Linköping University | Department of Computer and Information Science
Master's thesis, 30 ECTS | Datateknik
2021 | LIU-IDA/STAT-A--2021/044--SE

Text Mining Methods for Biomedical Data Analysis

Text Mining Metoder för Biomedicinsk Data Analys

Rakhshanda Jabeen

Supervisor : Sebastian Sakowski
Examiner : Anders Nordgard

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Abstract

Biological data topic modelling has become a very prevalent topic among researchers in recent times. However, analysing countless research papers and gathering consensus regarding biomedicine is a near impossible task for any researcher due to the complexity and quantity of material that is published. This thesis is devised to focus on two objectives that can help the researchers in this domain based on data related to five major DNA repair pathways. The first objective is to propose an unsupervised approach to examine the hidden structures and analyse research trends in temporal biomedical text data. The second objective is to find DNA repair markers involved in immune defence and retrieve potential PPIs, GIs, and disease-gene associations reported in the literature. We have used Latent Dirichlet Allocation (LDA) to discover hidden themes and semantically coherent topics from text. We have clustered the documents based on LDA topic models to analyse the research trend and used the Mann- Kendall test to understand the trends of the topics. Hybridization of text mining methods with classical co-occurrence statistical approach and association rule mining was used to discover potential PPIs, GIs, and disease-gene association in the text. The results for PPIs and GIs were then evaluated with an external biological database of PPIs.

Acknowledgments

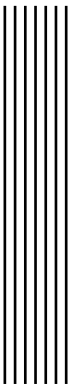
"To Him belongs the dimension of the heavens and the earth, it is He who gives life and death, and He has power over all things." (Al-Quran) All acclamations are to ALLAH, the most merciful and compassionate, who has empowered and enabled me to accomplish this task.

First of all, I would like to thanks my supervisor Sebastian Sakowski for his guidance and support in carrying out my master's thesis. I want to show my sincere gratitude to Prof. Tomasz Popławski from the Department of Molecular Genetics, University of Lodz, for his support and discussion throughout the project. Furthermore, I would also like to thank my examiner, Anders Nordgard, from Linköping University, for his valuable inputs.

Contents

List of Figures

List of Tables



Acronyms

API Application Program Interface. 3, 6

BoW bag-of-words. viii, 3, 10, 19, 39

FDR false discovery rate. 3, 40, 41

FPM frequent pattern mining. 3, 16–18

GI Genetic Interaction. iii, v, 3, 4, 19, 22, 23, 31–33, 39–42, *Glossary:* Genetic interactions

GO Gene Ontology. 3, 5, 23, 24, 26, 28, 31–33, 38, 40, 42

LDA Latent Dirichlet Allocation. iii, 3, 5, 11, 12, 19, 20, 27, 28, 30, 38–42

MK Mann-Kendall. 3, 13, 21, 30, 39, 42

NER named entity recognition. vii, 3, 14, 21–23, 25, 40

NLP Natural Language Processing. 3, 22, 26

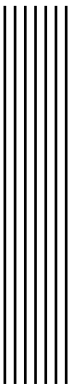
PMI pointwise mutual information. 3, 15, 22–24, 31, 33, 40

POS parts of speech. 3, 14, 21, 41, 43

PPI Protein-Protein Interaction. iii, v, 3–5, 19, 22–24, 31–33, 39–42, *Glossary:* Protein-protein interactions

UMLS Unified Medical Language System. 3, 22–25, 40

VSM vector space model. 3, 9



Glossary

Autoimmune disease A condition in which immune system mistakenly attacks body. 3, 35

Biomarker An indicator of normal/abnormal condition or disease in the body. 3, 33, 38, 39

DNA Deoxyribonucleic Acid (DNA) is a polymer that encodes genetic information in living organisms. 3

DNA damage An abnormal alteration in the chemical structure of DNA. 3

DNA repair A collection of processes by which a cell identifies and corrects damage to the DNA molecules that encode its genome. 3, 31

DNA repair marker Enzymes that recognize and corrects physical damage in DNA. 3

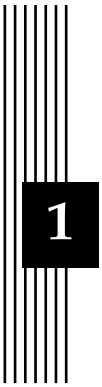
DNA repair pathways A mechanism of DNA damage, repair and genetic mutations. 3, 39

Gene A basic physical unit of inheritance, made up of DNA. 3

Genetic interactions (GIs) capture functional association between genes. 3

Protein A molecule (product of gene) that facilitates most of the biological processes in the cell. 3

Protein-protein interactions (PPIs) detect the physical interactions between gene products. 3



1 Introduction

1.1 Motivation

With the rapid growth in biological technology, the study of biomedical science has witnessed an exponential growth of knowledge over time—researchers from all around the world share discoveries, new phenomena, and experimental results in life science journals in the form of electronic texts. A massive amount of biological knowledge exists in life science research articles, and the domain experts and researchers cannot keep track of developments in the field and trending subject areas. Manual transformation and retrieval of the knowledge from enormous unstructured biomedical text is a laborious task. Text mining methods with Natural Language Processing (NLP) and machine learning approaches provide a solution to the information overloading problem [21].

DNA is a genetic material and damage in DNA causes numerous human pathologies such as *cancer*, *premature aging* and *chronic inflammatory conditions*. In contrast, there are DNA damage sensors proteins encoded in the gene of the organisms that repair DNA damage, and such proteins are usually referred to as DNA repair markers. Several recent studies have shown a strong connection between the DNA repair process and the immune system. This interesting interplay has shifted the viewpoint in clinical studies for autoimmunological diseases, and cancer treatment [38]. Molecular biologists and clinical researchers yearn to track the new research done in DNA repair pathways and discover potential associations between autoimmune diseases and DNA repair markers.

Recently, scientific communities have developed a great interest in biological-data topic modeling and information extraction methods of text mining to extract specific biological information from the literature. One of the most researched topics in biomedical text mining is the retrieval of protein-protein interactions (PPIs) and genetic interactions (GIs) from literature. PPIs and GIs play a vital role in cellular functions and biological processes, and the discovery of such interactions leads to a better understanding of disease mechanism and development of drugs for efficient treatment of diseases[53]. Other than this, several researchers used text mining methodologies to retrieve disease-gene associations and drug-

disease associations from the biomedical text based on their co-occurrence in a fixed context.

The primary motivation of the thesis is to propose an unsupervised approach to examine the hidden structures in biomedical literature concentrated in DNA repair pathways and identify and analyze the research trends over time. The reason for digging this information is to recognize paradigmatic shifts in evolving research trends and direct researchers towards the most trending areas of interest in this domain. Another motivation of this project is to leverage text mining methods to discover DNA repair markers involved in DNA damage repair and immune defense and discover PPIs, GIs and disease-gene associations reported in the biomedical literature.

