations," 2018, (Accessed: November 9, 2018). [Online]. Available: https://allennlp.org/elmo

[30] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE transactions on neural networks*, vol. 11, no. 3, pp. 574–585, 2000.

[31] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[32] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[33] S. Logeswari and K. Premalatha, "Biomedical document clustering using ontology based concept weight," in *2013 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2013, pp. 1–4.

[34] X. Zhang, L. Jing, X. Hu, M. Ng, and X. Zhou, "A comparative study of ontology based term similarity measures on pubmed document clustering," in *International Conference on Database Systems for Advanced Applications*. Springer, 2007, pp. 115–126.

[35] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[36] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 39.

[37] C. Van Rijsbergen, *Information Retrieval*. Butterworths, 1979, (Accessed: October 1, 2018). [Online]. Available: http://www.dcs.gla.ac.uk/Keith/Preface.html

[38] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[39] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, November 2008.

[40] U. S. National Library of Medicine, "Unified Medical Language System (UMLS)," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.nlm.nih.gov/research/umls/

[41] U. S. National Library of Medicine, "UMLS - Metathesaurus," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

[42] U. S. National Library of Medicine, "The UMLS Semantic Network," 2018, (Accessed: November 9, 2018). [Online]. Available: https://semanticnetwork.nlm.nih.gov/

[43] U. S. National Library of Medicine, "The SPECIALIST NLP Tools," 2018, (Accessed: November 9, 2018). [Online]. Available: https://lexsrv3.nlm.nih.gov/Specialist/

[44] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[45] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[46] U. S. National Library of Medicine, "MetaMap - A Tool For Recognizing UMLS Concepts in Text," 2018, (Accessed: November 9, 2018). [Online]. Available: https://metamap.nlm.nih.gov

[47] U. S. National Library of Medicine, "MetaMap 2012 XML Explained," 2018, (Accessed: November 9, 2018). [Online]. Available: https://metamap.nlm.nih.gov/Docs/MM12_XML_Info.shtml

[48] U. S. Preventive Services Task Force, "Final Recommendation Statement: Lung Cancer: Screening," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/lung-cancer-screening

[49] S. Shah, "A simple Python wrapper for MetaMap." 2018, (Accessed: November 9, 2018). [Online]. Available: https://github.com/setu4993/Python-Wrapper-for-MetaMap

[50] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60, (Accessed: October 3, 2018). [Online]. Available: https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf

[51] Stanford CoreNLP, "corenlp.run," 2018, (Accessed: November 9, 2018). [Online]. Available: http://corenlp.run

[52] Python Software Foundation, "pycorenlp - PyPI," 2018, (Accessed: November 9, 2018). [Online]. Available: https://pypi.org/project/pycorenlp/

[53] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180, (Accessed: October 3, 2018). [Online]. Available: https://nlp.stanford.edu/~manning/papers/tagging.pdf

[54] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993, (Accessed: October 3, 2018). [Online]. Available: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1246&context=cis_reports

[55] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 740–750, (Accessed: October 3, 2018). [Online]. Available: https://cs.stanford.edu/~danqi/papers/emnlp2014.pdf

[56] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 363–370, (Accessed: October 3, 2018). [Online]. Available: https://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

[57] G. Angeli, M. J. J. Premkumar, and C. D. Manning, "Leveraging linguistic structure for open domain information extraction," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 344–354, (Accessed: October 3, 2018). [Online]. Available: https://nlp.stanford.edu/pubs/2015angeli-openie.pdf

[58] U. S. National Library of Medicine, "Home - PubMed - NCBI," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/

[59] D. A. B. Lindberg, "Internet access to the national library of medicine." *Effective clinical practice: ECP*, vol. 3, no. 5, p. 256, 2000.

[60] U. S. National Library of Medicine, "Home - PMC - NCBI," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[61] U. S. National Library of Medicine, "MEDLINE ® Description of the Database," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

[62] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[63] U. S. National Library of Medicine, "Home - MeSH - NCBI," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.ncbi.nlm.nih.gov/mesh

[64] U. S. National Library of Medicine, "Open Access Subset," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[65] S. Zhu, J. Zeng, and H. Mamitsuka, "Enhancing medline document clustering by incorporating mesh semantic similarity," *Bioinformatics*, vol. 25, no. 15, pp. 1944–1951, 2009.

[66] A. Moschitti, "Text Categorization Corpora," 2018, (Accessed: November 9, 2018). [Online]. Available: ttp://disi.unitn.it/moschitti/corpora.htm

[67] A. Moschitti and R. Basili, "Complex linguistic features for text classification: A comprehensive study," in *European Conference on Information Retrieval*. Springer, 2004, pp. 181–196.

[68] National Institute of Standards and Technology, "Text REtrieval Conference (TREC) 2005 Genomics Track," 2018, (Accessed: November 9, 2018). [Online]. Available: https://trec.nist.gov/data/t14_genomics.html

[69] W. Hersh, A. Cohen, J. Yang, R. T. Bhupatiraju, P. Roberts, and M. Hearst, "Trec 2005 genomics track overview," in *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, 2005.

