Visual Analytics on Biomedical Dark Data

Shashwat Aggarwal [a], Ramesh Singh [b]

[a] University of Delhi

[a] National Informatics Center

1. INTRODUCTION

In today's data centralized world, the practice of data visualization has become an indispensable tool in numerous domains such as *Research*, *Marketing*, *Journalism*, *Biology* etc. Data visualization is the art of efficiently organizing and presenting data in a graphically appealing format. It speeds up the process of decision making and pattern recognition, thereby enabling decision makers to make informed decisions.

With the rise in technology, the data has been exploding exponentially, and the world's scientific knowledge is accessible with ease. There is an enormous amount of data available in the form of scientific articles, government reports, natural language, and images that in total contributes to around 80% of overall data generated as shown in an excerpt from The Digital Universe [1] in Figure 1. However, most of the data lack structure and cannot be easily categorized and imported into regular databases. This type of data is often termed as Dark Data. Data visualization techniques proffer a potential solution to overcome the problem of handling and analyzing overwhelming amounts of such information. It enables the decision maker to look at data differently and more imaginatively. It promotes creative data exploration by allowing quick comprehension of information, the discovery of emerging trends, identification of relationships and patterns etc.

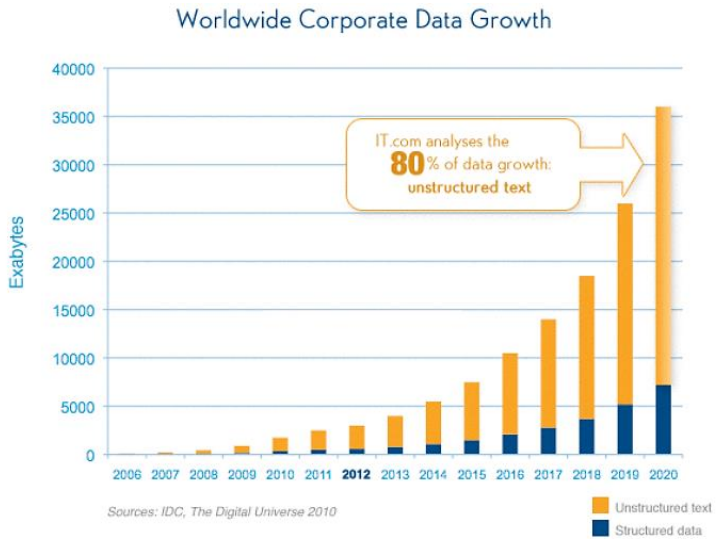


Figure 1. Worldwide growth of corporate data categorized by structured and unstructured/dark data over the past decade.

Data visualization techniques proffer a potential solution to overcome the problem of handling and analyzing overwhelming amounts of such information. It enables the decision maker to look at data differently and more imaginatively. It promotes creative data exploration by allowing quick comprehension of information, the discovery of emerging trends, identification of relationships and patterns etc. Over time, data visualization techniques have been used in a great number of domains, but the domain that has received major attention recently is text. Professional users like scholars, research institutions, and funding agencies have become more and more interested in the textual domain. There are numerous techniques used for visualization of text as summarized in Figure 2. All these techniques can be categorized into various subdomains depending on their use case, for instance, geometric, clustering based, graph-based etc.

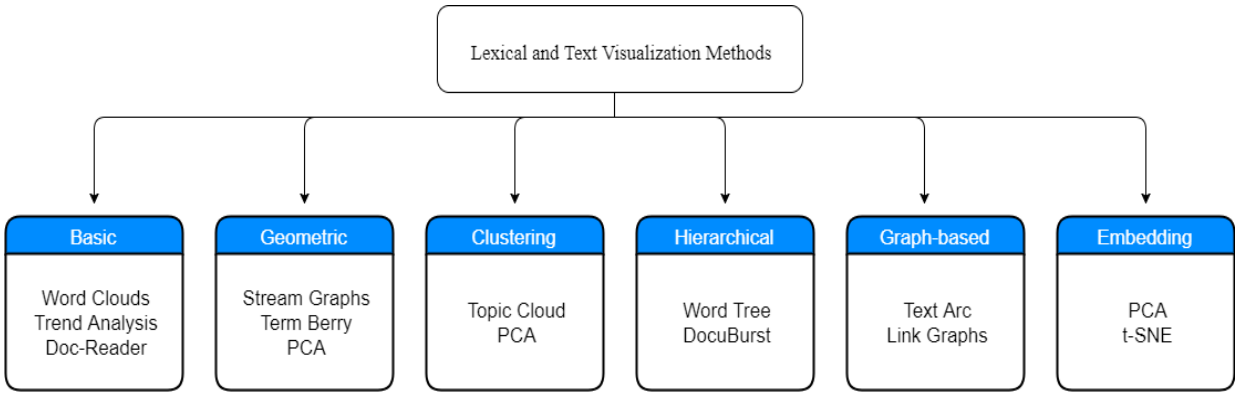


Figure 2. Hierarchy of Lexical and Text Visualization Techniques used.

1.1 INTRODUCTION TO THE DEEP DIVE PROJECT

Deep Dive [2], a project led by Christopher Ré and his group at Stanford has proved to be beneficial in extracting value out of dark data. It is a dark data management system developed to bring dark data to light by creating structured data (SQL tables) from unstructured information (text documents) and integrating such data with an existing structured dataset. Deep Dive is primarily used to extract sophisticated relationships between entities and make inferences about facts involving those entities. It helps in processing a wide variety of dark data by organizing results in a dataset. With the data in a dataset, a variety of standard tools that consume structured data; e.g., visualization tools like Tableau or analytics tools like Excel can be utilized to extract value from the data. In this paper, we try to extract value and get insights from the PubMed dataset provided by Deep Dive in a Deep Dive-ready DB Dump format which can be loaded directly into a RDBMS like MySQL or PostgreSQL etc.

1.2 MISCELLANEOUS DARK DATA ANALYTICS PROJECTS

1.2.1 Snorkel

Snorkel [3] is a system for rapidly creating, modeling, and managing training data, that enables users to train state-of-the-art models without hand labeling any training dataset. It focuses on accelerating dark data extraction for various domains like health, finance etc. in which large labeled training sets are not easily available. Snorkel is based on a new data programming paradigm, in which the user focuses on writing a set of labeling functions, that programmatically label data. The resulting labels are noisy, but Snorkel automatically models this process - learning, essentially, which labeling functions are more accurate than others and then uses that to train an end-to-end model. By modeling the noisy training set created from those potentially low-quality labeling functions defined by the user, it can create high-quality end models. We see Snorkel as providing a general framework for many weak supervision techniques, and as defining a new programming model for weakly-supervised machine learning systems.

1.2.2 MacroBase

MacroBase [4], a data analytic monitoring engine developed by Stanford Future Data Systems and the Stanford DAWN Project, available under the permissive Apache V2.0 License, is designed to prioritize human attention in large-scale datasets and data streams. Unlike a traditional analytics engine, MacroBase is specialized for one task: finding and explaining unusual or interesting trends in data. For large data volumes, manual inspection of data is very difficult and untenable. MacroBase counteracts this problem by providing an efficient, accurate and modular approach to highlight and aggregate important and unusual trends in the data. MacroBase can deliver very high speedups over other alternatives by optimizing the combination of explanation and classification tasks and by leveraging a new reservoir sampler for fast data streams.

1.2.3 Apache MADlib

Apache MADlib [5] is an open-source machine learning project in SQL developed for scalable in-database analytics. It supports various open-source databases such as Postgres, Greenplum, and Apache HAWQ for efficiently storing and handling large datasets. It provides data-parallel implementations of mathematical, statistical and machine learning methods for both structured and unstructured data. MADlib operates on data locally through powerful implementations on various machine learning, graph, statistical algorithms like classification, regression, clustering, topic modeling, descriptive statistics etc.

1.3 OVERVIEW OF PUBMED DATASET

“PubMed comprises of more than 28 million citations and abstracts for biomedical literature from MEDLINE, life science journals, and online books.” (https://www.ncbi.nlm.nih.gov/pubmed/, 2018). It is an open source database developed and maintained by the National Center for Biotechnology Information (NCBI). In addition to free access to MEDLINE, PubMed also provides links to free full-text articles provided by PubMed

Central and third-party websites, advanced search capabilities, clinical queries search filters, special query pages and other related resources. “PubMed is a key information resource in biological sciences and medicine primarily because of its wide diversity and manual curation. It comprises of an order of three billion bases of human genome, rich meta-information (e.g. MeSH terms), detailed affiliation, etc., summing up to a total of 70GB database.” [6]

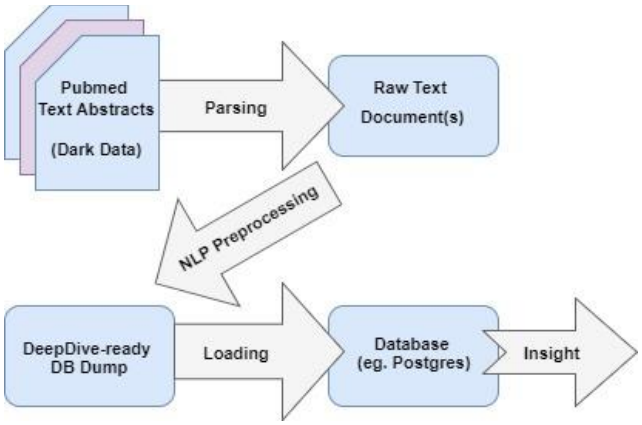


Figure 3. Pre-processing pipeline adopted by Deep Dive to convert PubMed dataset into a structured format.

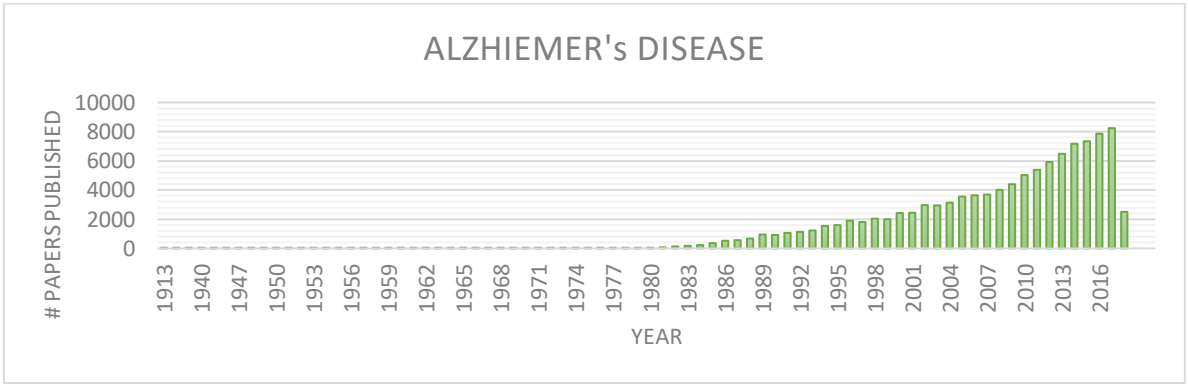
As of 1 April 2020, PubMed has more than 30 million records from 5500 journals, dating back to 1966, with the earliest publication available from the year 1809. “13.1 million of PubMed's records are listed with their abstracts, and 14.2 million articles have links to full-text with around 500,000 new records being added each year. Around 12% of the records in PubMed correspond to cancer-related entries, which have grown from 6% in the 1950s to 16% in 2016. Other significant proportions of records correspond to “Chemistry” (8.69%), “Therapy” (8.39%) and "Infection" (5%).” [7]. PubMed provides efficient methods to search the database by using author names, journal names, keywords or phrases, MeSH terms or any combination of these. It also enables users to download the fetched citations and abstracts for queried terms in various formats such as plain text form (both Summary and Abstract), XML form, PMID form, CSV form and MEDLINE form. The results are sorted according to one of the followings: publication date, author name, title of paper or journal name.

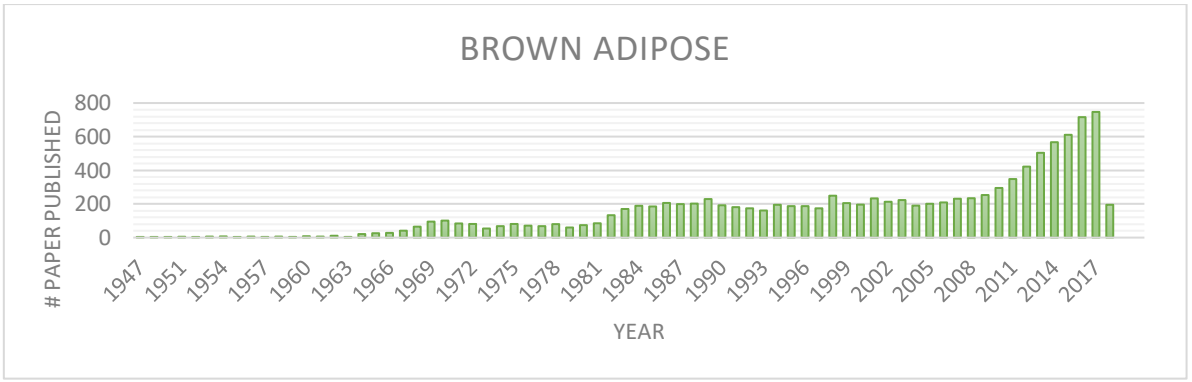
Table 1: Fundamental Statistics of PubMed Dataset as on March 2014; (http://deepdive.stanford.edu/opendata/#pmc-oa-pubmed-central-open-access-subset, 2014)

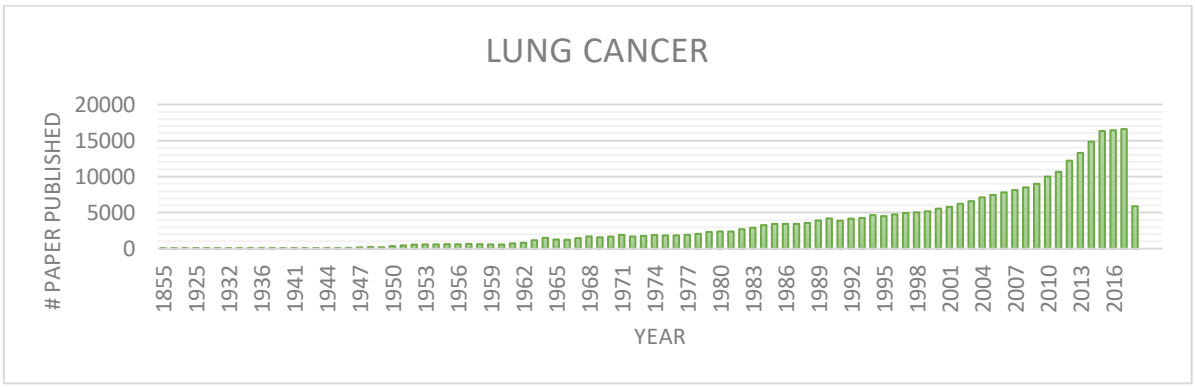
Size	70 GB	**# Sentences**	110 Million
# Documents	359,324	**# Distinct Entities**	412,593,720
# Words	2.7 Billion	**# Distinct Subjects**	412,593,720
# Distinct Literals	1,842,783,647	**# Distinct Objects**	436,101,294

Deep Dive [2] is an open source dark data analytical platform developed by Christopher Ré and his group at Stanford which hosts pre-processed copies of several open source text

databases (e.g. PubMed). The processing pipeline used by Deep Dive to create a structured database from the research documents present in PubMed is shown in Figure 3. The pipeline consists of several stages, namely: a) scraping of data in HTML/XML format (Parsing), b) striping it into plain text, c) applying basic NLP pre-processing such as tokenization, stemming, POS tagging etc to clean up the data, and d) loading the Deep Dive ready DB dump in a structured database like Postgres. The fundamental statistics of PubMed database provided by Deep Dive are reported in Table 1. These statistics help in providing an abstract overview of the database like for example the frequency count at various levels (literals, words, sentences, documents).







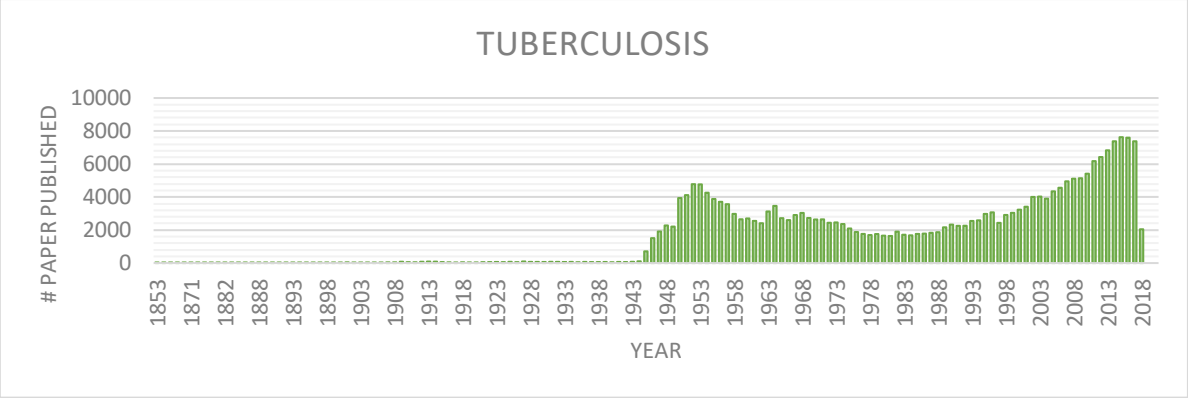


Figure 4. Timeline of PubMed Articles for queried topics: Alzheimer's Disease, Brown Adipose, Lung Cancer and Tuberculosis, respectively.