1.2 Aim

The objective of this thesis can be divided into two categories. The first objective of the thesis is to provide an unsupervised machine learning approach to examine the biomedical literature and identify research trends. This can be further subdivided into the following sections:

- Identify semantically coherent and intelligible topics in research papers
- Use visualization methods to interpret the topics
- Identify research trends over time periods
- Propose a method to visualize the trends in research papers

Semantically coherent topics imply such topics that depict strong semantic relation between top keywords of topics and summarizes the content of documents in few terms. The main focus of this section will be to discover such underlying topics from research papers in an unsupervised manner and, based on these topics, propose an overall trend analysis of research. The second objective of this project is to leverage information extraction methods of text mining and statistical approaches to discover the following biological information from the research papers:

- Find DNA repair markers that are involved in DNA damage response and immune defense
- Retrieve potential PPIs and GIs reported in research papers
- Discover disease-gene associations discussed in research papers

1.3 Research questions

The thesis will be focused on answering the following research questions:

1. Recommend an unsupervised machine learning approach to discover hidden themes and semantically coherent topics from unstructured biomedical text?
2. How can we analyze and visualize the evolving research trends in unstructured text data over time?
3. Recommend text mining methods to retrieve relationships between biological entities such as PPIs, GIs and disease-gene associations from biomedical literature?

1.4 Related Work

In machine learning methods, topic modeling is frequently used to overview the hidden semantic structures in the voluminous textual data. Many researchers have developed an interest in biological-data topic modeling due to an exponential growth in the biomedical literature to help researchers interpret trending topics and summarize the vast textual data [29]. Evangelopoulos et al. [18] recommended that topic modeling compared with hard clustering is a better choice to capture the temporal trends in the textual data. Blei et al. [9] developed a generative probabilistic model Latent Dirichlet Allocation (LDA) and describes the use of the model for summarization and classification of electronic corpora.

For preprocessing of data, Schofield et al. [49] suggests that except for the removal of commonplace words, removing domain specific stopwords has a shallow impact on the quality of the model, whereas stemming can even worsen the quality of inferred topics as topic model inference often places word sharing morphological roots in the same topic. Röder et al. [47] proposed a coherence metric named Cv to evaluate the quality of topic models. Their algorithms compute the coherence of topics based on the high probable words in the topics. Sharma et al. [52] used three different topic modeling techniques together with LDA to analyze and visualize the evolution of research topics in the domain of machine learning. They used Mann-Kendall [32][27] test to identify significant trends in topics.

Nowadays, text mining computational methods are commonly used to extract the proteins, and their interactions from biomedical literature and data repositories (Papanikolaou et al. [41]). The systems proposed to identify relationships between entities in the text generally assume that associated entities occur in the same sentence (Ray et al.[45] and Barnickel et al.[6]). Bunescu et al. [12] integrated information extraction methods with co-occurrence statistics to retrieve PPIs from biomedical literature and rank these interactions based on the information-theoretic measure pointwise mutual information. Al-Aamri et al. [1] proposed a system to automatically extract genetic interactions to predict disease-gene associations from text based on co-occurrence frequency at three different levels, i.e., sentence level, abstract level, and semantic level with network analysis. Min et al. [23] developed a tool *PPI Finder* to mine PPIs from PubMed [43] articles based on the co-occurrence frequency of proteins and shared Gene Ontology (GO) terms between proteins. Another approach by Sun et al. [57] also utilizes GO terms annotation of genes to predict an association between genes.

Cellier et al. [13] utilized a hybridization of sequential pattern mining and information extraction methods to discover genetic interactions and contextual information of genes from biomedical literature. Their proposed system automatically produces patterns conveying gene interactions and their characterization at a sentence level. Alves et al. [5] used association rule mining with FP-grow algorithm [22] to perform gene association analysis on gene expression data. Hristovski et al. [25] leveraged sequential pattern mining on MEDLINE articles to discover disease candidate genes from biomedical literature.

For the computation of statistical significance of association rules, Alvarez et al. [4] proposed a method to compute chi-squared χ^2 test statistics as a function of interestingness measures of association rules.

2 Data

2.1 Data

The source of data for this project is PubMed [43]. PubMed is one of the most comprehensive sources of information in biomedicine. It is a system of National Library of Medicine (NLM)[1] containing more than 32 million citations for biomedical literature from MEDLINE (Medical Literature Analysis and Retrieval System Online), life science journals, and online books. Figure 2.1 represents the growth of the number of new citations and the total number of citations in the PubMed database over the past six decades. It is evident from the figure that there is an exponential growth of knowledge in biomedicine and clinical studies.

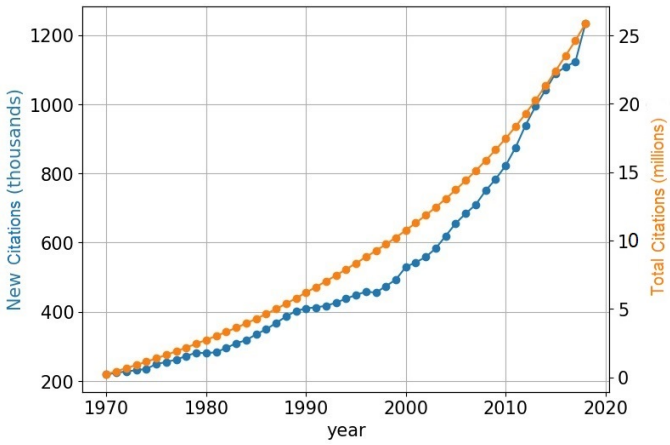


Figure 2.1: Growth of the number of citations in PubMed databse

National Center for Biotechnology Information[2] (NCBI) provides several public Application Program Interfaces (APIs) to programmatically access and download data from PubMed. We have used ENTREZ[3] programming E-utilities (*esearch* & *efetch*) to access and download

[1] www.nlm.nih.gov/

[2] www.ncbi.nlm.nih.gov/

[3] An integrated Information Retrieval System to explore links between biological sequences and texts.

data from PubMed. Research papers related to five major DNA repair pathways are downloaded from PubMed using the following keywords:

1. Base excision repair (BER)
2. Nucleotide excision repair (NER)
3. Non-homologous end joining (NHEJ)
4. Mismatch Repair (MMR)
5. Homologous recombination repair (HRR)

We exclusively fetched the titles, abstracts, year of publication, and a unique id assigned by PubMed to each published paper. The downloaded dataset comprises 97881 records. For this project, titles and abstract text of the research papers are used for topic modeling and information extraction tasks. Time information such as the publishing year of the paper is used for trend analysis. Figure2.2 represents a sample record from the dataset.