[70] P. Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: a framework for mining clinical notes through symptom relation graphs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1167–1175.

[71] P. A. McKee, W. P. Castelli, P. M. McNamara, and W. B. Kannel, "The natural history of congestive heart failure: the framingham study," *New England Journal of Medicine*, vol. 285, no. 26, pp. 1441–1446, 1971.

[72] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature communications*, vol. 5, p. 4212, 2014.

[73] U. S. Preventive Services Task Force, "Final Update Summary: Abnormal Blood Glucose and Type 2 Diabetes Mellitus: Screening," 2018, (Accessed: November 9, 2018). [Online]. Available: https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/screening-for-abnormal-blood-glucose-and-type-2-diabetes

[74] US Preventive Services Task Force, *Guide to clinical preventive services: report of the US Preventive Services Task Force*. DIANE publishing, 1989.

[75] S. Tulkens, S. Šuster, and W. Daelemans, "Using distributed representations to disambiguate biomedical and clinical concepts," *arXiv preprint arXiv:1608.05605*, 2016, (Accessed: November 14, 2018). [Online]. Available: https://arxiv.org/abs/1608.05605

[76] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," *Neurocomputing*, vol. 266, pp. 336–352, 2017.

[77] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Biomedical Natural Language Processing," 2018, (Accessed: November 9, 2018). [Online]. Available: http://bio.nlplab.org/#word-vectors

[78] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[79] S. Shah, X. Luo, S. Kanakasabai, R. Tuason, and G. Klopper, "Neural networks for mining the associations between diseases and symptoms in clinical notes," *Health Information Science and Systems*, vol. 7, no. 1, p. 1, November 2018, (Accessed: November 28, 2018). [Online]. Available: https://doi.org/10.1007/s13755-018-0062-0

[80] B. K. Nallamothu and T. S. Baman, *Dilated and Restrictive Cardiomyopathy*. John Wiley & Sons, Ltd, 2013, ch. 14, pp. 178–186, (Accessed: November 14, 2018). [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118484784.ch14

[81] I. Yoo, X. Hu, and I.-Y. Song, "A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method," in *BMC bioinformatics*, vol. 8, no. 9. BioMed Central, 2007, p. S4.

[82] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[83] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of artificial intelligence research*, vol. 11, pp. 95–130, 1999.

[84] R. Knappe, H. Bulskov, and T. Andreasen, "Perspectives on ontology-based querying," *International Journal of Intelligent Systems*, vol. 22, no. 7, pp. 739–761, 2007.

[85] J. Gu, W. Feng, J. Zeng, H. Mamitsuka, and S. Zhu, "Efficient semisupervised medline document clustering with mesh-semantic and global-content constraints," *IEEE transactions on cybernetics*, vol. 43, no. 4, pp. 1265–1276, 2013.

[86] C. Görg, H. Tipney, K. Verspoor, W. A. Baumgartner, K. B. Cohen, J. Stasko, and L. E. Hunter, "Visualization and language processing for supporting analysis across the biomedical literature," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 420–429.

[87] S. Shah and X. Luo, "Exploring diseases based biomedical document clustering and visualization using self-organizing maps," in *Proceedings of the 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2017, pp. 1–6.

[88] S. Shah and X. Luo, "Comparison of deep learning based concept representations for biomedical document clustering," in *Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, March 2018, pp. 349–352.

[89] X. Luo and S. Shah, "Concept embedding-based weighting scheme for biomedical text clustering and visualization," *Applied Informatics*, vol. 5, no. 1, p. 8, Nov 2018. [Online]. Available: https://doi.org/10.1186/s40535-018-0055-8

[90] E. MacIntosh, N. Rajakulendran, Z. Khayat, and A. Wise, "Transforming health: Shifting from reactive to proactive and predictive care," 2014.

[91] K. S. Yarnall, K. I. Pollak, T. Østbye, K. M. Krause, and J. L. Michener, "Primary care: is there enough time for prevention?" *American journal of public health*, vol. 93, no. 4, pp. 635–641, 2003.

[92] E. W. Campion, "A symptom of discontent," *New Englang Journal of Medicine*, vol. 344, no. 3, pp. 223–225, 2001, (Accessed: November 14, 2018). [Online]. Available: https://doi.org/10.1056/NEJM200101183440311

[93] K. S. Collins, *The Commonwealth fund survey of physician experiences with managed care*. Commonwealth Fund, 1997.

[94] Y. Shahar, S. Miksch, and P. Johnson, "The asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines," *Artificial intelligence in medicine*, vol. 14, no. 1-2, pp. 29–51, 1998.

[95] R. Serban, A. ten Teije, F. van Harmelen, M. Marcos, and C. Polo-Conde, "Extraction and use of linguistic patterns for modelling medical guidelines," *Artificial intelligence in medicine*, vol. 39, no. 2, pp. 137–149, 2007.

[96] M. Peleg and S. W. Tu, "Design patterns for clinical guidelines," *Artificial Intelligence in Medicine*, vol. 47, no. 1, pp. 1–24, 2009.