The dataset we use is pre-processed by Deep Dive by converting the unstructured abstracts into a structured format by following a basic pipeline, as illustrated in Figure 3, namely: a) scraping of data in HTML/XML format (Parsing), b) striping it into plain text, and c) applying basic NLP pre-processing such as tokenization, stemming, POS tagging etc to clean up the data. Some basic statistics of PubMed dataset are given in Table 2, that summarizes the high-level structure of the dataset.

Table 2: Basic Statistics of PubMed Dataset as on March 2014; [8].

Size	70 GB	**# Sentences**	110 Million
# Documents	359,324	**# Distinct Entities**	412,593,720
# Words	2.7 Billion	**# Distinct Subjects**	412,593,720
# Distinct Literals	1,842,783,647	**# Distinct Objects**	436,101,294

Each record in the dataset is a sentence. The first column is doc_id, which is a text string specifying the ID of a document; the second column is sentence_id, which is an integer specifying the ID of a sentence. These two columns together form the primary key. The remaining columns are tokens containing tokenized words of PubMed abstracts, their lemmas, part of speech (POS) tags and Named Entity Recognition (NER) tags respectively. Table 3 provides an overview of the contents of the PubMed dataset.

Table 3: Summary of PubMed Dataset

Column	**Datatype**
doc_id	text
sentence_text	text
tokens	text[]
lemmas	text[]
pos_tags	text[]
ner_tags	text[]
dep_types	text[]
dep_tokens	int[]

2. VISUAL EXPLORATION OF PUBMED

We use the rich corpus of the PubMed which comprises of more than 28 million citations and abstracts for biomedical literature from MEDLINE, life science journals, and online books. [NCBI]. It is an open source database developed and maintained by NCBI. In addition to free access to MEDLINE, PubMed also provides links to free full-text articles provided by PubMed Central and third-party websites, and other facilities such as clinical queries search filters, special query pages, etc. PubMed is a key information resource in biological sciences and medicine primarily because of its wide diversity and manual curation. It comprises of an order of three billion bases of the human genome, rich meta-information (e.g., MeSH terms), detailed affiliation, etc., summing up to a total of 70GB database. [Roberts, 2001]. As of 1 June 2019, PubMed has more than 28 million records from 5500 journals, dating back to 1966, with the earliest publication available from the year 1809. PubMed supports efficient methods to search the database by using author names, journal names, keywords, and phrases, or any combination of these. It also enables users to download the fetched citations and abstracts for queried terms in various formats such as plain text form (both Summary and Abstract), XML form, PMID form, CSV form and MEDLINE form.

In Table 2, we present some basic statistics to describe the PubMed database the size of the database, number of documents present inside the database, number of sentences and terms summed across all the documents present in the database. We also compute the number of distinct literals, entities, subjects and objects present within PubMed using a natural language processing (NLP) based pipeline: a) scraping of data in HTML/XML format (i.e., Parsing), b) striping it into plain text, and c) applying NLP pre-processing techniques such as sentence

segmentation and word tokenization to compute the number of distinct literals, stemming and lemmatization for normalization of terms and phrases, POS (part of speech) tagging and Dependency Parsing to identify subjects and objects within each sentence, and finally, NER (named entity recognition) and Coreference Resolution to identify the entities present inside the database.

In order to visually explore and analyse the biomedical research document present within PubMed, we firstly use Word Clouds, a visual representation of text data, typically used to depict the prominent words across the data with the prominence of words measured relative to their frequency counts, to summarize and get a general overview of the contents of PubMed documents. Word clouds are commonly used in the field of text mining and information retrieval for abstracting, visualizing, and comparing textual databases and have demonstrated to be useful in various research settings. We use the open source tool, Wordle [Jonathan Feinbergz, 2014] to create cloud visualization to summarize the contents of the PubMed corpora.

Figure 5 presents the word cloud on the PubMed dataset. Words like *patient*, *cells*, *data*, *cancer*, *gene* are more prominent signifying a substantial proportion of study related to cancer and genomics being conducted compared to other domains. Furthermore, various words such as *DNA*, *tumour*, *acid*, and *receptors* highlight the other significant areas where research has been done or is going on. One interesting fact that can be observed from these clouds is the prominence of words, *High Population* and *Children* providing a high-level indication of the major disease cause and majority group affected by those diseases. Alongside the word cloud, in Figure 1 we also plot the streamline graph for the seven most frequent terms across the database depicting the variation of their relative frequency distribution across the set of documents.

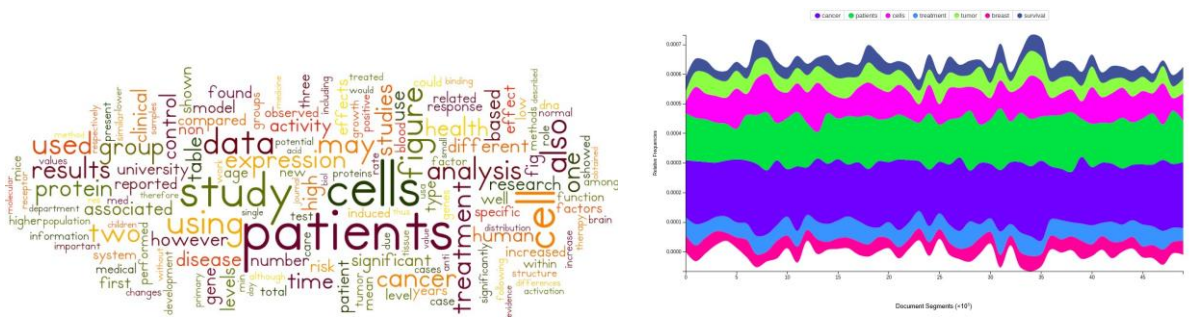


Figure 5: Left: Word Cloud of the PubMed Corpora. Right: Streamline Graph for top 7 most frequent words in PubMed.

Another widely used text representation technique which apart from considering the term frequencies, also encodes their importance inside a document, is *TF* ×*IDF*. The TF gives the frequency of the term within a particular document, and the IDF gives the inverse document frequency of a term, i.e., a measure of the importance of term across several documents. *TF*×*IDF*, term frequency-inverse document frequency, is a statistical measure used to evaluate how important a word is to a document in a collection. The importance is directly proportional to the term frequency of the word in the document but is offset by the frequency of the word in the corpus.

In Figure 6, we visualize the *TF* ×*IDF* plot computed over 50,000 full text articles retrieved from PubMed Central. The x-axis represents the normalized decimal term representations while the y axis gives their corresponding *TF* ×*IDF* scores. All the *TF* ×*IDF* scores are calculated by calculating the term document matrix using Sklearn *TF* ×*IDF* vectorizer [Pedregosa et al., 2011]. The top 2 most significant components of the corresponding vector representation of the terms obtained from *TF* ×*IDF* vectorizer are computed using PCA (principal component analysis) dimensionality reduction technique [Jolliffe, 2002] and are visualized in Figure 6. We extract a number of relevant words (A.K.A. keywords) in accordance to the *TF* ×*IDF* scores computed earlier, with the number determined by a certain threshold score. We visualize the distribution of the number of keywords extracted via *TF* ×*IDF* score for different document lengths. score for different document lengths. For space constraints and sparsity reasons, we binned the document lengths by quartile (i.e., the bins are not of equal range but contains the same number of documents 25% each). Figure 3 displays the box-and-whisker plot computed over 50,000 full text articles from PubMed Central showing the distribution of keywords across different document lengths (binned by quartile). Figure 7 also shows the swarm plot, which gives a better representation of the distribution of keywords, visualizing all observations along with the underlying distribution.

As it can be seen from the plots, the median value of number of keywords increase with the document length till the third quartile after which there is a drop in the median value, which is mainly because of *TF* × *IDF* scoring and is intuitive since a given word is more likely to be found in a relatively longer document as compared to a shorter document but is not necessarily a keyword. We can also observe that the variation of number of keywords is less in first and the last bins as compared to the second and the third bin, indicating that documents which are either too short or long approximately contain a constant number of relevant words while moderate length documents have a high variability in their distribution of the number of keywords.

Figure 6: *TF* ×*IDF* Score Distribution Plot over 50,000 full-text articles retrieved from PubMed Central.

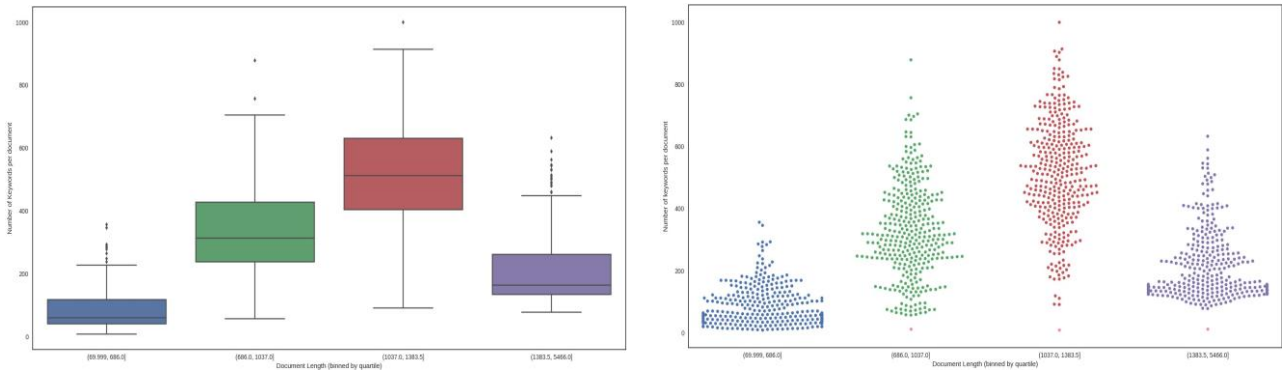


Figure 7: Left: Box-and-Whisker plot and Right: Swarm Plot, computed over 50,000 full text articles from PubMed Central showing the distribution of keywords across different document lengths (binned by quartile.)

Although the raw term frequencies, word cloud visualizations and *TF* ×*IDF* scores work quite well in practice (e.g., in summarization of general overview of the database), however, these techniques fail to capture the ordered relationship between terms and sentences. To understand the contextual relationship between the different terms we use DocuBurst. DocuBurst [Collins et al., 2009] is an online document visualization tool used for creating interactive visual summaries of documents, exploring keywords to uncover document themes or topics, investigating intra-document word patterns, such as character relationships, comparing documents, etc. which takes advantage of the human-created structure in lexical databases.

DocuBurst visualize nouns and phrases in a hierarchically structured manner centered around a root word which is selected either as the most prominent word in the database or as queried by the user. DocuBurst uses a pre-existing ontology, WordNet [Fellbaum, 1998], to group words having related meanings together. It creates a radial, space-filling layout of hyponymy (IS-A relation) with interactive techniques of zoom, filter, and details-on-demand for the task of document visualization.

We generate DocuBurst graphs over 50,000 full-text articles that we retrieved earlier from PubMed Central. We limit our database to this subset of PubMed to meet the software and memory requirements of the tool utilized. Alongside the DocuBurst graphs, both word score and word clouds for the selected word (shown in pink) and its co-occurring words are also displayed to summarize the content better. Figure 8 shows the DocuBurst graph with part chosen as the root word. DocuBurst hierarchically structures the radially surrounding hyponyms such as organ, structure, tissue, and system around the root word. Each surrounding hyponym is further sub-structured with its related hyponyms, thereby forming chains of correlated keyword terms which reveal the coherent document themes present in the database. Along with the DocuBurst graph, Word Clouds depicting words having strong correlations with terms on the DocuBurst graph are also shown. From these word clouds, we can infer that words like *Cancer* and *Hypertension* are highly correlated with terms related to body parts. The word *Cancer* can be seen to be highly connected with term *tissues* and also with other terms like

University, *Research* and specific country names such as *China*, *Germany* indicating significant work related to cancer research being carried out by Universities in these countries. Similarly, it can also be observed that terms like *Hypertension* are highly correlated to terms highlighted in pink which are mostly related to *mind* thus, indicating some of the body organs affected due to hypertension.

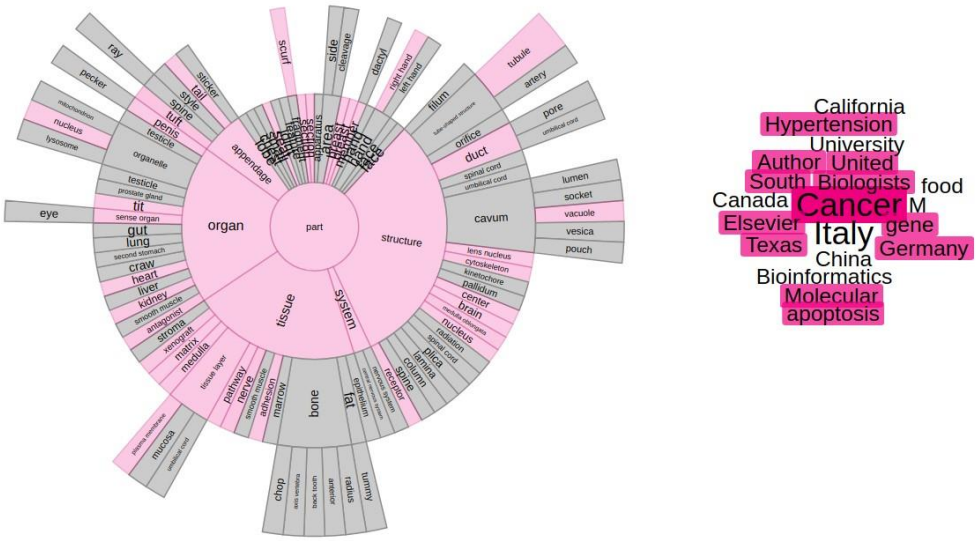


Figure 8: DocuBurst plot on the PubMed database with *part* as the root word.

3. LEXICAL AND TEXT VISUALIZATION METHODS

3.1 WORD CLOUDS

Clouds are graphical visualizations of the word frequencies where words within a collection of text are drawn with size relative to their frequencies. The larger the word in the cloud, the more common it is in the document(s). In the last few years, word clouds have become a standard tool for abstracting, visualizing, and comparing text documents and have been demonstrated to be useful in various research settings.

We use Word Clouds to summarize the large corpora of Pubmed and get a high-level idea about the contents of Pubmed articles. Word Clouds are generated on the entire corpora of Pubmed effectively utilizing the techniques of Clustering, Hashing, Indexing and Merging to efficiently pre-process and scale up to the size of corpora. Open Source tools, Wordle and WordArt, are used to create word cloud visualizations in an appealing way to summarize the contents of the Pubmed Abstract Corpora. We follow a simple pipeline for the generation of word clouds on the Pubmed dataset. Firstly, the collection of words is extracted from the *tokens* column of the dataset, common stop-words such as "a", "the", "is" are removed from the collection and the remaining words are grouped by their stems using the Porter Stemming Algorithm. The most common variation of the word is used in the final word cloud.