PMID	Year	Title	Abstract
22794911	2012	A candidate gene study of one-carbon metabolism pathway genes and colorectal cancer risk.	The risk of colorectal cancer (CRC) may be influenced by aberrant DNA methylation and altered nucleotide synthesis and repair, possibly caused by impaired dietary folate intake as well as by polymorphic variants in one-carbon metabolism genes. A case-control study using seventy-one CRC patients and eighty unrelated healthy controls was carried out to assess the genetic association of fifteen SNP and one insertion in nine genes belonging to the folate pathway. Polymorphism selection was based on literature data, and included those which have a known or suspected functional impact on cancer and missense polymorphisms that are most likely to alter protein function. Genotyping was performed by real-time PCR and PCR followed by restriction analysis. The likelihood ratio statistic indicated that most of the polymorphisms were not associated with the risk of CRC. However, an increased risk of CRC was observed for two variant alleles of SNP mapping on the transcobalamin 2 gene (TCN2): C776G (rs1801198) and c.1026-394T>G (rs7286680). Considering the crucial biological function played by one-carbon metabolism genes, further investigations with larger cohorts of CRC patients are needed in order to confirm our preliminary results. These preliminary results indicate that TCN2 polymorphisms can be a susceptibility factor for CRC.

Figure 2.2: A sample of data from downloaded articles

We have concatenated the title and abstracts of all papers and created one text document to address all research questions. Figure 2.3 represents the distribution of the number of words of the papers. The average length of the titles & abstract text in the dataset is 217.

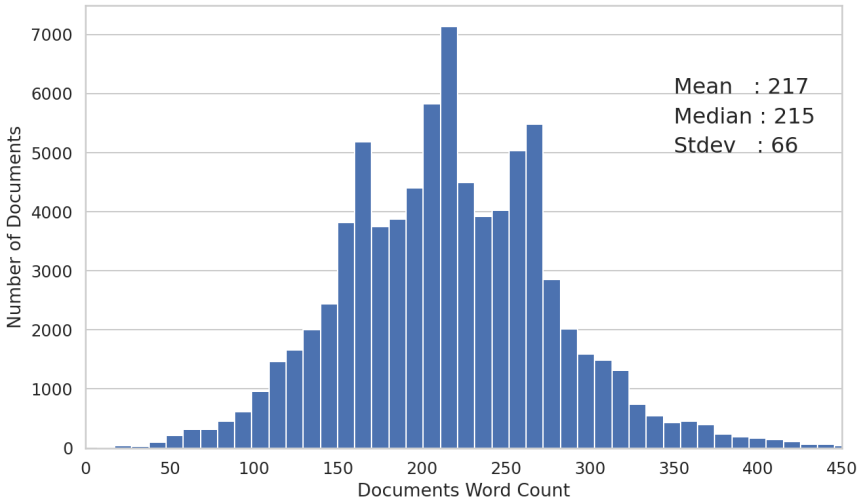


Figure 2.3: Distribution of Number of Words

Figure 2.4 is a visual interpretation of the number of articles published per year. By looking at the bar plot, we can infer that this research area has gained popularity in the last two decades.

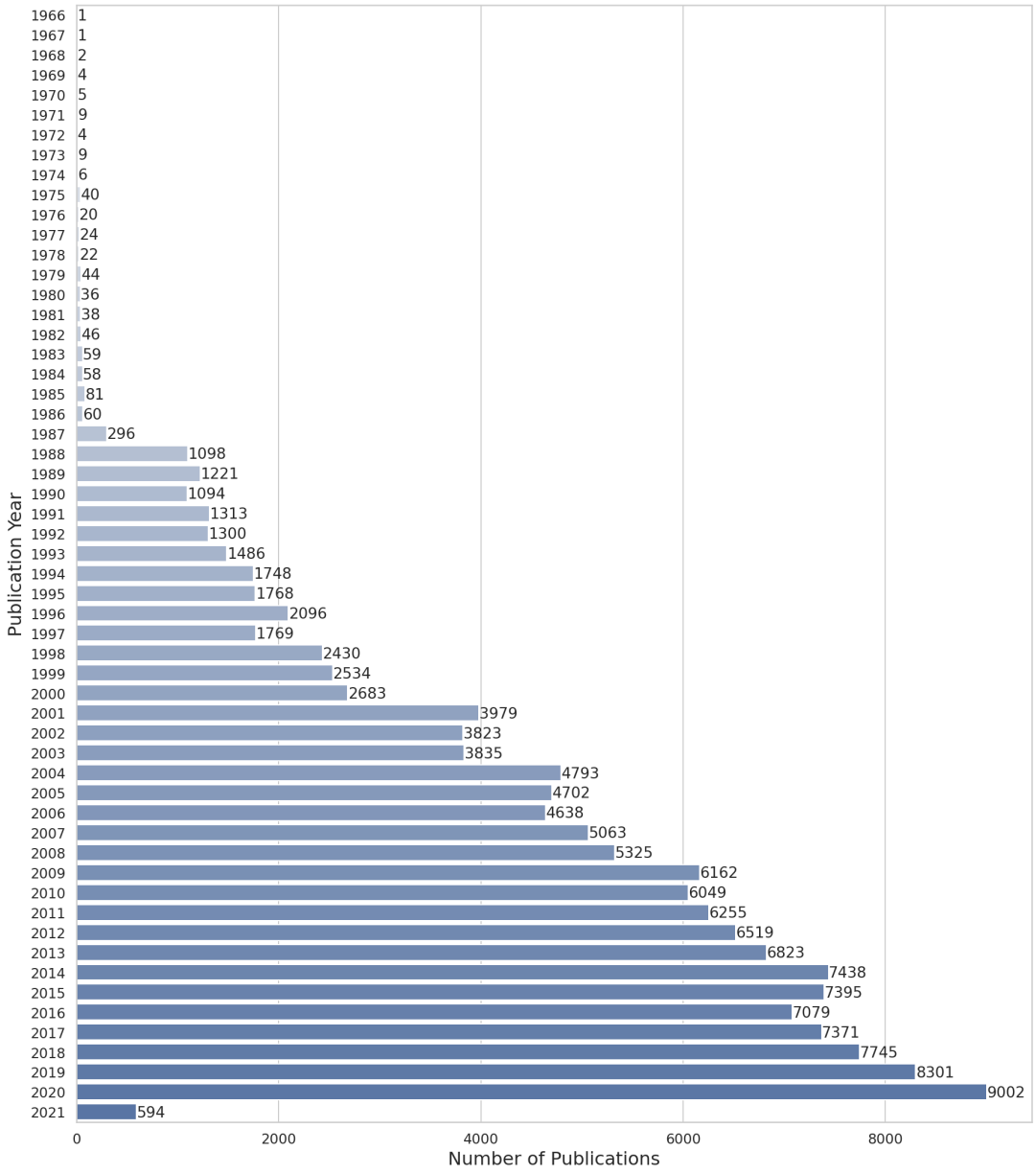


Figure 2.4: Number of publications per year

3 Theoretical Background

3.1 Text Representation

Text mining has grasped many researchers' attention in the domain of machine learning to automatically extract structured and human-readable information from unstructured text. A fundamental problem in text mining is dealing with the representation of data that makes it mathematically computable. In the following sections, we will discuss the types of text representation used in this thesis.