[97] S. Meystre and P. J. Haug, "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation," *Journal of biomedical informatics*, vol. 39, no. 6, pp. 589–599, 2006.

[98] R. Rosales, F. Farooq, B. Krishnapuram, S. Yu, and G. Fung, "Automated identification of medical concepts and assertions in medical text," in *AMIA Annual Symposium Proceedings*, vol. 2010. American Medical Informatics Association, 2010, p. 682.

[99] Q. Li and Y.-F. B. Wu, "Identifying important concepts from medical documents," *Journal of biomedical informatics*, vol. 39, no. 6, pp. 668–679, 2006.

[100] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," *Journal of Machine Learning Research*, vol. 8, pp. 2265–2295, October 2007.

[101] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.

[102] Y. Zhu, E. Yan, and F. Wang, "Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec," *BMC medical informatics and decision making*, vol. 17, no. 1, p. 95, 2017.

[103] S. Shah and X. Luo, "Extracting modifiable risk factors from narrative preventive healthcare guidelines for ehr integration," in *Proceedings of 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2017, pp. 514–519.

[104] HTACG, "HTML Tidy," 2018, (Accessed: November 9, 2018). [Online]. Available: http://www.html-tidy.org

[105] M. B. Barton, T. Miller, T. Wolff, D. Petitti, M. LeFevre, G. Sawaya, B. Yawn, J. Guirguis-Blake, N. Calonge, and R. Harris, "How to read the new recommendation statement: methods update from the us preventive services task force," *Annals of internal medicine*, vol. 147, no. 2, pp. 123–127, 2007.

[106] L. Richardson, "Beautiful soup," *Crummy: The Site*, 2013, (Accessed: October 13, 2018). [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/

[107] C. Maloney, E. Sequeira, C. Kelly, R. Orris, and J. Beck, "Pubmed central," 2013, (Accessed: October 3, 2018). [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK153388/

[108] H. L. Behforouz, P. K. Drain, and J. J. Rhatigan, "Rethinking the social history," *New England Journal of Medicine*, vol. 371, no. 14, pp. 1277–1279, 2014.

[109] P. Cavanagh, J. Derr, J. Ulbrecht, R. Maser, and T. Orchard, "Problems with gait and posture in neuropathic patients with insulin-dependent diabetes mellitus," *Diabetic Medicine*, vol. 9, no. 5, pp. 469–474, 1992.

[110] C. Macgilchrist, L. Paul, B. Ellis, T. Howe, B. Kennon, and J. Godwin, "Lower-limb risk factors for falls in people with diabetes mellitus," *Diabetic Medicine*, vol. 27, no. 2, pp. 162–168, 2010.

VITA

VITA

Setu Shah

Education

- Master of Science, December 2018
  Purdue University
  Department of Electrical and Computer Engineering
- Bachelor of Engineering, June 2016
  Savitribai Phule Pune University
  Department of Electronics and Telecommunication Engineering

Teaching Experience

- ECE 57000 – Programming Languages for Artificial Intelligence
  Fall 2017 and Fall 2018
- ECE 60800 – Computational Models and Methods
  Spring 2018

Academic Awards

- Outstanding Electrical and Computer Engineering Graduate Student
  Awarded by the Department of Electrical and Computer Engineering, in March, 2018.
  Purdue University, Indianapolis.
- IUPUI Graduate Travel Fellowship
  Awarded by the IUPUI Graduate Office for attending IEEE Healthcom 2017 in Dalian, China in November, 2017.
  Purdue University, Indianapolis.

- IUPUI Graduate Professional Educational Grant

  Awarded by the IUPUI Graduate and Professional Student Government for attending AAAI '18 in New Orleans, USA in February, 2018.

  Purdue University, Indianapolis.

- Best Final Year Project

  Awarded by the Department of Electronics and Telecommunications in April, 2017.

  Pune Vidyarthi Griha's College of Engineering and Technology,

  Savitribai Phule Pune University.

PUBLICATIONS

1. S. Shah, X. Luo, S. Kanakasabai, R. Tuason, and G. Klopper, "Neural networks for mining the associations between diseases and symptoms in clinical notes," in *Health Information Science and Systems*, vol. 7, no. 1, 2018.

2. X. Luo and S. Shah, "Concept Embedding based Weighting Scheme for Biomedical Text Clustering and Visualization," in *Applied Informatics*, vol. 5, no. 1, 2018.

3. S. Shah and X. Luo, "Comparison of Deep Learning based Concept Representations for Biomedical Document Clustering," in *Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 349-352.

4. S. Shah and X. Luo, "Exploring diseases based biomedical document clusteringand visualization using self-organizing maps," in *Proceedings of the 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2017, pp. 1-6.

5. S. Shah and X. Luo, "Extracting Modifiable Risk Factors from Narrative Preventive Healthcare Guidelines for EHR Integration," in *Proceedings of 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2017, pp. 514-519.

6. S. Shah and X. Luo, "Biomedical Document Clustering and Visualization based on the Concepts of Diseases," in *Proceedings of KDD 2017, Data Driven Discovery workshop*.