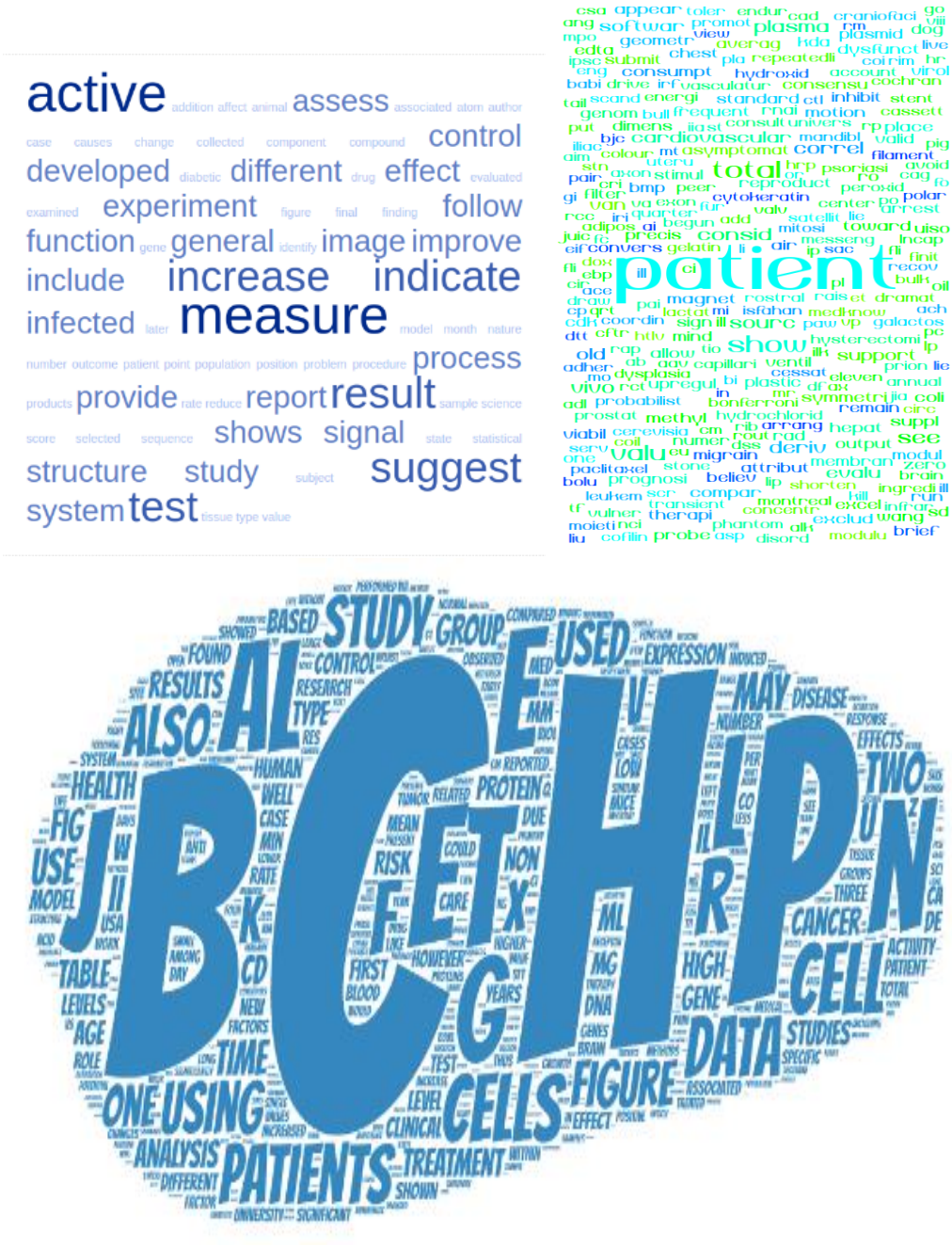


Figure 9(a): Word Clouds on 10 GB Subset of Pubmed Database for 50(left), 200(center), 500(right) most frequent words.

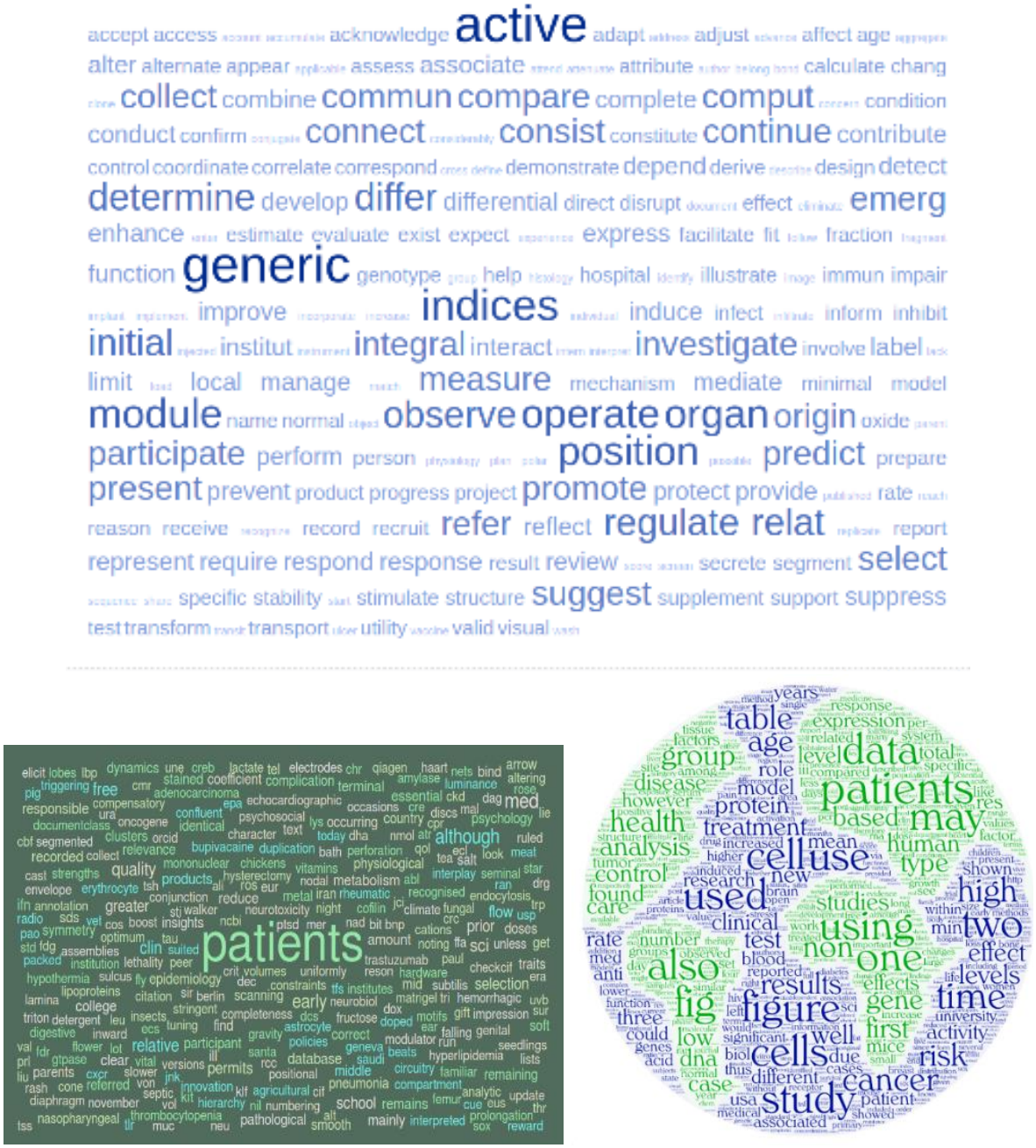


Figure 9(b): Word Clouds on 30 GB Subset of Pubmed Database for 50(left), 200(center), 500(right) most frequent words.

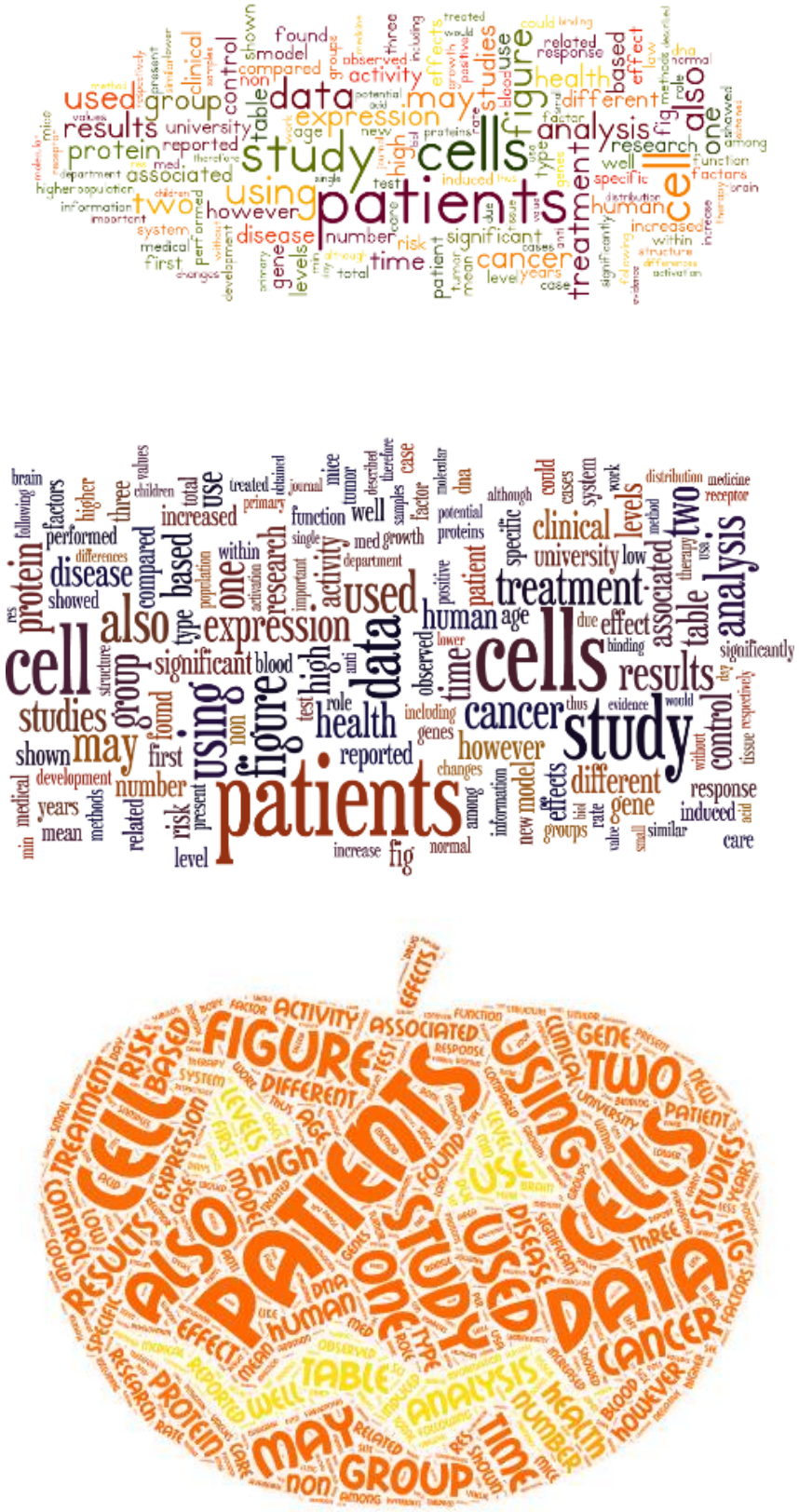


Figure 9(c): Word Clouds on 30 GB Subset of Pubmed Database for 50(left), 200(center), 500(right) most frequent words.

Word clouds of the Pubmed Corpora consists of the most important words across all the documents. Each word is printed in a given font style and scaled by a factor roughly proportional to its importance (i.e. frequency count). The printed words are arranged without overlap and tightly packed into some shape (usually a rectangle or a masked filter). Word Clouds are generated for three different configurations namely a) number of words sampled (50, 200, 500), b) minimum length of a word (greater than 0, greater than 3) and c) Image Filter used (plane rectangular or mask).

From the different word cloud settings shown in figure 9, it can be inferred that word clouds on about 10 GB of Pubmed dataset are somewhat gibberish, but they tend to become more meaningful with the size of the dataset. Words like *patient, cells, data, cancer, gene* become more prominent signifying a larger proportion of study related to cancer and genomics being conducted compared to other domains. Furthermore, various words such as *DNA, tumour, acid* and *receptors* highlight the other major areas where research studies have been and are carried out. One interesting fact that can be observed from the clouds is prominence of words such as *High Population* and *Children* giving a high-level picture about the major cause of diseases and majority group affected by these diseases.

3.2 TOPIC CLOUDS

A Topic Cloud is a pie chart consisting of topic slices, where each slice contains the most important words in that topic. The relative prominence of words/topics are made explicit by drawing the words/topics in sizes that are proportional to their importance in the document. A topic cloud is like a word cloud giving frequency of words or phrases, but the major difference that a topic cloud offers as compared to word cloud is the semantic grouping of words under corresponding topics, hence providing greater insights into the textual data with refined granularity.

We use Topic clouds as described in [Document Visualization using Topic Clouds Shaohua Li et al] to gain insights on the Pubmed dataset. We generate topic clouds using two different algorithms namely a) kmeans which is a popular clustering algorithm and b) TopicVec as described in [5]. The topic proportion in both cases is defined as the proportion of words in a cluster.

For K-means scenario, we choose k = 10 for topic clustering and conceptnet-numberbatch embedding, which is a combination of word2vec and glove for converting words to dense vectors to generate topic clouds for two different settings (top 1000 and 10000 words respectively) as shown in Figure 10(a and b) respectively. Figure 10(c, d) shows the topic clouds derived by TopicVec. We tune the *gamma* (γ) parameter of TopicVec, i.e., the maximal magnitude γ of topic embeddings according to the original paper to 3 and 5, respectively. In Figure 10(c), we find out that certain topics are highly similar, and all topics have similar proportions. In contrast, in Figure 10(d), clustering of words in different coherent topics tends to improve, and the topic proportions gradually decrease clockwise. The two topic clouds reveal that $\gamma = 5$ to be a better setting as compared to 3.

All the topic clouds elegantly summarize the contents of Pubmed articles, with each slice depicting the most important words in that topic. Words like *patients, cancer, disease, treatment* etc. are grouped under one topic whereas words such as *data, study, research* depict

another topic. On comparison of topic clouds for kmeans and TopicVec, we find that the topics produced by TopicVec are more even, with similar sizes; while the topics produced by K-means are more disproportionate, hence we say that TopicVec groups words into coherently less noisy topics as compared to kmeans.

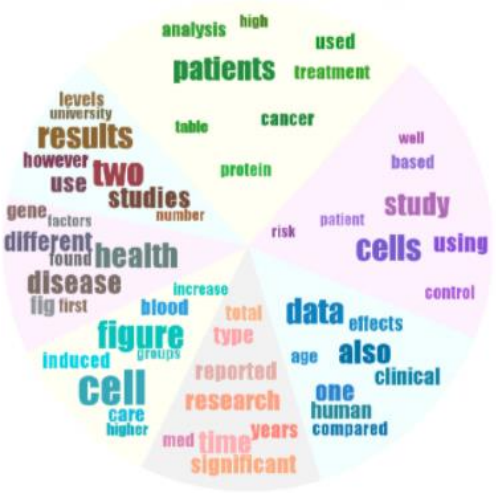


Figure 10(a): Topic cloud with Kmeans for top 1000 words (# Topics = 7)

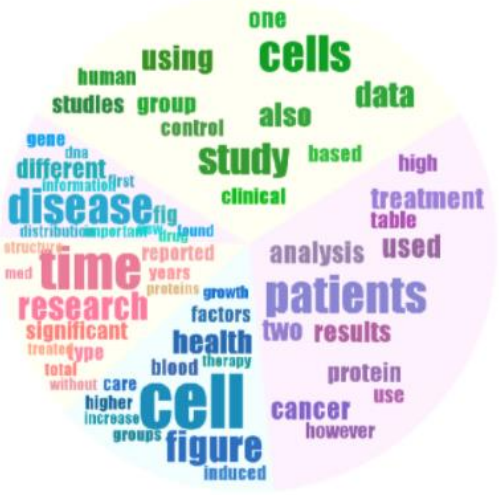


Figure 10(b): Topic cloud with Kmeans for top 10000 words (# Topics =7)

Figure 10(c): Topic Cloud with TopicVec (lambda = 3)

Figure 10(d): Topic Cloud with TopicVec (lambda = 5)

3.3 DOCUBURST

DocuBurst is an online document visualization tool used for creating interactive visual summaries of documents, exploring keywords to uncover document themes or topics, investigating intra-document word patterns, such as character relationships, comparing documents etc. which takes advantage of the human-created structure in lexical databases. It creates a radial, space-filling layout of hyponymy (IS-A relation) with interactive techniques of zoom, filter, and details-on-demand for the task of document visualization. DocuBurst is overlaid with occurrence counts of words in a document of interest to provide visual summaries at varying levels of granularity.

The DocuBurst visualizes hierarchically structured nouns. A root is the starting point or centre word of the DocuBurst. It defines the scope of our investigation. Techniques like DocuBurst use a pre-existing ontology, Wordnet [1], to group words having related meanings. The flowchart shown in figure 11(a) summarizes the method to interpret DocuBurst visualizations. Figure 11(b) displays the DocuBurst graph on the entire Pubmed dataset with *part* as the root word and the radially surrounding hyponyms such as organ, structure, tissue, system etc. Figure 7(b) also displays the filtered view of the above DocuBurst graph with *organ, structure and tissue* as the filtered root words.

Figure 11(c) shows the DocuBurst graphs for Pubmed query on *fibromyalgia* with *condition* and *disorder* as the root words respectively. Words related to *condition* depicts various conditions and illness whereas the word *disorder* visualize various disorders experienced by the patients of *fibromyalgia*. Both word score and word clouds for the selected word (shown in *pink*) and its cooccurring words are displayed to better summarize Pubmed content through DocuBurst visualizations. Finally, figure 11(d) displays a comparison between the documents of Pubmed returned from queries on *microbiology* and *fibromyalgia* and between words from Pubmed documents on microbiology and English Words respectively. Words displayed in green belongs to one category (viz *fibromyalgia, English words*) whereas words in blue belong to other the category (*microbiology*). Words displayed in red/orange are neutral and occur similarly in both types of texts.