3.1.1 Vector Space Model

One of the most commonly used text representation technique is the vector space model (VSM) [48]. In VSM each document in the corpus[1] is represented as a binary vector in the word space. Let us suppose that $V = \{w_1, w_2, w_3, ..., w_n\}$ be the terms in the corpus and $D = \{d_1, d_2, d_3, ..., d_m\}$ be the documents in the corpus. Then we can create a document-term matrix M of order $m \times n$ on the basis of presence of a term in the document. If a term k is present in a document d then $d[k] = 1$ and $d[k] = 0$ otherwise. This representation of text documents is widely used in information retrieval [16].

Let us consider a sample example of a corpus with three documents, $D1$: "Ali is happy because Sara is happy", $D2$: "Ali and Sara are happy, $D3$: "Ali is not happy because Sara is not happy". Initially, a vocabulary of terms present in the corpus is built, and a document-term matrix M is computed. In this matrix, rows correspond to the documents, and columns correspond to the terms. Table 3.1 is an illustration of the binary representation of the documents in the word space.

Table 3.1: Document-term matrix of vector space model

	ali	and	are	because	happy	is	not	sara
D1	1	0	0	1	1	1	0	1
D2	1	1	1	0	1	0	0	1
D3	1	0	0	1	1	1	1	1

[1] collection of documents

3.1.2 Bag-of-Words Model

A common approach for extracting vectors from text is the traditional bag-of-words (BoW) model [60]. In this representation, each document in the corpus is considered a bag of words. Weights are assigned to the terms in a document based on the term's frequency in the document. This notion is generally known as the term weighing. For example, suppose we have M documents in the corpus and N terms in the vocabulary. Then a document-term matrix of order $M \times N$ can be computed as discussed in 3.1.1 but here, each cell in the matrix represents the term's occurrence frequency in the corresponding document. A BoW representation of the documents is illustrated in the table 3.2 by using the same example as discussed in 3.1.

Table 3.2: Example BoW representation of count vectors

	ali	and	are	because	happy	is	not	sara
D1	1	0	0	1	2	2	0	1
D2	1	1	1	0	1	0	0	1
D3	1	0	0	1	2	2	2	1

This representation of the documents does not account for the exact ordering of the words, but the term's occurrence frequency is substantial. The drawback of this representation is that it fails to preserve the semantic and syntactic relationships between words and suffers from sparsity and high dimensions curse [51]. A document can also be described as a set of *n-grams*, a contiguous sequence of n words in the text. In this representation, weights are assigned to n words based on their co-occurrence in the document [50]. This text representation model helps capture the dependencies between the terms that are occurring together in the text. BoW model is a special case of *n-grams* model also known as unigram model.

3.1.3 Co-occurrence Word Vectors in a Fixed Context Window

Firth [19] is frequently cited for saying, "You shall know a word by the company it keeps." Therefore mostly word vectors are driven by the idea that words that share semantic relationships will occur in the same context, and this representation of text can preserve the semantic and syntactic dependencies of the words in the text.

A word-word co-occurrence matrix M is based on the co-occurrence frequency of words in a fixed context. Each row and column in the matrix corresponds to a word in the corpus. The idea behind this approach is that for each word w_i in the corpus consider a fixed context window of size n. Then n preceding and successive words of the word w_i, that is $w_{i-n}, w_{i-(n-1)}, ..., w_{i+1}$ and $w_{i-1}, w_{i+2}, ..., w_{i+n}$ are the words in the context window of w_i. The cell M_{ij} in the matrix M is the number of times a word w_j appears in the context window of w_i. An illustration of the co-occurrence word-word matrix with a context window of size 2 is represented in table 3.3 by using the same example as discussed in 3.1 [54].

Table 3.3: Co-occurrence word-word matrix with a context window of 2

	and	not	are	is	ali	happy	because	sara
and	0	0	1	0	1	0	0	1
not	0	0	0	2	1	2	1	1
are	1	0	0	0	0	1	0	1
is	0	2	0	0	2	4	3	2
ali	1	1	0	2	0	1	0	1
happy	0	2	1	4	1	0	2	4
because	0	1	0	3	0	2	0	2
sara	1	1	1	2	1	4	2	0

3.2 Topic Modeling

Topic models are probabilistic generative models extensively used in text mining and information retrieval systems. It is an unsupervised machine learning approach to discover the underlying semantic structure of a given collection of text documents [8]. The two most widely used probabilistic topic modeling techniques are Latent Dirichlet Allocation (LDA) [9] and probabilistic Latent Semantic Analysis (pLSA)[24]. These methods provide a soft clustering of the documents in the corpus, assuming that every document belongs to each topic to some degree.

3.2.1 Latent Dirichlet Allocation (LDA)

LDA was initially proposed by Blei et al. [9]. They introduced it as a three-level Bayesian hierarchical model. The principle of the method is that each document m in the corpus M can be expressed as a finite mixture of K topics, and each topic k can be expressed as a finite mixture of words. This approach allows a soft clustering of the corpus, making each document a combination of several topics. Plate notation of LDA as demonstrated in figure 3.1, is a compact way to visually represent the dependencies of the model parameters. M, N, and K represent the number of documents, the number of words in the corpus, and the number of topics, respectively. All nodes in the circles represent the random variables of the model and they are defined as follows:

$\alpha-$ parameter of the Dirichlet prior on topics distribution per document
$\beta-$ parameter of the Dirichlet prior on words distribution per topic
θ_m- topic distribution of the document m
ϕ_k- word distribution of the topic k
$z_{mn}-$ topic for the n^{th} word in the document m
$w_{mn}-$ word in the corpus vocabulary

The only observable variable of the model is w, which is grayed in the diagram; all the other variables are latent. Plates (rectangles) represent the repetition of variables and, depending on where they fall inside plates, indicate whether they are applied at the document level, topic level, or word level.