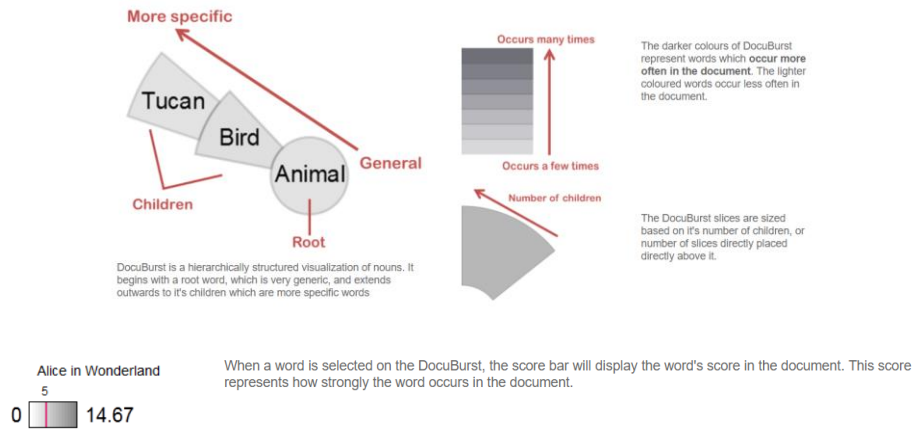
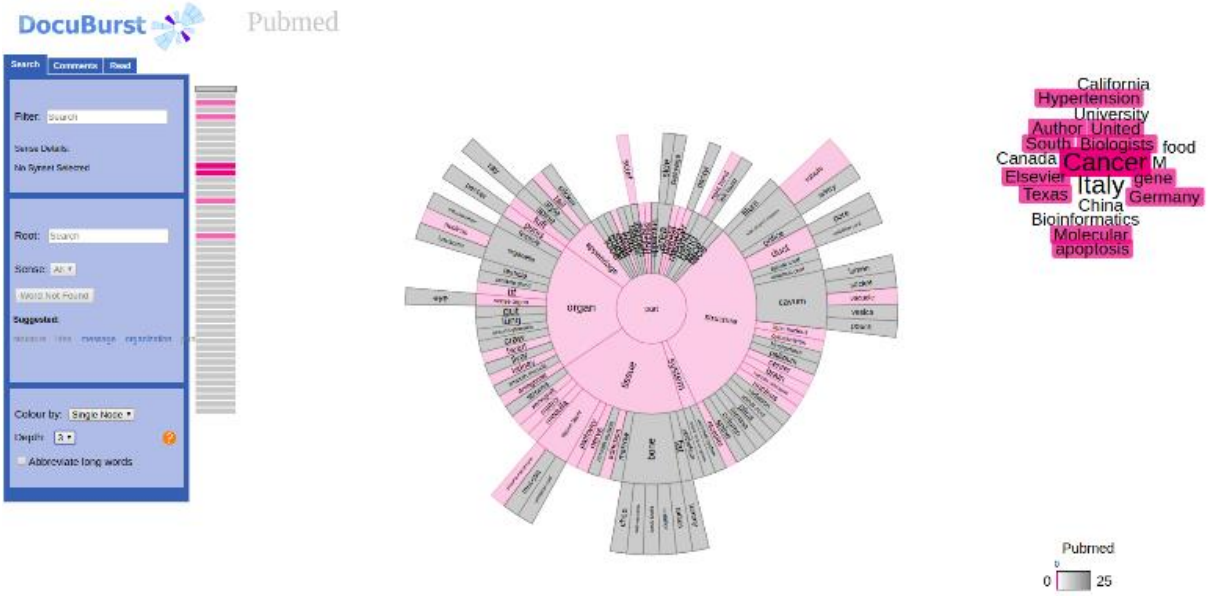
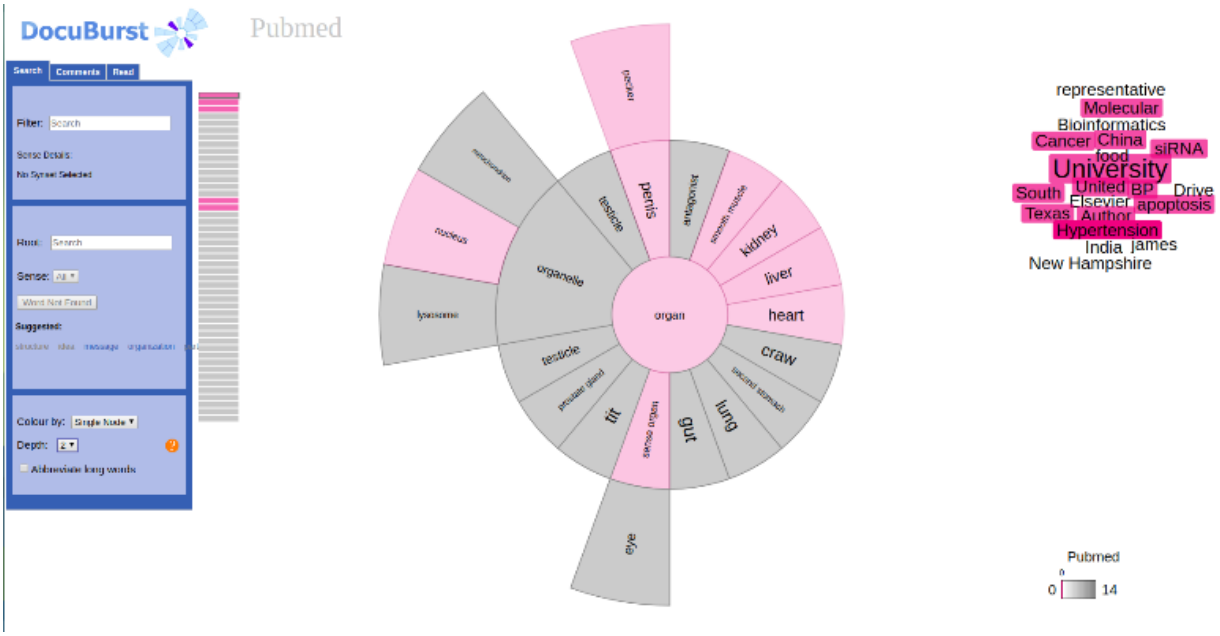


Figure 11(a): Flowchart describing interpretation of Visualizations (top) and Scoring of Words in DocuBurst (bottom).

Although DocuBurst is a good visualization to get a fundamental understanding about unstructured texts is has a major problem, Wordnet. The categories in Wordnet, like most ontologies, are not completely intuitive - e.g., few users will naturally understand the distinction between an “entity” and an “object” and many words such as names of various drugs, chemicals and diseases are missing from it.





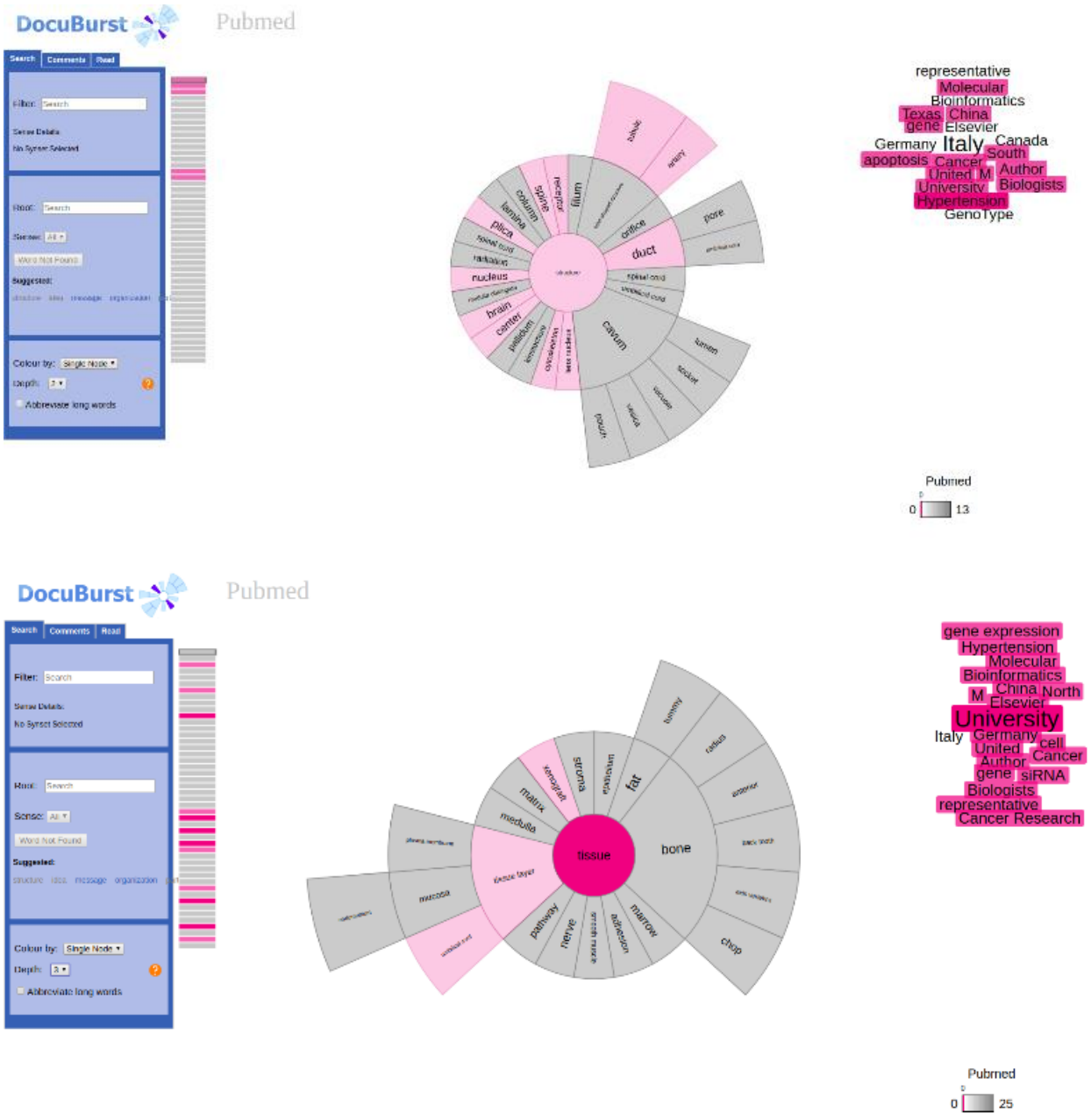
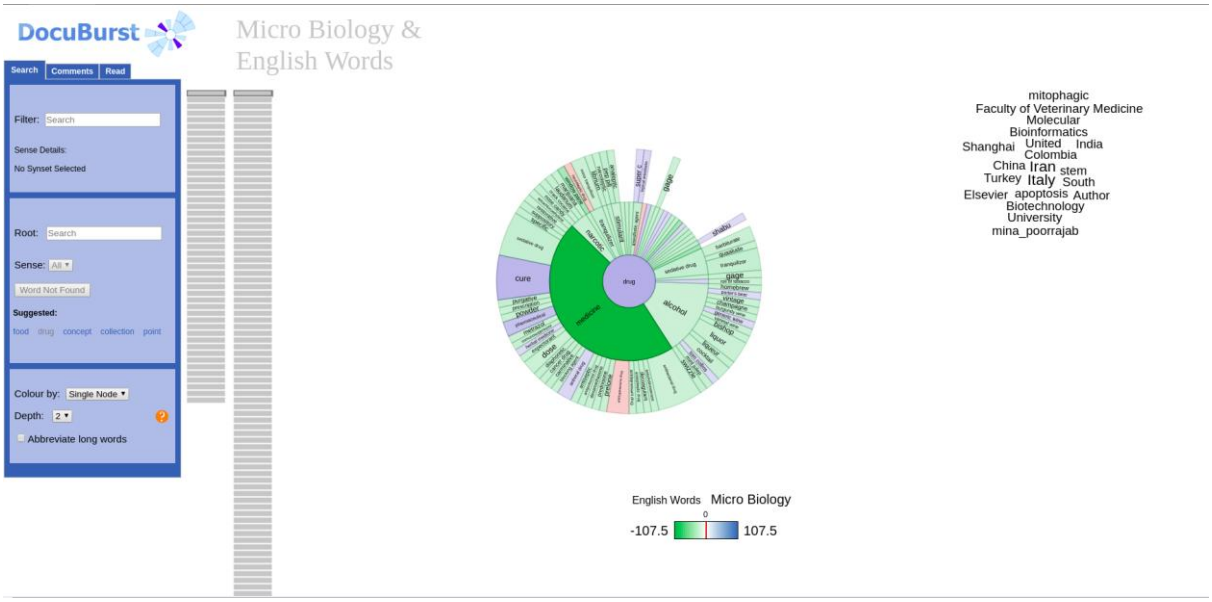


Figure 11(b): DocuBurst graph on the entire Pubmed dataset with *Part* as the root word and *Organ, Structure and Tissue* as filtered root words respectively.

Figure 11(c): DocuBurst graphs for Pubmed query on *Fibromyalgia* with condition (left) and disorder (right) respectively



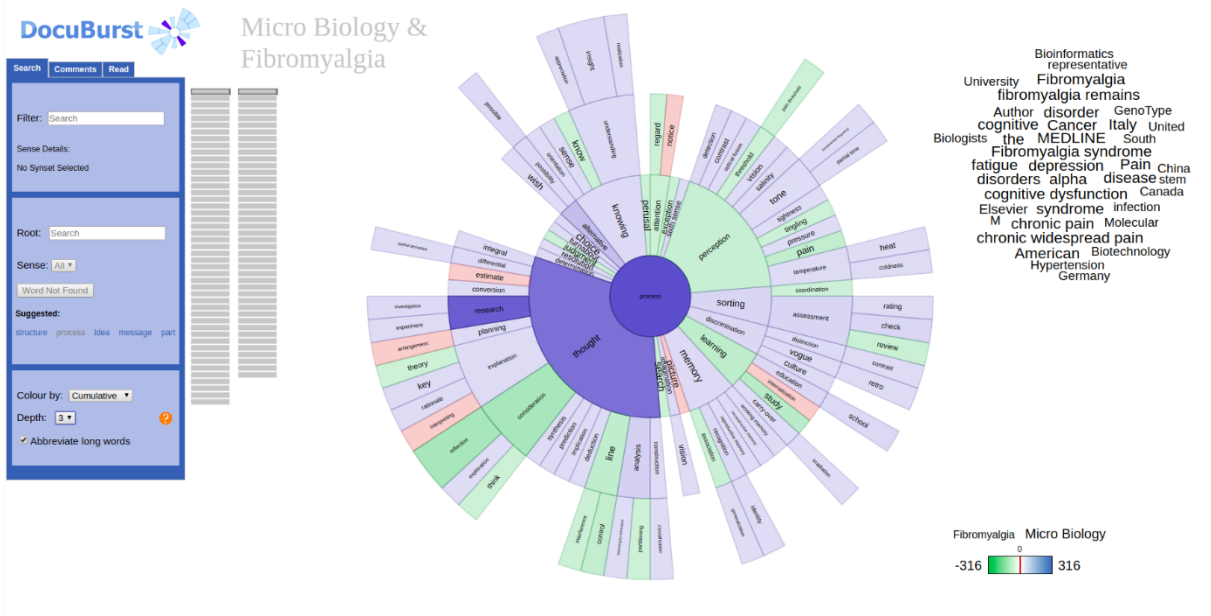


Figure 11(d): Comparison between the documents of Pubmed returned from queries on *Microbiology* and F*ibromyalgia* (top) and between words from Pubmed documents on *Microbiology* and English Words (bottom) respectively.

3.4 WORD TREES

Word trees are hierarchical graphs that show where a chosen word or phrase appears in the body of a collection of texts. It lets you pick a word or phrase and shows you all the different contexts in which it appears. The contexts are arranged in a tree-like branching structure to reveal recurrent themes and phrases. Unlike word clouds, word trees visually display the connection of words in the dataset, providing some context to their use. Words that show up more frequently in combination with the pre-selected word(s) are displayed in larger font size. With the right phrase, a word tree can reveal the heart of a data set.

Since the Pubmed dataset provided by Deep Dive contains pre-processed tokenized words instead of full abstracts, therefore for the generation of Word trees we sample out 5000 abstracts of documents by querying Pubmed for two different two topics namely Microbiology and Fibromyalgia. Figure 12(a and b) show the word trees for both the queries respectively. From these examples we infer that words like DNA are commonly followed by names of variety of drugs and chemicals, displaying a high correlation among these terms. Further, as seen from figure 12(b) there is a high cooccurrence between words/phrases like *fatigue* and *sleep disturbances*, signalling lack of sleep to be a major fatigue cause. Phrases like *treatment of fibromyalgia* displays various possible treatments on each branch of the tree, thus revealing lots of useful insights from the Pubmed dataset.

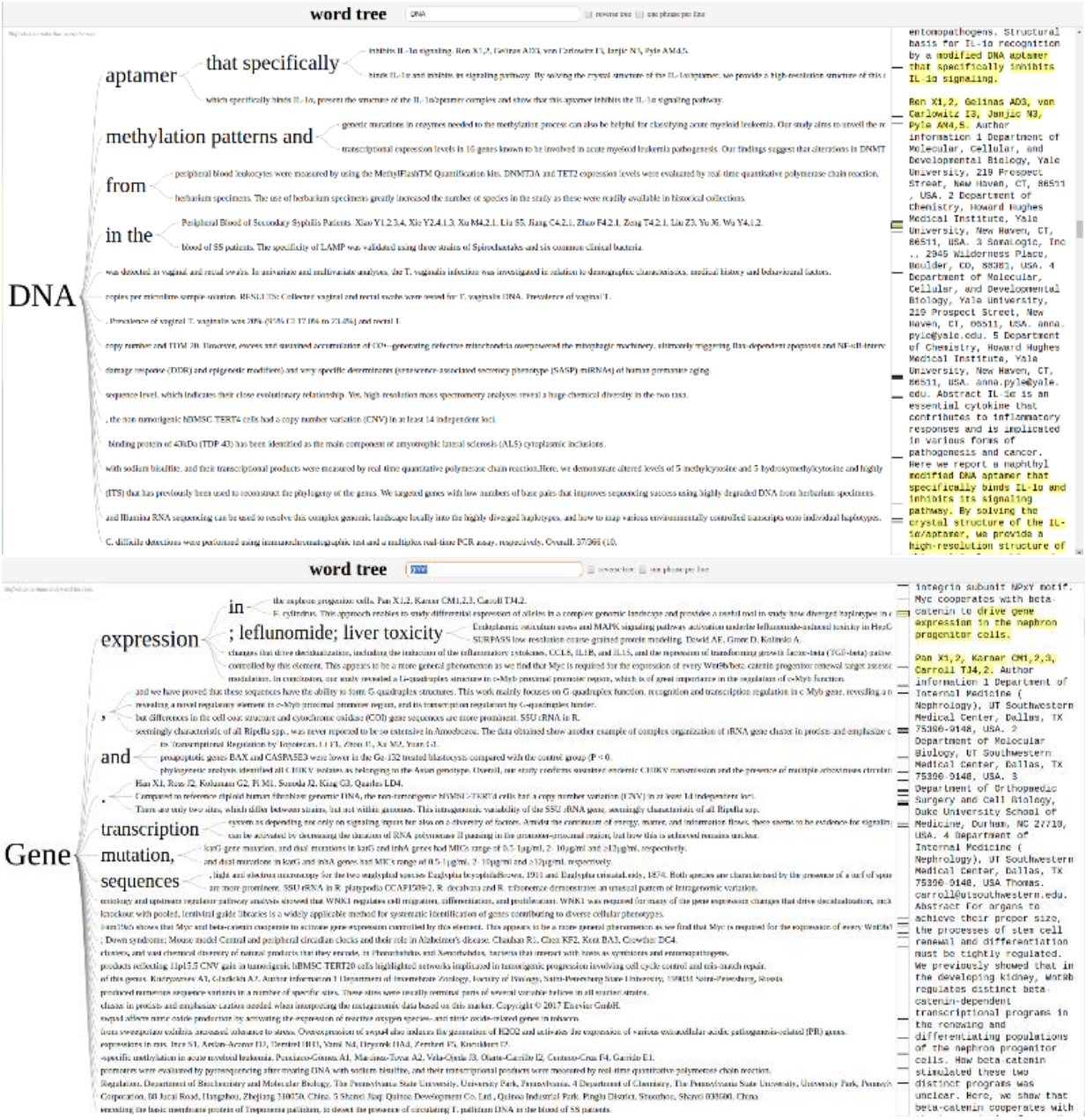


Figure 12(a): Word trees with root words *DNA* and *Gene* drawn from abstracts from a Pubmed query on M*icrobiology*.

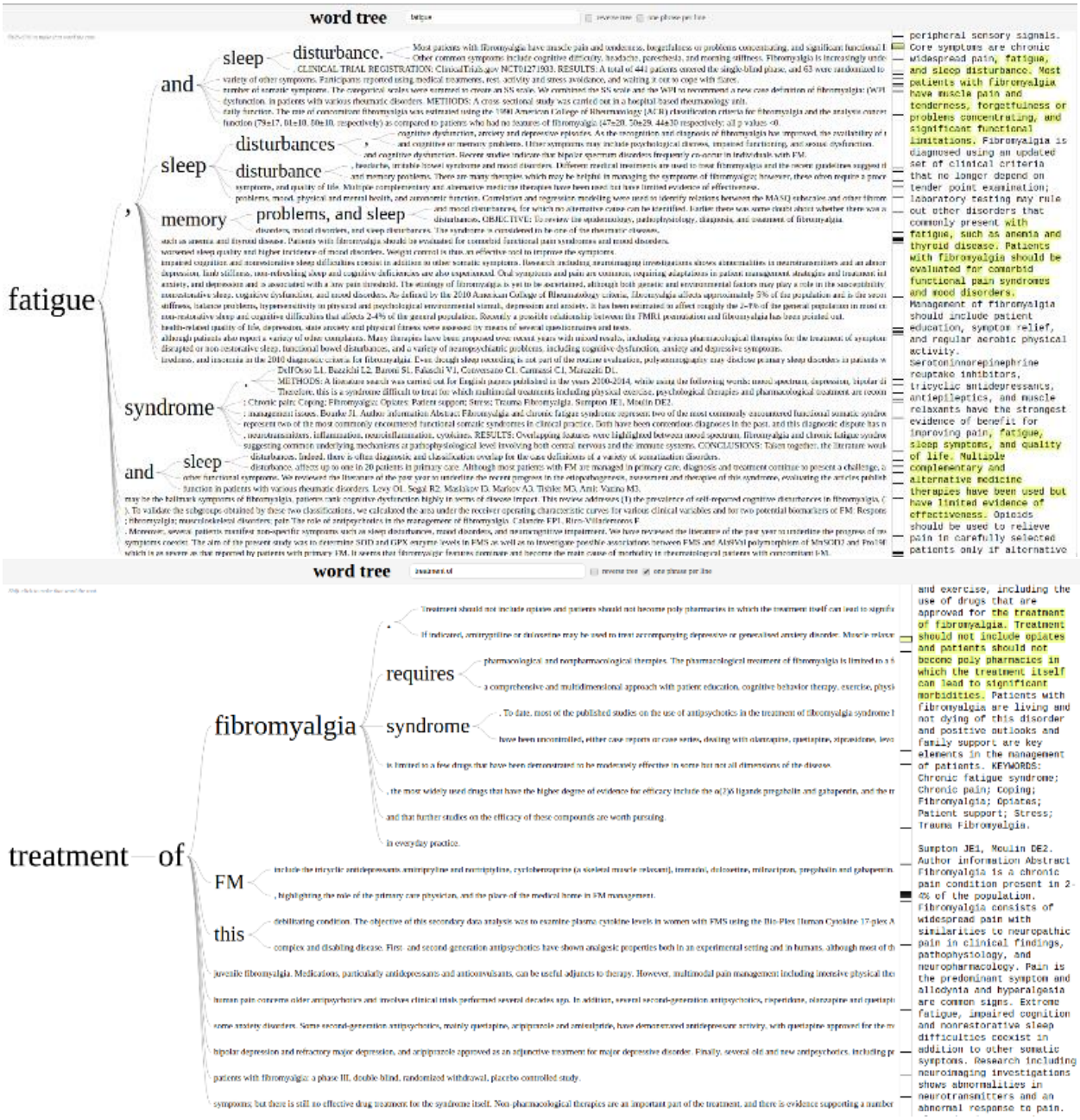


Figure 12(b): Word trees with root words *fatigue* and *treatment* drawn from abstracts from a Pubmed query on *Fibromyalgia.*

4. FURTHER INSIGHTS

To gain further useful insights from the PubMed dataset, we perform a few more different kinds of visualizations namely, Graph-based and Geometric. Firstly, we plot the Link Graphs [10] and Text Arcs [11] for the PubMed corpora which provide a compact representation of interconnection and interrelation between different topical segments. Words are treated as nodes, while the arcs/links show contextual co-occurrences and dependencies between words. Figure 13 and 14 display the Link graphs and text arcs with their filtered views along-side for a particular chosen word as center-root to enhance their readability.

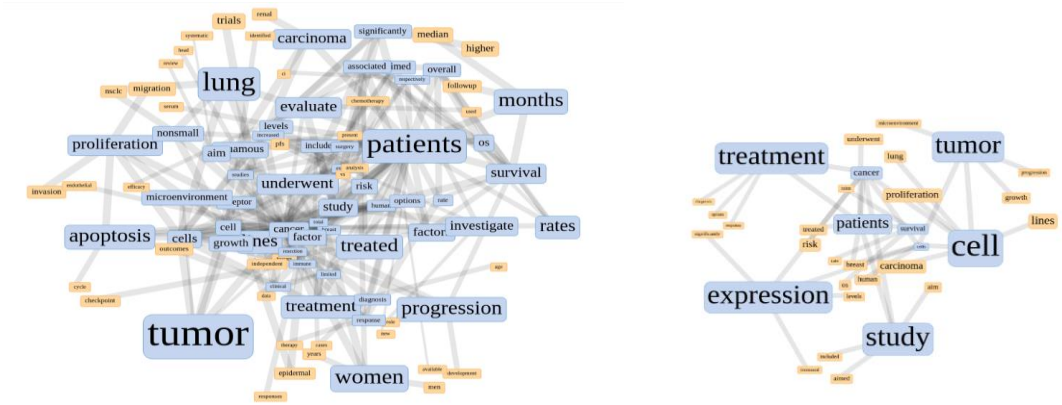
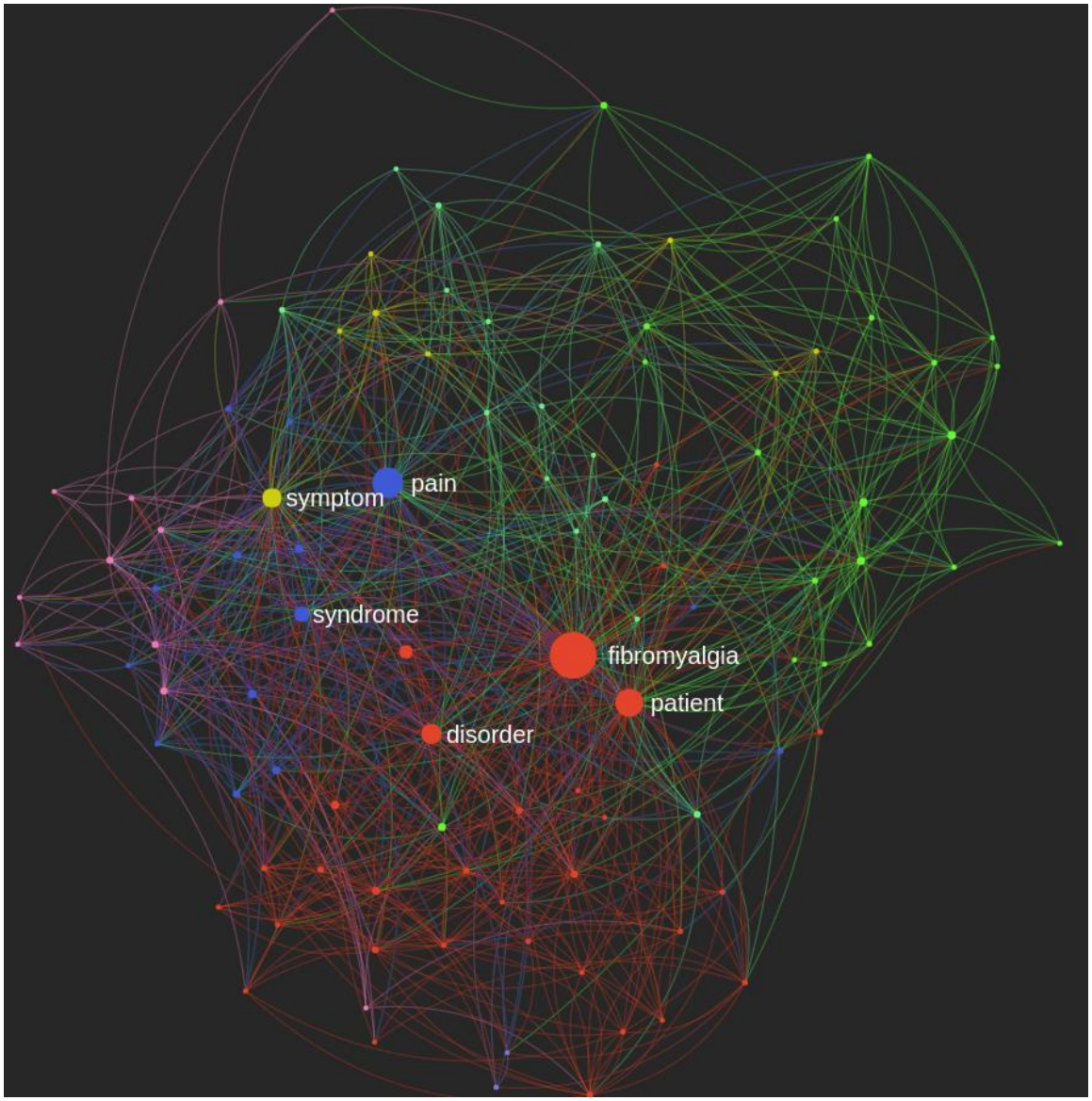


Figure 13: Link Graphs for a) top 100 words and b) filtered view with *patient* as the root node.



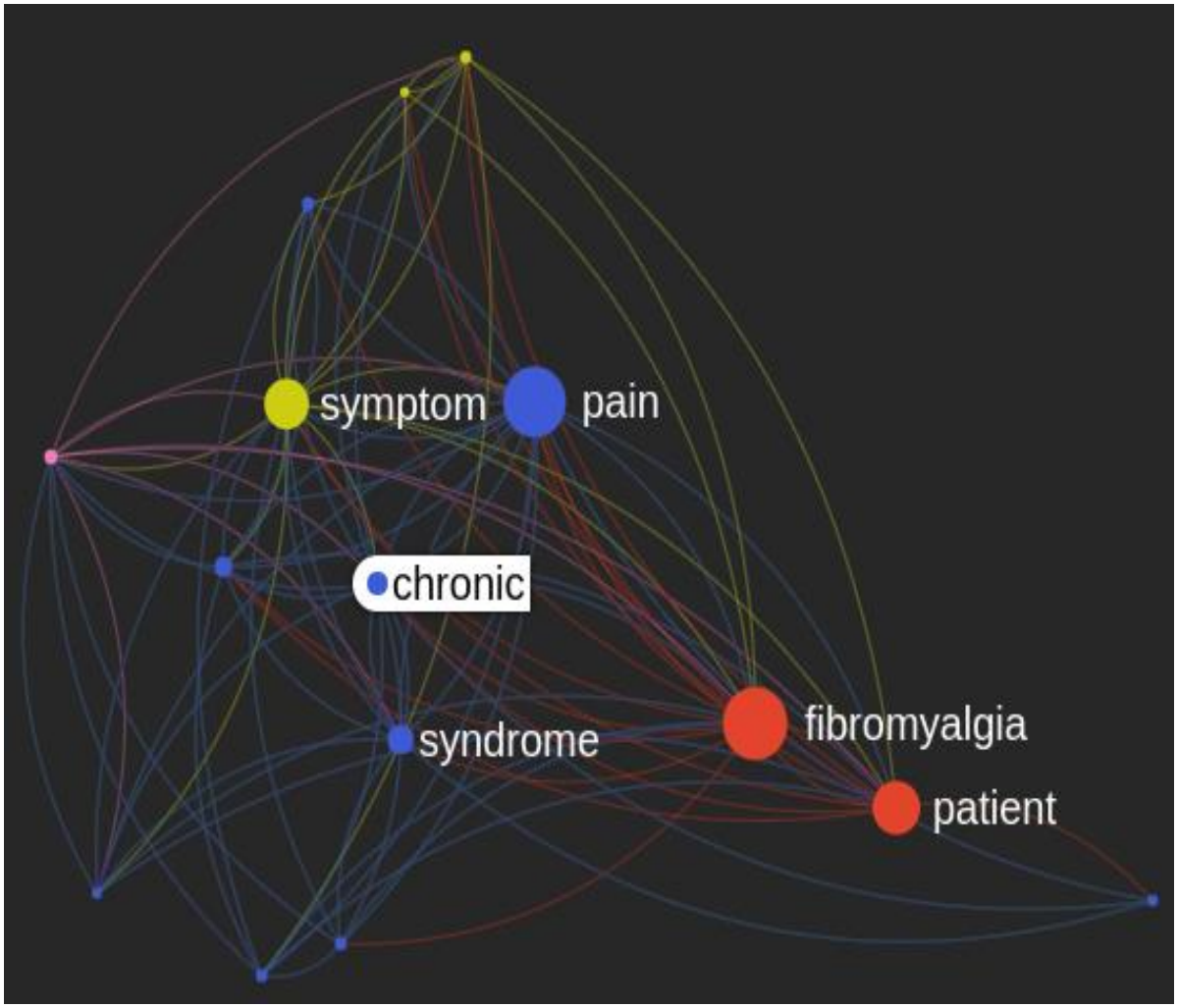


Figure 14: Text Arc for a) entire dataset and b) filtered view with *chronic* as the root node.

Secondly, we plot Stream graphs and trend lines [47] as a part of geometric visualizations to observe hidden pattern from raw and relative word frequencies. The stream graphs and trends are shown in figure 14. We further, generate a new type of a plot called Term Berry [47] as shown in figure 15(a) which displays the frequencies of individual terms across the documents with terms grouped according to their semantic closeness. Lastly, we generate a graph called as a Magnet graph [47] as shown in figure 15(b) to get a different perspective from our visualizations. The magnet graph basically places the top 20 most frequent terms along the circumference of its disc with the name of the dataset at the center. Each term exerts an attractive force towards the center term with magnitude weighted by their frequency counts. The displacement of the center term from its position to certain terms shows the dominance of those terms over other terms. In the given plot, terms like *cancer, patients, cells, study* dominate other terms like *tissue, lung, molecular* etc.