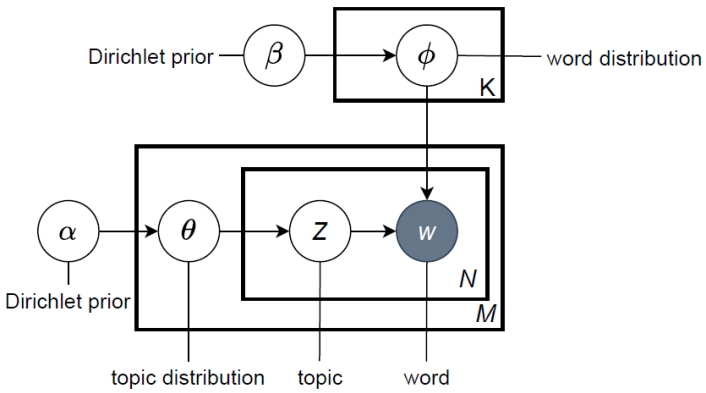


Figure 3.1: Graphical representation of LDA [9]

Sparse Dirichlet priors are used for modeling the topics distribution over documents and words distribution over topics as proposed in the original paper [9]. The parameters α and β controls the sparsity of topics distribution and words distribution, respectively. Bayesian inference can be used to get the posterior distribution of the latent variables θ_m, $z_{m,n}$, and ϕ_k

given the words. The generative process of LDA to infer K topics from the corpus with M documents is as follows [28]:

1. For each topic $k \in \{1, 2, ..., K\}$
   - Draw a distribution over words $\phi_k \sim Dirichlet(\beta)$
2. For each document $m \in \{1, 2, ..., M\}$
   - Draw a distribution over topics $\theta_m \sim Dirichlet(\alpha)$
   - For each position n in document m
     - Draw a topic assignment $z_{m,n} \sim Multinomial(\theta_m)$
     - Draw a word from vocabulary $w_{m,n} \sim Multinomial(\phi_{z_{m,n}})$

Direct inference of the posterior distributions is complex; thus, Gibbs sampling [20] can be used to approximate the posterior distributions [28]. Table 3.4 is an illustration of the LDA topic distribution on a general domain corpus with three topics and top five words. Each word in the topic is assigned a weight; this weight indicates the word's importance in the topic. It can be rightly inferred that topic-1 covers the information about education or an educational institute, while topic-2 is mostly related to science.

Table 3.4: LDA topic-word distributions with 3 topics

Topic-1		**Topic-2**		**Topic-3**	
word	weight	word	weight	word	weight
college	0.07	space	0.09	court	0.03
students	0.05	solar	0.04	judge	0.01
teacher	0.03	nasa	0.02	case	0.007
education	0.003	satellite	0.007	prison	0.003
syllabus	0.003	moon	0.003	culprit	0.001

3.3 Evaluation of the LDA Topic models

Evaluation of the topic models is a challenging task as it is an end-to-end unsupervised learning method. Generally, topics coherence metrics are used to quantify the semantic similarities between words of each topic of a probabilistic topic model. The two most common and widely used topic coherence metrics are Cv metric [47], and UMass metric [35]. These metrics imitate human judgment while measuring the semantic relations between high probability terms of the topics.

3.3.1 Cv Metric

Röder et al. [47] proposed a unifying framework to quantify the coherence score of topics. The workflow of framework comprises of four steps, i.e., segmentation ($\mathcal{S}$), probability calculation ($\mathcal{P}$), confirmation measures ($\mathcal{M}$) and aggregation (Σ). In the first step, the set of words is segmented into subsets of pair of words. For instance if $T = \{t_1, t_2, t_3, t_4, t_5\}$ is a topic with top five high probable words, then T can be segmented into subsets of pair of words as shown in 3.1.

$$S_i = \{T^{'} = (t_i), T^{*} = T\}, \qquad 1 \leqslant i \leqslant 5 \tag{3.1}$$

In the second step, the joint probability of words is estimated based on the reference corpus. This probability is computed using the co-occurrence frequency of the words within a fixed context window divided by the total number of documents in the corpus. In the third

step, the segmented subsets $S_i = (T', T^*)$ and their corresponding probabilities are used to compute the set of confirmation measures $(\mathcal{M})$. The confirmation measure quantifies how strongly T' supports T^*. In a general sense, how often a term t_i co-occur with other terms in the topics relative to when it occurs alone in the corpus. In the final step, confirmation measures of all subset pairs S_i are aggregated into a single coherence score (c) by taking mean. Topics with high values of **Cv** score are considered semantically coherent and self-explanatory.

3.4 Mann-Kendall Test

Mann-Kendall (MK) is a non-parametric test developed by Mann [32], and Kendall [27] to detect increasing or decreasing monotonic trends in time-series given that there is no serial correlation in data. In this test, the null hypothesis H_0 is that there exists a trend in the time-series, and the alternative hypothesis H_a is that there is no trend in the time-series. The MK test statistics (S) is computed as defined in equation 3.2, where X_i and X_j are sequential values of the time-series in the time-periods i and j respectively [52].

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(X_j - X_i), \qquad j > i \tag{3.2}$$

where,

$$sign(X_j - X_i) = \begin{cases} 1 & \text{if } (X_j - X_i) > 0 \\ 0 & \text{if } (X_j - X_i) = 0 \\ -1 & \text{if } (X_j - X_i) < 0 \end{cases}$$

If n is the number of data points in time-series, q is the number of tied groups[2], and t_p is the number of data points in p^{th} group, then the variance of S and standardized test statistic (Z_{MK}) is computed as defined in the equations 3.3 and 3.4 respectively.

$$Var(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p}^{q} t_p(t_p-1)(2t_p+5)\right] \tag{3.3}$$

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & \text{if } (X_j - X_i) > 0 \\ 0 & \text{if } (X_j - X_i) = 0 \\ \frac{S+1}{\sqrt{Var(S)}} & \text{if } (X_j - X_i) < 0 \end{cases} \tag{3.4}$$

A positive value of Z_{MK} indicates an increasing trend in the data, and a negative value indicates a decreasing trend in the data. Time series does not exhibit any trend if the value of Z_{MK} is zero. If the number of observations in the time series is greater than 10, then the distribution of Z_{MK} is approximately standard normal. Based on the 0.05 significance level of the test statistics, if the *p-value* is less than 0.05, then the null hypothesis is rejected in favor of the alternative hypothesis. Thus we will say that there is a trend in data.

3.5 Information Extraction

Information extraction is a process of extracting structured information from unstructured or semi-structured text [26]. It is a two-phase process; first: named entity recognition in the text, second: semantic relations extraction from the identified entities. Figure 3.2 demonstrates a visual interpretation of information extraction workflow. The first two steps of the first phase

[2]A tied group is the number of samples in data having the same value.

are related to the prepossessing of the text. In **Sentence segmentation**, text documents are divided into individual sentences. Sentence segmentation is usually performed by separating text from full stops (".") in the text. In the **tokenization**[3] phase segmented sentences are tokenized to words.