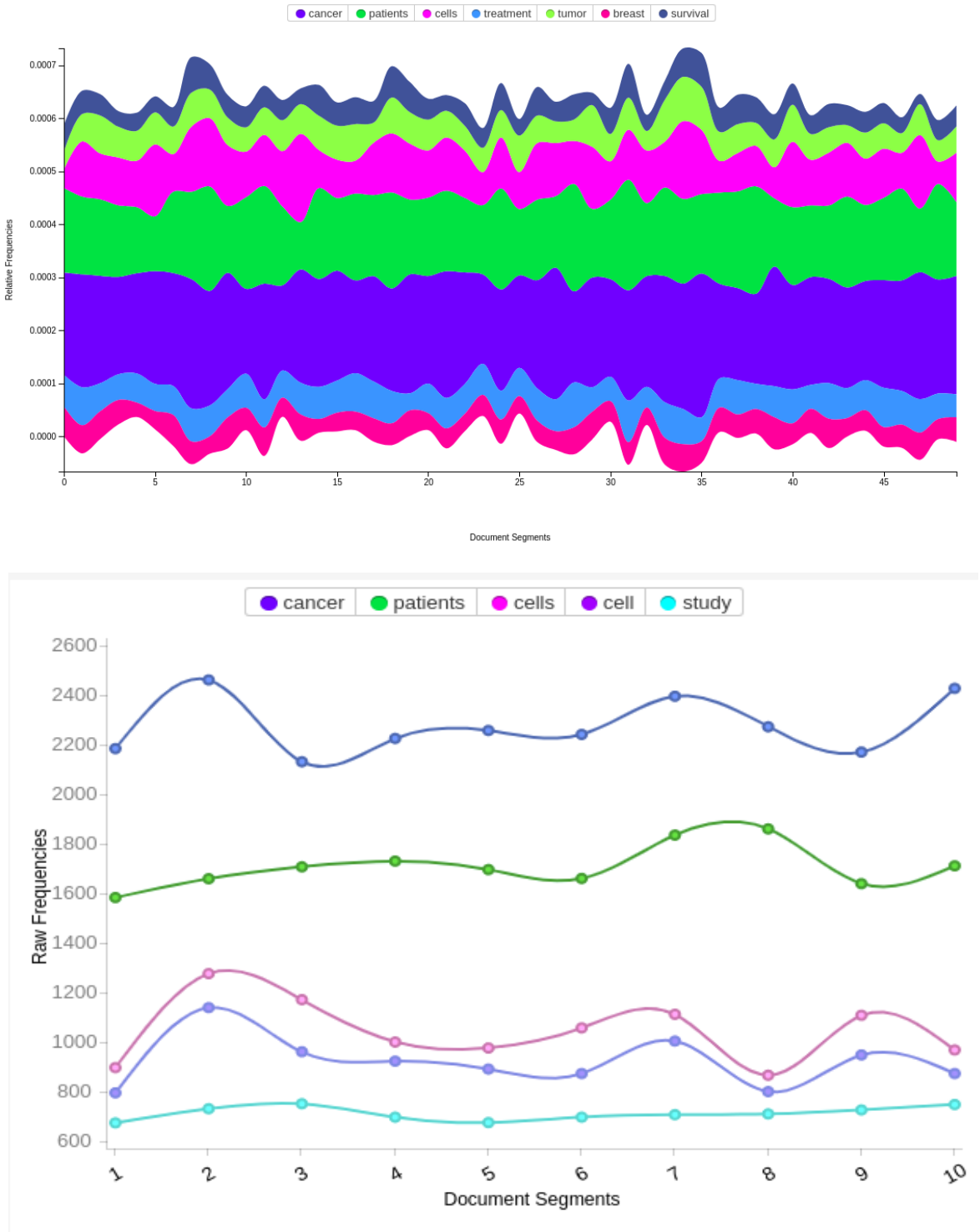


Figure 15: Stream Graphs and Trend Line Analysis for top 5 most frequent words.

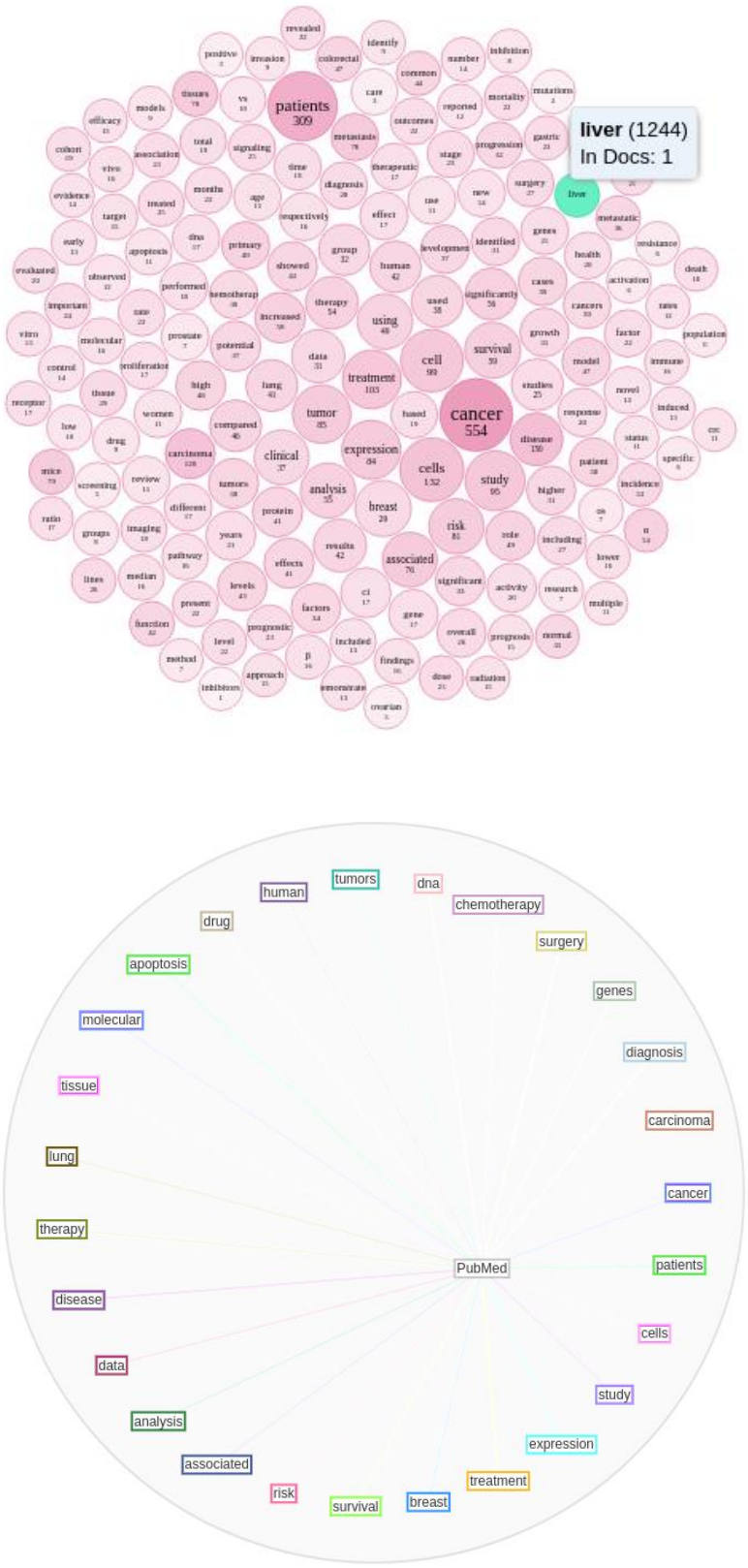


Figure 16: Term Berry (Top) and Magnet Graph (Bottom) on PubMed Dataset.

5. COMPARISON OF LEXICAL AND TEXT VISUALIZATION TECHNIQUES

We visualize the PubMed Dataset through lexical and text visualizations which are broadly categorized into six different categories [49] as depicted in figure 17. First, the basic category represents the visualizations that deal with raw or relative word frequencies across the documents. They provide an outline of content present in various articles. Secondly, categories like geometric techniques deal with the visualizations of geometric transformations and projections of data, while categories like clustering depend upon unsupervised machine learning procedures and are used to discover hidden patterns inside the data. Further, there are hierarchical and graphical techniques used to visualize datasets using hierarchical partitioning and interconnected networks. Lastly, there are embedding methods which utilize distributed representation of words in term of vectors to visualize their contextual dependence on each other. Our objective through these visualizations is to gain greater insights and hence draw conclusions from contents of PubMed database. These visualizations enable us to directly interact and easily deal with such large highly non-homogenous and noisy datasets. They provide a qualitative overview for further quantitative analysis over the dataset.

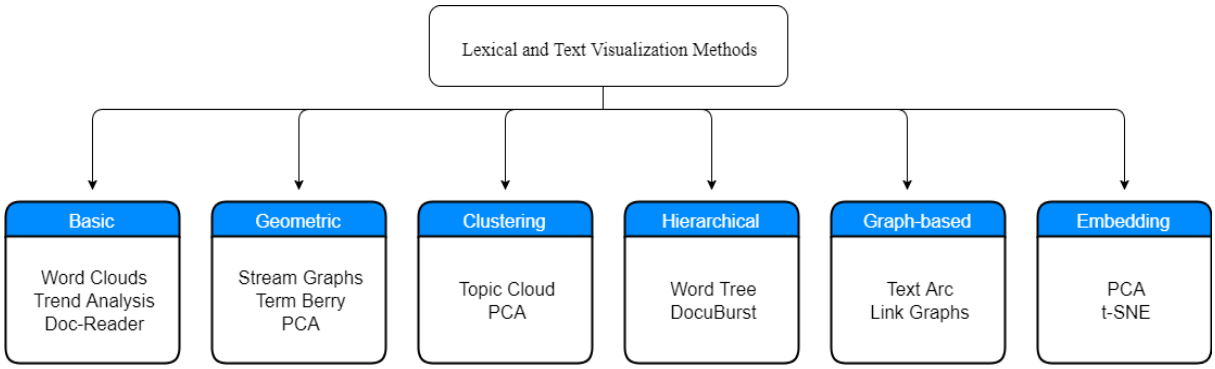


Figure 17: Hierarchy of Lexical and Text Visualization Techniques used.

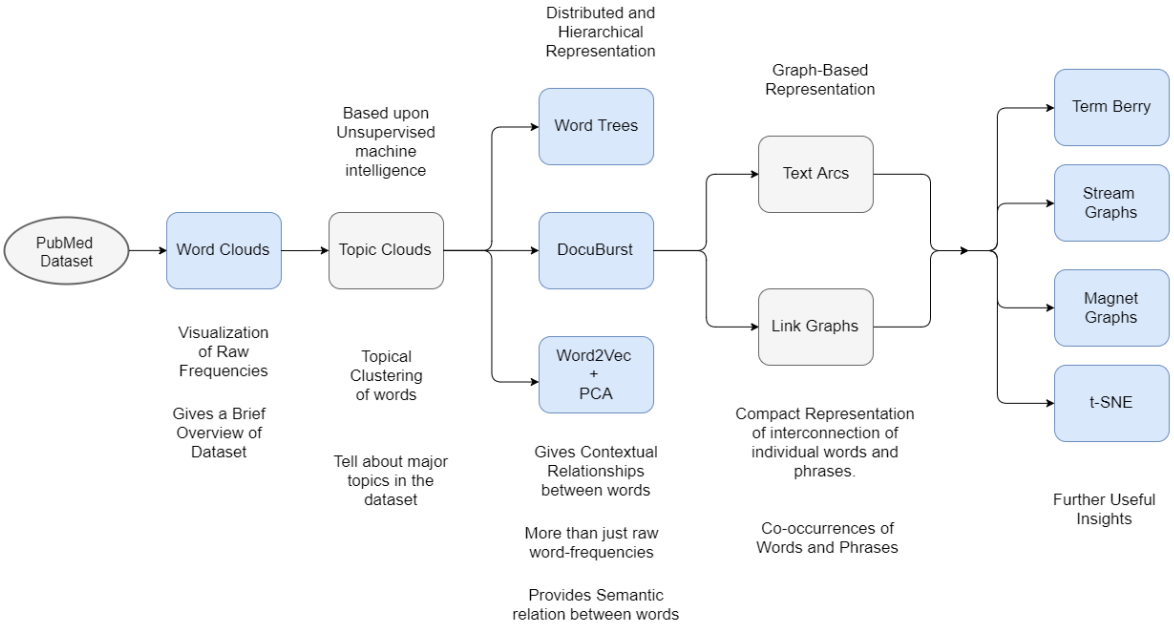


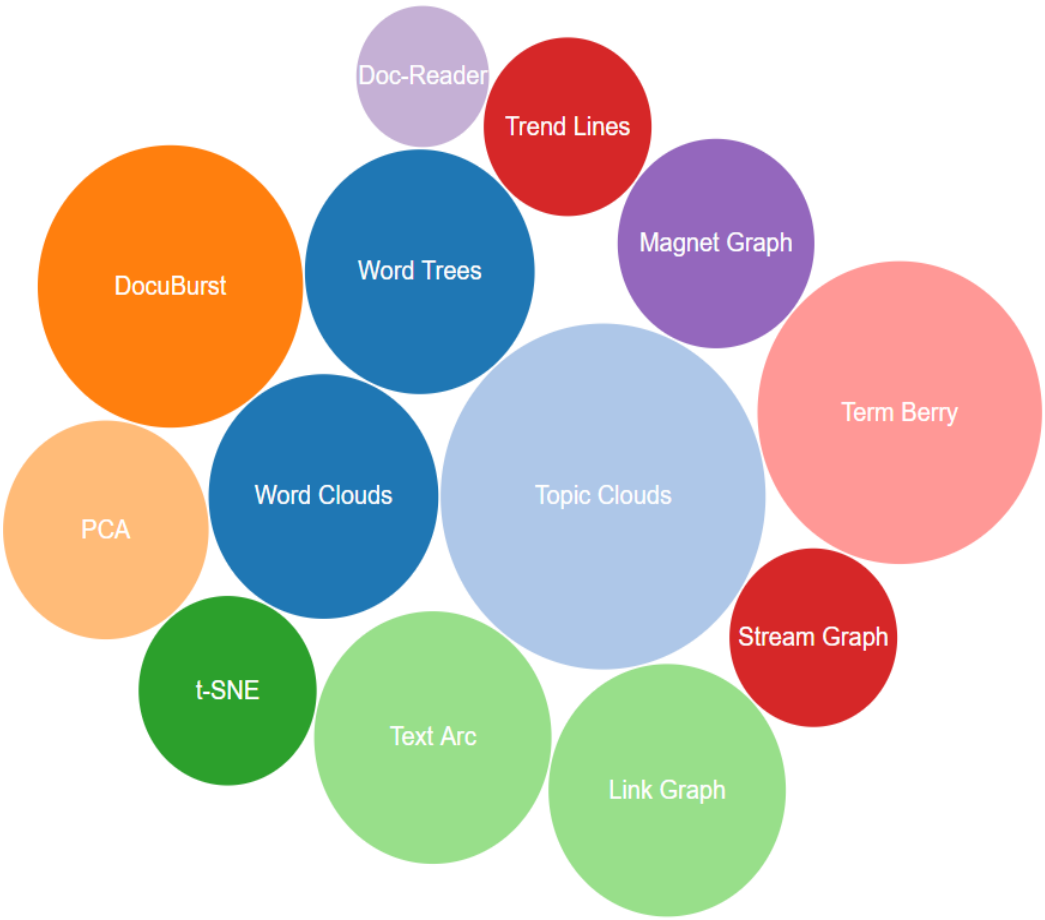
Figure 18: Data visualization pipeline with main advantages of each technique used.

Table 4 summarises the advantages and drawbacks of each visualization technique we encountered during the analysis of PubMed dataset. Figure 18 depicts data visualization pipeline we follow to gain insights from PubMed with short notes around them representing their main advantages. We display the importance of each visualization technique in terms of the amount of insights we gather from that particular visualization using a bubble chart generated with [13], as shown in figure 19(a). The greater the diameter of a bubble, the more we felt that visualization was able to reveal about the dataset. We also visualize the relative importance of each visualization technique using a TreeMap [15], as shown in figure 19(b).

Table 4: Comparison of Lexical and Text Visualization Method

Visualization Method	Pros.	Cons.
Word Clouds	• Give brief overview of content • Show the most frequent words in all the documents • Is scalable with large size datasets • Generates Interactive and easy to read visualizations • Fast and efficient	• Does not display the interconnection between those words • Contextual relationship between the words could not be inferred
Topic Clouds	• A word cloud with semantic grouping of words into major topics • Provides greater insights into the textual data with refined granularity	• Does not display the interconnection between those words • Contextual relationship between the words could not be inferred
Word Trees	• Hierarchical representation i.e. parent-child relationship • Shows the contextual dependencies between words • Fast and efficient	• Due to large branching factors, trees grow quickly, using lot of real estates (i.e. space) • Difficult to read with increase in branching
DocuBurst	• Radially symmetric, space-filling layout • Shows the contextual dependencies between words • Great to navigate hierarchical data • More effective at visualizing “large” datasets	• Hard to estimate the exact frequency of word using arc length • Small, periphery arcs can be hard to read/ analyze

	• More intuitive and easier to read than Word Trees	
PCA, t-SNE	• Geometric; Embedded representation of words in a text • Distributed representation; Words defined in terms of vectors • Avoids localist/discrete representation of words. • Grouping of semantically related words into clusters	• Time Consuming. • Difficult to read and analyze for large datasets • Techniques like t-SNE can be misleading and can produce mysterious visualizations
Text Arcs, Link Graphs	• Graphs with words as nodes, arcs/links to show their contextual cooccurrence • The interrelation between different topical segments. • Networking of textual data items. • Different point of view	• Difficult to read with increase in number of edges/nodes • Too dense representation inside a small screen area



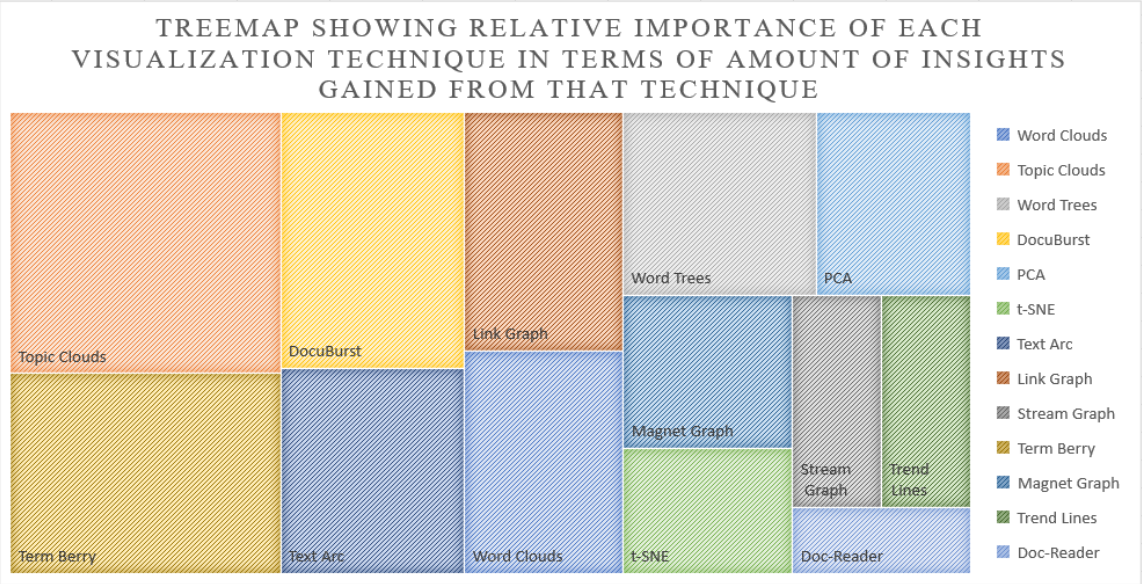


Figure 19: Bubble Chart (Top) and TreeMap (Bottom) visualizing the relative importance of each visualization technique.

6. TOPIC MODELLING ON PUBMED

In the previous sections, we utilized numerous visualization techniques to visually explore and get a general overview of the PubMed corpora. In this section, we use the MALLET library to perform topic modeling to extract knowledge from the biomedical database by identifying and clustering the major topics inherent in the database.

MALLET is a Java-based package that includes sophisticated tools and a wide variety of algorithms for performing statistical natural language processing tasks like document classification, information extraction, topic modeling, etc., for analyzing large collections of dark data and extracting information out of it. It is co-written by Andrew McCallum and his group at the University of Massachusetts Amherst, as well as contributions from Fernando Pereira, Ryan McDonald, and others at the University of Pennsylvania. The topic modeling toolkit in MALLET contains several efficient, sampling-based implementations of Latent Dirichlet Allocation (LDA), Pachinko Allocation, and Hierarchical LDA, etc.

We use Latent Dirichlet Allocation from the topic modeling toolkit to map terms and words present in the database into a low dimensional continuous space by exploiting the word collocation patterns. We then use Kmeans, a famous clustering algorithm to group words into semantically related clusters according to their cosine similarity measure. Typically, only a small number of topics are present in each document, and only a small number of words have a high probability in each topic. So, we represent these topics along with top-n most frequent terms in each topic using a Topic Cloud.

A Topic Cloud is a pie chart visualization of inherent topics inside a database which consists of a number of topic slices, where each slice contains the most important words in that topic. The relative prominence of words in a topic is made explicit by scaling their sizes in proportion to their confidence score as computed by

LDA. A topic cloud is like a word cloud giving the frequency of words or phrases, but the major difference that a topic cloud offers as compared to word cloud is the semantic grouping of words under similar topics, hence capturing the contextual relationship between terms and providing more significant insights into the text data with refined granularity.
We analyse top *n* most significant topics inherent in PubMed where *n* specifies the number of topics considered. Figure 20 shows the topic clouds for *n=5* and *n=7* topics respectively. Both the topic clouds elegantly summarize the contents of PubMed documents, with each slice depicting the most important words in that topic. Words like *patients, cancer, disease, treatment* etc. are grouped under one topic whereas words such as *data, study, research* depict another topic.

On comparison of both the topic clouds, we find that the topics produced for *n=7* criterion are more homogenous, with uniform proportions, while the topics generated for *n=5* criterion are more disproportionate; indicating *n=7* to be a coherently less noisy criterion as compared to the *n=5* criterion. In addition to topic clouds representing top five and seven most significant topic, we also compute top 20 topics inherent in PubMed as reported in Table 5 along with proportion for each topic representing the confidence score as computed by LDA. The topic with most significant proportions majorly focuses upon recent research and development strategies such as *molecular and therapeutic research* primarily concerned with *Cancer*, indicating substantial research work related to Cancer.

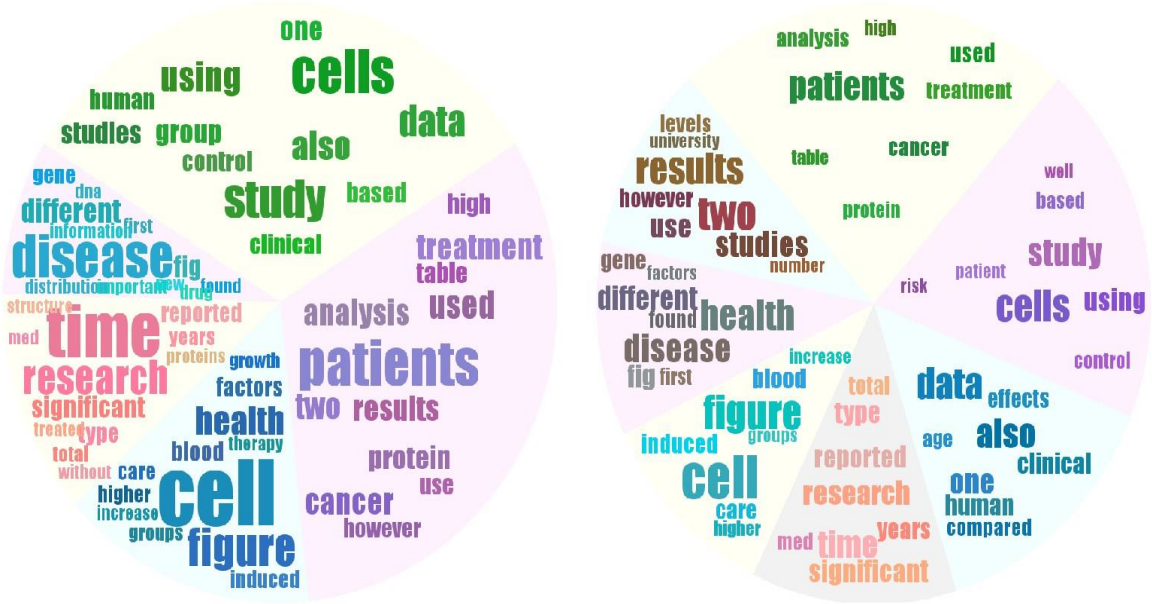


Figure 20: Topic Cloud for Left: *n=5* and Right: *n=7* topics respectively.

Table 5: Top 20 topics along with their proportion in PubMed; Computed using MALLET.

Topic Id	Proportion	Major Words in topic
1	**0.26558**	cancer clinical review studies therapeutic development disease treatment potential current molecular research provide including recent strategies data role approach diseases
2	0.11401	patients survival cancer analysis prognostic stage study factors group lymph significantly tumor clinical multivariate risk significant gastric metastasis ratio compared
3	0.10156	cells cell cancer expression apoptosis proliferation growth migration protein pathway effect lines study effects human signaling inhibition invasion tumor induced
4	0.09279	cancer risk women breast study years age screening mortality association incidence factors data population higher increased men diagnosis health rates
5	0.08076	cancer care patients health quality study life treatment patient oncology medical information studies data intervention research systematic clinical guidelines palliative
6	0.07577	data imaging model method analysis sensitivity based accuracy volume study values methods detection performance results diagnostic mri compared images test
7	0.07482	case tumors pancreatic thyroid tumor cases diagnosis carcinoma patient lesions malignant adenocarcinoma report rare imaging biopsy benign metastatic petct primary
8	0.07106	expression mir cancer gene tissues cell genes analysis tumor lung crc protein study normal significantly tissue mirnas correlated prognosis rna
9	0.06988	patients treatment therapy months radiation radiotherapy dose chemotherapy median treated brain metastases local cancer lung tumor received control survival followup
10	0.06737	protein signaling dna cell proteins binding role pathway activity cells transcription human cellular regulation function receptor activation gene complex show
11	0.06581	cells tumor immune cell mice expression tumors pdl human microenvironment model macrophages bone receptor stem antibody responses antibodies antitumor immunotherapy
12	0.05701	patients treatment phase response leukemia survival trial lymphoma disease acute therapy study months safety chemotherapy aml median cell efficacy transplantation
13	0.05561	breast cancer treatment resistance inhibitors growth receptor egfr therapy lung patients kinase factor combination tumor cell nsclc response tnbc chemotherapy
14	0.04979	mutations mutation genetic dna hpv gene genes sequencing patients variants brca methylation cervical samples identified kras human genomic testing polymorphisms

15	0.04636	drug cancer cells delivery nanoparticles tumor vivo therapy cell activity imaging vitro anticancer release high compounds results nps surface potential
16	0.04391	patients surgery group postoperative complications surgical outcomes study cancer underwent resection days performed time laparoscopic bleeding hospital compared approach cases
17	0.04026	prostate patients cancer levels serum pca detection psa plasma bladder diagnosis antigen blood circulating healthy biopsy men clinical significantly androgen
18	0.03793	liver hcc carcinoma skin oral melanoma hepatocellular cell squamous protected article reserved rights copyright oscc head neck hepatic hepatitis cutaneous
19	0.03643	mice group rats effects stress study oxidative intestinal effect groups treatment exposure significantly inflammatory radiation levels increased compounds control damage
20	0.0324	metabolic metabolism levels obesity diabetes glucose vitamin weight acid mice metabolites fatty increased lipid acids bmi loss fat tissue mitochondrial

7. NETWORK AND CITATION VISUALIZATION

Network visualization also called Network graphs are often used to visualize complex and convoluted relations between an enormous amount of entities. They represent information in a hierarchically structured manner through an interconnected network of entities highlighting the correlation between them. At its most basic level, a network graph consists of nodes and edges. Nodes represent the entities, and edges represent the relationship between those entities. Edges in the graph can be directed or undirected. Directed edges indicate the flow of information from one node to another. Undirected edges, on the other hand, show the presence of a bidirectional relationship between the two nodes.

One of its variants, Citation networks has been used extensively in the field of bibliometrics [Belter, 2012]. Citation networks proffer quick summarization and visualization of the structure inherent to a set of publications. The resulting visualizations provide insights not only into the present state of scientific research but in identifying potential future research directions and collaboration opportunities. Bibliometric or citation networks can be classified into *direct citation networks, co-citation networks,* and *bibliographic coupling relations*. Direct citation networks, also known as cross citation networks represent research documents citing each other directly as nodes in the network. These networks only offer a direct indication of relatedness between the entities. These are usually very sparse networks and hence relatively uncommon in research settings.

The second variant of citation networks, i.e., co-citation networks represents co-cited research documents (i.e., a pair of documents that are cited by some other group of common documents) as network entities. The higher the number of research documents citing the two documents, the stronger is the relatedness between them. [White and McCain, 1998] used these co-citation networks to study the researchers in the field of information science. And in the

final variant, bibliographic coupling, two documents are said to be coupled if both cite a common research document [Kessler, 1963]. In other words, the more common the references two documents have, the stronger is the coupling relation between them.

We use VOSviewer, an open source software tool used for constructing and visualizing bibliometric networks, [Van Eck and Waltman, 2010] to construct networks of scientific publications and journals. Each visualization map consists of a network of objects of interests, also known as entities. Entities can be research documents, authors, or keywords which are interconnected with other entities through edges representing citation (co-citation and bibliographic coupling), co-authorship, or co-occurrence links. Each link has a strength associated with it, represented by a positive numerical value. The higher this value, the stronger is the link between the connected entities. Furthermore, each entity is grouped into a non-overlapping and exhaustive cluster. Entities have various attributes associated with them for instance, the weight attribute of the entity or the distance between two entities. The weight of an entity indicates the importance of that entity in the network. An entity with a higher weight is regarded as more important than an entity with a lower weight and hence shown more prominently. The distance between any two entities indicates the strength of the relationship between the corresponding entities. The closer the entities are to each other, the stronger they are correlated with each other.

We create two types of visualizations, *Co-Authorship* and *Co-occurrence Word* Networks. The CoAuthorship networks link the authors of various biomedical research publications in Pubmed based upon the number of publications they have co-authored. These networks help in obtaining significant insights on possible communities and group of researchers who are involved in contributing in their field and may prove to be of significant help to researchers of the same or even different field which are closely related to closely focus on the works of major contributors and head their research forward. The Co-occurrence word networks on the other hand link keywords and term phrases which co-occur together. These networks reveal the semantic correlation among different terms along with major terms highlighted within each cluster. Various useful insights for example in case of PubMed names of major medicines used to cure a disease, or side-effects of treatment, or the possible age groups or gender targeted because of disease can be inferred from these networks with ease.

To create these networks, we query the PubMed database on two different topics, *Alzheimer's Disease* and *Tuberculosis* respectively. We obtain the resulting abstracts for each topic from the PubMed site in MEDLINE format. While creating these networks, some parameters need to be selected. Like for the coauthorship networks, we choose the parameter of *full counting,* i.e., each link contributes equally and Authors as the unit of analysis. We choose the minimum number of documents an author must have published to be ten as the threshold to limit our network to mentions of authors who have contributed at least ten documents related to the topic for which the network map is created. Finally, from the authors shortlisted we select top 500 authors to be visualized in our network based upon their total link strength which indicates the total strength of the co-authorship links of a given researcher with other researchers. Similarly, for the Co-occurrence based Word Networks, we first select the option to ignore copyright statement to get rid of the unwarranted text, we extract text from both title and abstract fields of the MEDLINE document, and we select the option of full counting, to count all the occurrences of a term in a document. We filter the less significant terms from more significant terms by setting the minimum number of occurrences of a term

barrier to 8. For each of the filtered term, VOSviewer calculates a relevance score, which represents the specificity of a term towards the topics covered by the text data. We select the top 80% most relevant terms depending upon the relevance score metric to be displayed in our network. We select minimum cluster size to be 2 and *Association Strength* to be the normalization method for the layout algorithm for both visualizations discussed above.

Figure 21(a) shows the Co-Authorship networks for PubMed documents related to the topic of Alzheimers Disease. From the figure, we can observe that some authors like *Bennett Da, Blennow K, Perry G, Zhang Y, Wang Y* have bigger node sizes as compared to other authors thereby indicating a higher proportion of work contributed by these authors in the queried field of work. From the same figure, we can also observe the potential clusters of authors depending upon the papers they have co-authored and the areas of their study. Authors in the purple and yellow cluster appear to work only with authors within their clusters while the work of authors of red, blue and cyan clusters are uniformly interspersed between different clusters. Also, from the distances between two authors in the visualization and the thickness of links connecting them, the relatedness of the authors can be inferred. In general, the closer two authors are located to each other or thicker the link connecting them is, the stronger their relatedness.

Similarly, Figure 21(b) shows the Co-Authorship network for PubMed documents related to the topic of *Tuberculosis*. From the figure, prominent authors can be observed based on their node sizes like *Harries AD, Van Soolingen, Wang J, Narayanan PR,* etc. Various clusters can also be observed shown in distinct colors with the authors in the red cluster can be seen to be widely connected with authors present in other clusters, thereby indicating major source of work done by these authors related to tuberculosis.

In addition to the co-authorship networks, we also show the Co-Occurrence word networks in Figure 22. Various useful insights can be gathered from these networks. Firstly, from the co-occurrence word network shown in Figure 22(a) we can infer that *Alzheimers disease* is related to the brain due to presence cooccurring terms such as *brain, memory, cognition,* etc. We can also infer that *males* are more vulnerable to Alzheimer disease as compared to *females*. Potential age group suffering from Alzheimer disease can also be identified as the *middle age* to *old age* group. Various side effects of Alzheimer disease can also be found such as *memory disorders, dementia, depression* etc.