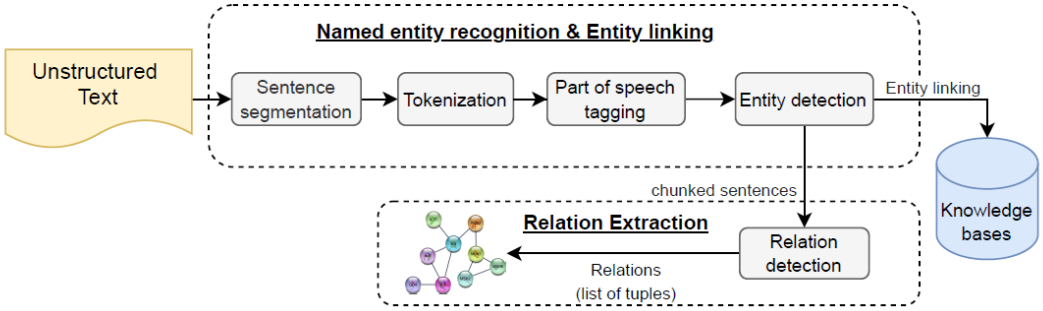


Figure 3.2: The schematic diagram of information extraction work flow

Several computational methods have been proposed for the other steps involved in the information extraction process. Significant methods include rule-based approaches, Hidden Markov Models, Conditional Random Fields, Support Vector Machines, Machine Learning, and Deep Learning approaches [37]. We will discuss the short descriptions of the other steps in the following sections.

3.5.1 Part of Speech Tagging

Parts of speech (POS) are the noun, pronoun, verb, adverb, adjective, prepositions, and conjunction etc in the linguistics. Given a sequence of words, POS tagging is the process of categorizing the sequence base of their POS. Machine learning or deep learning models can be trained to predict the POS of a word given a large annotated dataset.

3.5.2 Named Entity Recognition

The aim of named entity recognition (NER) in the biomedical text is to identify and categorize the entities such as genes, proteins, drugs, disease, and chemicals mentioned in the text. NER in the biomedical literature is one of the most daunting tasks due to the complexities of biomedical terminologies [40]. There is no standard way of mentioning biological entities; different authors may use different nomenclature, spellings, abbreviations, and formatting in their research papers. For example, "p53" and "TP53" are two different ways to mention the same protein. Standard models for NER comprises Maximum Entropy Markov Models, Conditional Random Fields, and Long short-term memory recurrent neural networks. Moreover, several open-source frameworks and tools that are trained on the domain-specific corpus are available for NER in the biomedical text in multiple programming languages.

3.5.3 Entity Linking

Entity linking or entity normalization is the process of matching the named entities with a reference knowledge base[4]. As discussed above, named entities in the biomedical literature are ambiguous. For example, genes or proteins may have multiple aliases and acronyms;

[3]Tokenization is the process of splitting a text document into smaller units such as words.
[4]Knowledge bases store the structured and unstructured information in a machine-readable way.

sometimes, the same entity may correspond to multiple concepts depending on the context. Therefore, entity disambiguation is a crucial step of information extraction from biomedical literature. Table 3.5 shows a few renowned knowledge bases to comprehend structured information about biomedical terminologies.

Table 3.5: Popular biomedical knowledge bases

Knowledge base	Link
Unified Medical Language System (UMLS)	`www.nlm.nih.gov/research/umls`
Uniprot knowledge base (UniProtKB)	`www.uniprot.org/`
Gene Ontology (GO)	`http://geneontology.org/`
Human Phenotype Ontology (HPO)	`https://hpo.jax.org/app/`
Medical Subject Headings (MeSH)	`www.nlm.nih.gov/mesh/meshhome.html`

3.5.4 Relation extraction

Relation extraction is the final phase of the information extraction workflow 3.2. The objective of this task is to identify semantic structures and associations between identified named entities. Types of relations to be extracted from unstructured text must be defined by the domain experts and analysts. Consider an example from a biomedical sentence: "*BRCA1 interacts with RAD51*", yielding a relation *interact(BRCA1, RAD51)*. In this example, *BRCA1* and *RAD51* are two named entities (genes), and interaction is a relation between these entities.

Several techniques have been proposed for relation extraction from the biomedical text in the last two decades. These techniques are categorized into four different groups, co-occurrence based, pattern based, rule based, and machine learning based approaches [36]. Machine learning and deep learning methods are providing a state-of-the-art solution to relation extraction problems. Generally, a large set of annotated or labeled data is needed to train machine learning models for relation extraction. On the other hand co-occurrence approach is a classical technique used to discover association amongst entities. The intuition of this approach is that if two entities co-occur in a context frequently, then there might be a relation between these entities. Several statistical ranking measures such as pointwise mutual information (PMI), chi-square (χ^2) or log-likelihood ratio uses the entities co-occurrence frequency to identify whether terms in a context possess a strong relation or their co-occurrence is due to chance [33].

3.6 Association Rule Mining

Association rule mining is one of the most extensively used data mining techniques initially proposed by Agrawal et al. [2]. It aims to extract interesting correlations, frequent patterns, and associations among sets of items in data repositories. One of the most famous applications of association rule mining is market basket analysis. Nowadays, association rules are widely used in various domains such as telecommunication networks, market risk management, inventory control, and biomedical [61]. The most famous algorithms for association rule mining from traditional databases are the apriori algorithm [2] and the FP-grow algorithm[22].

Rule Mining Concepts

In this section we will discuss the concepts related to rule mining. For this purpose assume that, $I = \{i_1, i_2, i_3, ..., i_m\}$ be the set of m items and $T = \{t_1, t_2, t_3, ..., t_n\}$ be the set of n transactions called database.

Association Rule: The formal definition of an association rule proposed in [2] states that a rule is defined as an implication of the form $X \rightarrow Y$, where X and Y are disjoint subsets of I. The left-hand side and the right-hand side of the rule are named antecedent and consequent, respectively.

Support of an itemset: *Support of an itemset X, $sup(X)$ is defined as the proportion of transactions containing X.*

$$Sup(X) = P(X) = \frac{|X \subseteq t|}{|T|}, \qquad \forall t \in T$$

Since databases are large and users concern only the frequent itemsets in the transactions; thus a minimum support constraint (*minSup*) is used to drop the infrequent itemsets. An itemset X is called *frequent itemset* if $Sup(X)$ exceeds or equals the provided threshold of *minSup*.