Similarly, from the co-occurrence word network for *Tuberculosis* shown in Figure 22(b) many important terms such as certain types of tuberculosis, like *abdominal tuberculosis, pulmonary tuberculosis, neck tuberculosis,* etc can be identified. The network also lists the names of certain *vaccines* and *resistance techniques* related to tuberculosis. Finally, on studying the network in depth, names of certain places like *India, North Carolina, England,* etc can be found in association with terms like *healthcare workers, survey, treatment success rate,* etc indicating that these places are playing a major role in spreading information and public awareness related to disease and are providing proper treatment to people affected by tuberculosis.

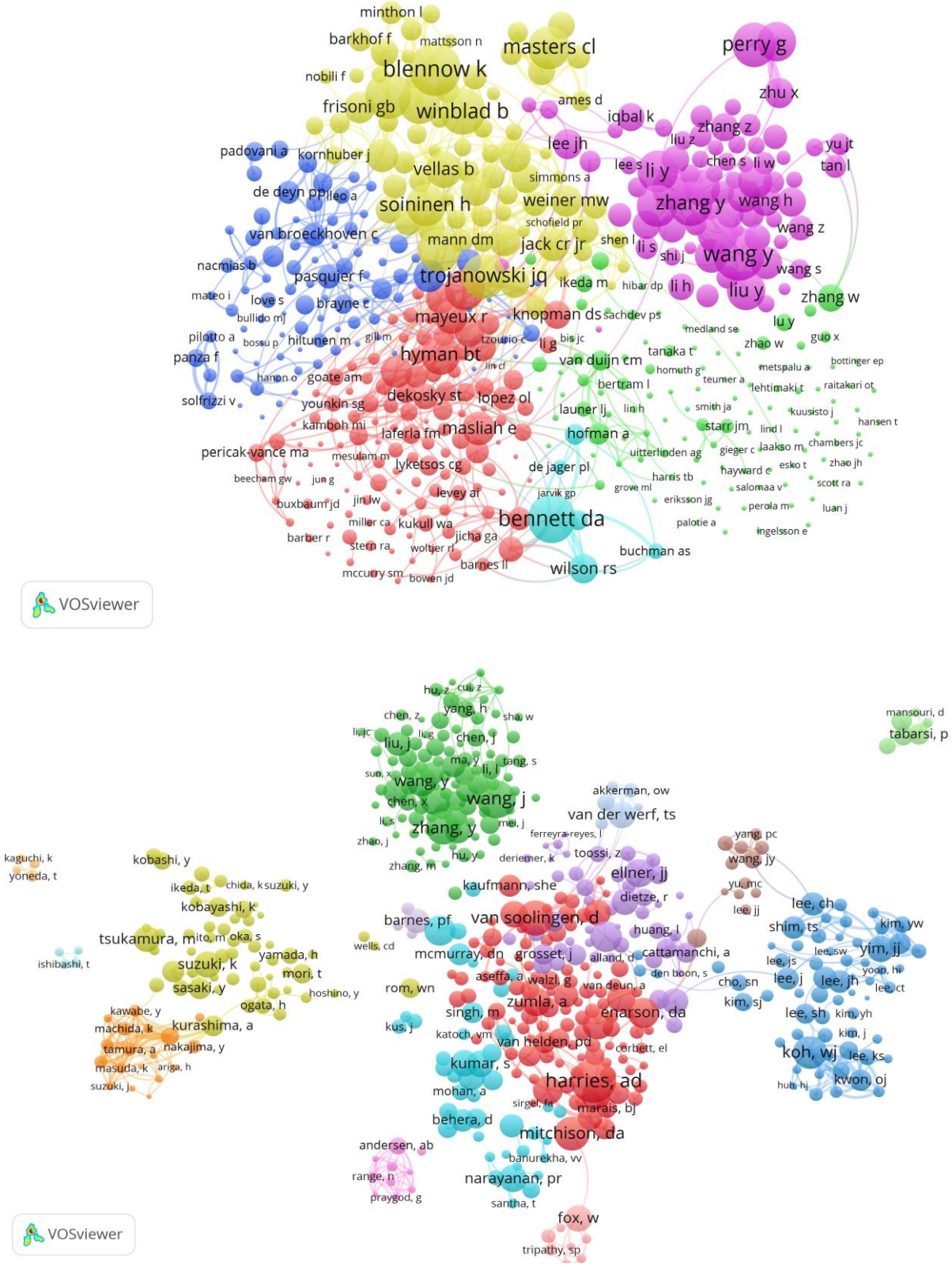


Figure 21: Co-Authorships Networks for (a) Alzheimers Disease (top) and (b) Tuberculosis (bottom) generated using VOSviewer, network and citation viewer.

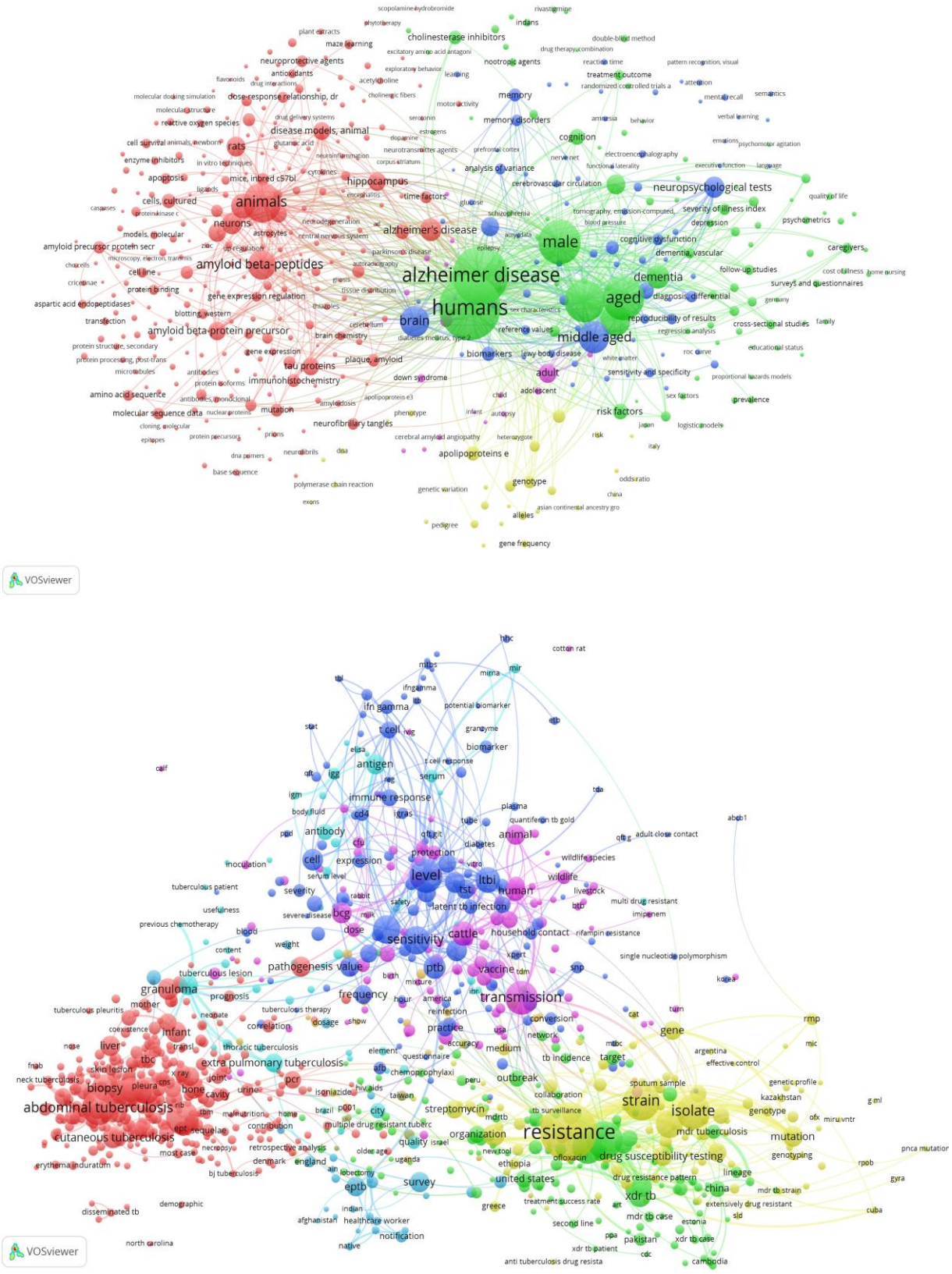


Figure 22: Co-Occurrence Word Networks for (a) Alzheimers Disease (top) and (b) Tuberculosis (bottom) generated using VOSviewer, network and citation viewer.

8. CONCLUSION

The main motivation behind the present work was to diminish the limitation of human cognition and perception, in handling and examining enormous amounts of information. The present analytical work endeavored to exploit various data visualization tools and natural language processing techniques to proffer a potential solution to overcome the problem of analyzing overwhelming amounts of information. In this empirical research, we utilized the rich corpus of the PubMed to visually explore and analyze lexical and textual biomedical dark data to mine knowledge out of it. We employed various text summarization and visualization techniques like computation of raw term and document frequencies, word clouds, *TF* ×*IDF* scores, DocuBurst, etc., to get a general overview of the PubMed database. We then utilized the MALLET library to perform topic modeling to extract knowledge from the biomedical database and visualized the inherent major topics inherent in PubMed using Topic Clouds. Finally, we used network and citation visualization techniques, to construct bibliometric networks, i.e., Co-Authorship and Co-Occurrence word networks for studying relationships between different entities like scientific documents and journals, researchers, and, keywords and terms. All of the techniques that were employed to explore visually and mine knowledge from dark data, i.e., PubMed proved to be of great help in allowing quick comprehension of information, the discovery of emerging trends, and identification of relationships and patterns within the database.

REFERENCES

[1] IDC's Digital Universe Study, Dec 2012, https://india.emc.com/leadership/digital-universe/index.htm
[2] Gartner Inc., https://www.gartner.com/it-glossary/dark-data, Accessed on 10-14-2017.
[3] Deep Dive, http://deepdive.stanford.edu/, Accessed on 10-14-2017.
[4] Macro Base, https://hazyresearch.github.io/snorkel/MacroBase, Accessed on 10-23-2017.
[5] Apache Madlib, http://madlib.apache.org/, Accessed on 10-23-2017.
[6] PubMed, https://www.ncbi.nlm.nih.gov/pubmed/, Accessed on 10-23-2017.
[7] SM. Douglas, GT. Montelione and M. Gerstein. 2005. PubNet: a flexible system for visualizing literature derived networks. Genome Biology, (Jun. 2005), R80. DOI: https://doi.org/10.1186/gb-2005-6-9-r80
[8] RJ Roberts. 2001. PubMed Central: The GenBank of the published literature. In Proceedings of the National Academy of Sciences. 381–382.
[9] JR McEntyre, S Ananiadou, S Andrews, WJ Black, R Boulderstone, P Buttery, D Chaplin, S Chevuru, and et al. 2010. UKPMC: A full-text article resource for the life sciences. Nucleic Acids Research. 39(Database issue): D58–D65.
[10] NLM Catalogue: Journals referenced in the NCBI Databases. NCBI. 2011.
[11] Reyes-Aldasoro C. 2017. "The proportion of cancer-related entries in PubMed has increased considerably; is cancer truly "The Emperor of All Maladies"?". PLOS ONE. 12 (3): e0173671. DOI: 10.1371/journal.pone.0173671. PMC 5345838. PMID 28282418.

[12] PubMed Preprocessed Dataset, March 2014, http://deepdive.stanford.edu/opendata/#pmc-oa-pubmed-central-open-access-subset, Accessed on 10-8-2017.
[13] P. Morville. 2005. Ambient Findability: What We Find Changes Who We Become. O'reilly Media, Inc.
[14] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. Dbpedia : a crystallization point for the web of data.
[15] H. Siirtola, T. Laivo, T. Heimonen, and K. Raiha. 2009. Visual perception of parallel coordinate visualizations. In IV '09: Proceedings of the 2009 13th International Conference Information Visualisation. IEEE Computer Society, Washington, DC, USA, 3–9.
[16] K. A. Olsen, R.R. Korfhage, KM Sochats, MB Spring, and JG Williams. 1993. Visualization of a document collection: the vibe system. Inf. Process. Manage. 29, 1, 69–81.
[17] J. A. Wise, J. J Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. 1995. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In IEEE Information Visualization '95, N. D. Gershon and S. Eick, Eds. IEEE Computer Soc. Press, 51–58.
[18] M. A. Hearst. 1995. Tilebars: visualization of term distribution information in full text information access. In CHI '95: Proc. of the SIGCHI Conf. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 59–66.
[19] M. A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics 23, 1 (March), 33–64.
[20] M. Analyzer. 2010. Matheo analyzer, database analysis and information mapping. http://www.matheo-analyzer.com/.
[21] Questel. 2010. Qpat, intellectual property patent and trademark searching. http://www.qpat.com/.
[22] J. T Stasko, C. G¨Org, and Z. Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization 7, 2, 118–132.
[23] Wanner. 2008. Towards content-oriented patent document processing. World Patent Information 30, 1, 21–33.
[24] S. Koch, H. Bosch, and T. Ertl. 2009. Towards content-oriented patent document processing. IEEE Symposium on Visual Analytics, Science and Technology, 203–210.
[25] D. Newman, A. Asuncion, P. Smyth, and M. Welling. 2009. Distributed Algorithms for Topic Models. JMLR 10, 1801–1828.
[26] K. Kucher and A. Kerren. 2015. "Text visualization techniques: Taxonomy visual survey and community insights", Proc. IEEE Pacific Vis. Symp., pp. 117-121. http://textvis.lnu.se/
[27] Word Clouds, http://www.betterevaluation.org/en/evaluation-options/wordcloud, Accessed on 10-20-2017.
[28] M. Bekos, T. van Dijk, M. Fink, P. Kindermann, S. Kobourov, S. Pupyrev, J Spoerhase, and A. Wolff. 2016. Improved Approximation Algorithms for Box Contact Representations. Algorithmica, 77(3), pp.902-920.
[29] J. Feinberg, Wordle, http://www.wordle.net/, Accessed on 10-20-2017.
[30] Word Art, https://wordart.com/, Accessed on 10-20-2017.

[31] S. Li, and T.S. Chua. 2017. Document Visualization using Topic Clouds. preprint arXiv:1702.01520. ARXIV. 2017arXiv170201520L
[32] S. Li, T.S. Chua, J. Zhu, and C. Miao. 2016. Generative topic embedding: a continuous representation of documents. In Proceedings of the ACL 2016.
[33] R. Speer, J. Chin, and C. Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. AAAI Conference on Artificial Intelligence, pp. 4444-4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972
[34] S. Carpendale, C. Collins and G. Penn. 2009. DocuBurst: Visualizing Document Content Using Language Structure. Eurovis '09.
[35] Docuburst, https://vialab.science.uoit.ca/docuburst/help.php, Accessed on 10-22-2017.
[36] C. Fellbaum (eds). 1998. WordNet: An Electronic Lexical Database, The MIT Press.
[37] D.R. Harris, R. Kavuluru, J.W. Jaromczyk, and T.R. Johnson. 2017. Rapid and reusable text visualization and exploration development with delve. In: Proceedings of the Summit on Clinical Research Informatics. AMIA.
[38] Word Trees, http://www.betterevaluation.org/en/evaluation-options/wordtree. Accessed on 11-5-2017.
[39] M. Wattenberg, F. B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance, IEEE Transactions on Visualization and Computer Graphics, v.14 n.6, p.1221-1228, DOI: 10.1109/TVCG.2008.172
[40] J. Davies Word Trees, https://www.jasondavies.com/wordtree/, Accessed on 11-5-2017.
[41] Q. Le, and T. Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML14), pp. 1188–1196.
[42] R. Reh ˇ u˚ˇrek, & P. Sojka. 2010. A software framework for topic modeling with large corpora. LREC.
[43] M. Richardson. 2009. Principal Component Analysis, URL: http://people.maths.ox.ac.uk/richardsonm/SignalProcPCA.pdf, retrieved Nov 30, 2017.
[44] L.v.d. Maarten, and G. Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, Vol 9(Nov), pp. 2579—2605.
[45] M. Wattenberg, F. Viégas, and I. Johnson. 2016. How to Use t-SNE Effectively. Distill, DOI: http://doi.org/10.23915/distill.00002
[46] Sinclair, Stéfan, Geoffrey Rockwell and the Voyant Tools Team. 2012. Voyant Tools (web application).
[47] W.B. Paley. 2002. TextArc: Showing Word Frequency and Distribution in Text. In Proceedings of IEEE Symposium on Information Visualization, Poster Compendium, IEEE CS Press.
[48] P. Lanzi. 2015. Visualization Techniques in Data Mining, ppt @ UIC 583.
[49] D3.js - Data-Driven Documents, http//d3js.org/, Accessed on 12-4-2017.
[50] B. Johnson and B. Shnedierman. Tree-maps: A space-filling approach to the visualization of hierarchical information. In Visualization 1991, pages 284--291. IEEE, 1991.