Support of a rule: *Support of a rule $X \rightarrow Y$, is defined as the fraction of transactions that contains X and Y to the total number of transactions.*

$$sup(X \rightarrow Y) = sup(X \cup Y) = P(X \cap Y) \qquad \text{range: } [0, 1]$$

Confidence of a rule: *Confidence of a rule $X \rightarrow Y$, is the fraction of the number of transactions that contains X and Y to the number of the transaction containing X.* In other words, confidence of a rule is the conditional probability of Y given X and it provides the strength of a rule.

$$Conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} = \frac{P(X \cap Y)}{P(X)} \qquad \text{range: } [0, 1]$$

Lift of a rule: *Lift [11] of an association rule $X \rightarrow Y$ measures how often X and Y co-occur than expected if they are statistically independent.* This measure is used to infer the dependence between antecedent and consequent of a rule. If the rule has a lift of 1, then we can infer that X and Y are independent.

$$lift(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X) * sup(Y)} = \frac{P(X \cap Y)}{P(X) * P(Y)} \qquad \text{range: } [0, \infty]$$

3.6.1 Apriori Algorithm

In general association rule mining can be viewed as a two step process, frequent pattern mining (FPM) and rule generation. The apriori algorithm was initially proposed by Agrawal et al. [2] for FPM and rule generation from traditional databases. Briefly the apriori algorithm process is as follows:

1. Scan the database and generate the frequent itemsets of size one (F_1) provided the *minSup* threshold.
2. Using the frequent itemset F_k of size k, generate C_{k+1} candidates set of size $k+1$ by joining F_k by itself.
3. Go through all the generated candidates and select itemsets whose support equals or exceeds provided the *minSup* threshold.

Frequent Pattern Mining with Apriori Algorithm

FPM is the process of finding all frequent itemsets X in transactions for which $sup(X)$ equals or exceeds the user-provided *minSup* threshold. Apriori property for FPM states that, *every subset of a frequent itemset is frequent or every superset of an infrequent itemset is infrequent.* The algorithm uses a "bottom-up" iterative approach known as *candidate generation*. Table 3.6

represents the pseudo code of apriori algorithm for FPM. The *apriori-gen*(F_k) function in 3.6 used join and prune steps to generate the candidates for frequent itemset F_k. Given the input of k^{th} frequent itemset F_k, the function generates the $k+1^{th}$ sets of candidates C_{k+1}. The candidate generation procedure is described in table 3.7.

Table 3.6: Psseudo-code of Apriori algorithm for FPM [30]

Apriori Algorithm
Input: A transactional database T and $minSup$
Output: All frequent itemsets in T

F_1 = frequent items of size 1
for $k = 1, F_k! = \Phi, k++$ do:
    C_{k+1} = apriori-gen(F_k) # new candidates generated from F_k
    for all $t \in T$ do:
        C_t = subset(c_{k+1}, t) # candidates belonging to transaction t
        for all $c \in C_t$ do:
            $c.count++$ # increment the count of all candidates in C_{k+1}
            end
        $F_{k+1} = \{c \in C_{k+1} \mid c.count \geqslant minSup\}$
        end
    end
return $\bigcup_k F_k$

Table 3.7: Pseudo-code of apriori candidate generation function [3]

Apriori Algorithm: apriori-gen(F_k)
Input: k^{th} frequent itemset F_k
Output: A superset of F_k

$C_{k+1} = \Phi$ # Join step
for all $I, J \in F_k$ do:
    if $I_1 = J_1, ..., I_{k-1} = J_{k-1}$ & $I_K < J_K$ then:
        add $I_1, I_2, ..., I_K, J_k$ to C_{k+1}
for all $c \in C_K$ do: # prune step
    for all $k-$subsets s of c do:
        if $s \notin F_k$ then:
            remove c from C_{k+1}
return C_{k+1}

Rule Generation with Apriori Algorithm

Given a frequent itemset F_k, $2^k - 2$ association rules can be generated from it, where k is the number of items in F_k. All of these association rules are not equally interesting; thus a minimum threshold on *confidence* of the rule is used to prune the discovered association rules [5] and the process is called rule generation. The apriori property for rule generation states that, *If the rule $X \rightarrow F - X$ does not satisfy the **minConf** constraint, then neither of the subsets X' of X will satisfy the constraint* [3].

$$Confidence(X \rightarrow F - X) = \frac{support(F)}{support(X)} \geqslant \frac{support(F)}{support(X')} = Confidence(X' \rightarrow F - X')$$

Given a frequent itemset F_k where $k \geqslant 2$, we can use *gen-rule* function as described in table 3.8 to generate strong association rules from F_k.

Table 3.8: Pseudo code of apriori rule generation algorithm [3]

Algorithm: gen-rules $(F_k, x_k, minConf)$
Input: A frequent itemset F_k, a subset x_m of F_k and $minConf$ threshold
Output: All the rules of the form $x \rightarrow F_k - x$ with $x \subseteq x_m$ and confidence equal or above $minconf$.

$X = \{(m-1)$ itemsets $x_{m-1} | x_{m-1} \subseteq x_m\}$
for all $x_{m-1} \in X$ do:
  $confidence = \frac{support(F_k)}{support(x_{m-1})}$ #confidence of the rule $x_{m-1} \rightarrow F_k + x_{m-1}$
  if $confidence \geqslant minConf$ then:
    output $x_{m-1} \rightarrow F_k + x_{m-1}$
    if $m - 1 > 1$ then:
      call gen-rules$(F_k, x_{m-1}, minConf)$

An example of the association rule mining process with the apriori algorithm is presented in figure 3.3. The $minSup$ and $minConf$ thresholds for this example are 0.5 and 0.8 respectively. Therefore each itemset must appear in at least 50% of the transactions to be a frequent itemset, and each association rule must have a confidence above or equal to 80% . In FPM phase, candidate itemsets are generated, and their support is calculated. The itemsets whose support is not equal or greater than the $minSup$ threshold are eliminated and colored gold in the figure. The resulting frequent itemsets are combined to generate new candidates at each iteration. At the end of FPM phase, the last set of frequent itemsets is used to generate association rules. In the final phase, rules are prunes based $minConf$ constraint.

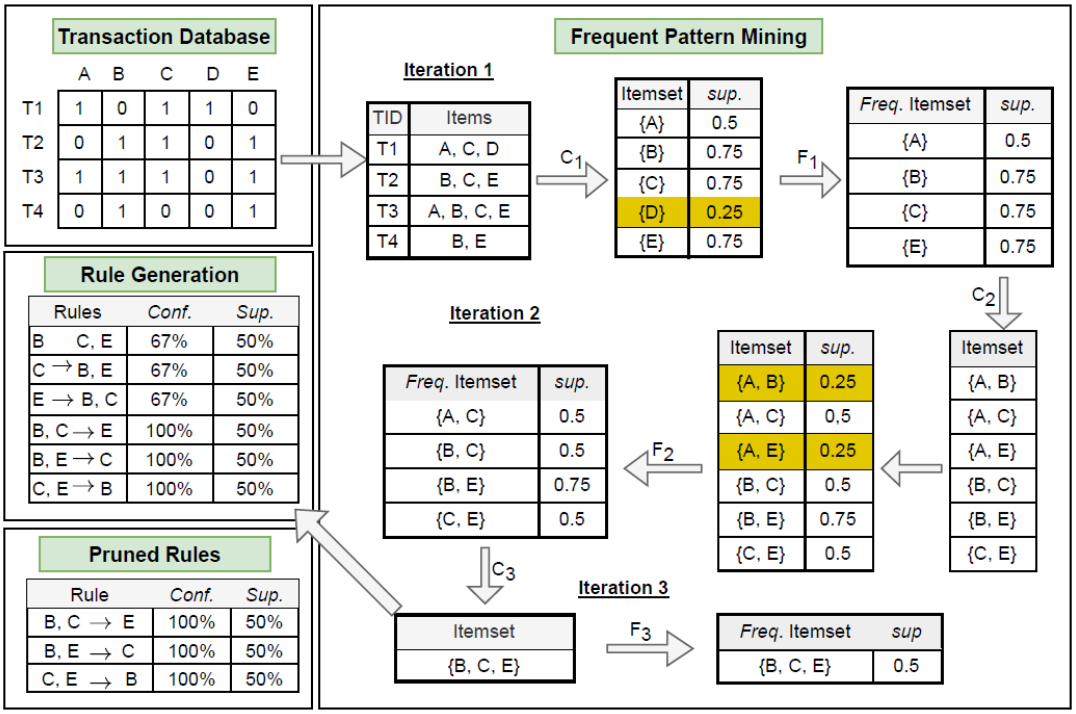


Figure 3.3: The process of association rule mining with apriori algorithm

4 Methodology

In this section, we will discuss methods used to achieve the objectives of the thesis. Traditional machine learning model LDA is used for topic modeling and trend analysis. For the discovery of PPIs, GIs and disease-gene association, we have opted for hybrid approaches of information extraction methods with co-occurrence statistics and association rule mining, respectively.

4.1 Topic Modeling and Trend Analysis of the Articles

LDA is a very appropriate choice to discover the trending research areas in the unstructured articles. In this section, we will discuss the LDA topic modeling and trend analysis of the titles and abstracts of the research papers.

4.1.1 Preprocessing

Most of the research papers related to the DNA repair pathway are published in the last two decades 2.4. Therefore, we dropped all the papers published before 2000 and obtained 80696 documents for topic modeling. In the preprocessing step, documents are split into word tokens and lowercased except genes and proteins in the text. This process makes it easier to identify the most discussed genes in a topic. The pipeline used to extract genes or proteins from the text will be discussed later in this chapter. Punctuations, tokens with less than two characters, and English stopwords were removed from the text. We have not performed any further preprocessing as suggested by Schofield et al.[49]. Their research quantifies the impact of different forms of preprocessing on the LDA topic model. In their experiments, they discovered that stemming of words even worsens the topic model quality, i.e., *topic coherence*, *model likelihood* or *classification accuracy* as topic model inference often places word sharing morphological roots in the same topic. Thus we have avoided performing such time-consuming tasks.

4.1.2 LDA Topic Modeling

LDA model takes in two input parameters, number of topics and a BoW representation of documents. Thus post preprocessing, we have generated a BoW representation of all the doc-

uments in the corpus as described in 3.1.2. The second major input parameter of LDA model is the number of topics that should be chosen before modeling the data. It is hard to evaluate how meaningful and interpretable topics are generated from LDA. However, the topic coherence measure as discussed in 3.3 can be used to choose the optimal number of topics required for the corpus. We have used Gensims's [46] topic coherence pipeline *coherencemodel* to find the optimal number of topics based on the Cv metric [47]. In figure 4.1 the x-axis represents the number of topics, and the y-axis represents the corresponding coherence score of the model. We can see that the model with 14 topics has the highest coherence score (0.53). Thus we have chosen LDA model with 14 topics for the representation of our data.

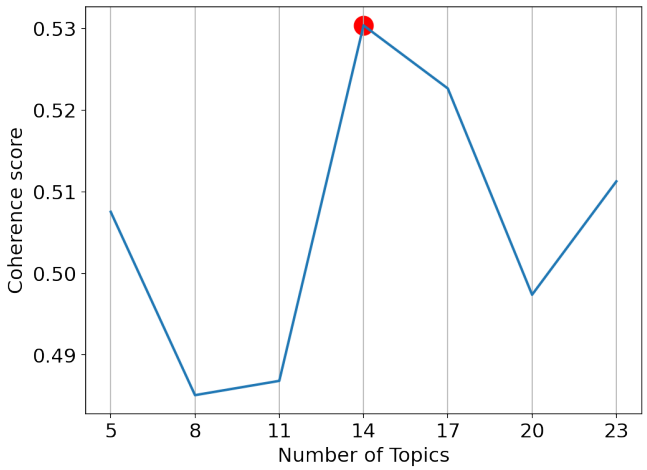


Figure 4.1: Coherence scores for six different number of topics

The LDA model outputs two matrices; a document-topic matrix of dimension $(M \times K)$, where M is the number of documents and K is the number of topics, and a topic-word matrix of dimension $(K \times N)$, where N is the size of corpus vocabulary. Each cell in the document-topic matrix represents the percentage contribution of a latent topic in a document. Similarly, each cell in the topic-word matrix represents the percentage contribution of words in each latent topic. Figure 4.2 represents an illustration of whole process of the LDA topic modelling.

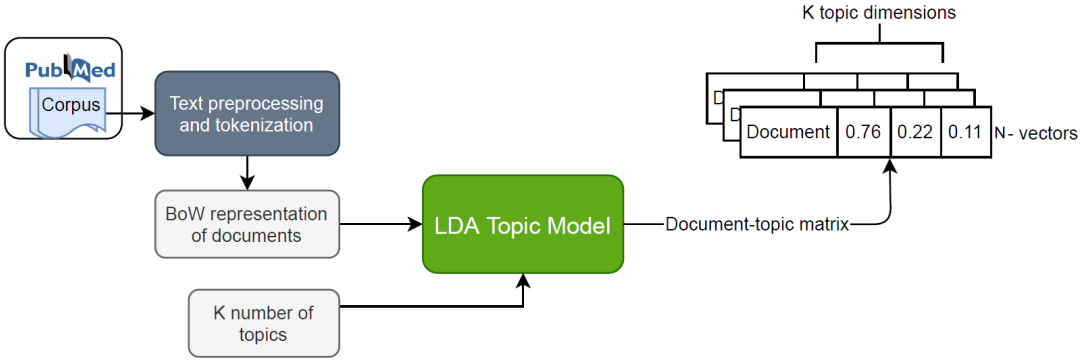


Figure 4.2: LDA topic modeling process

4.1.3 Trend Analysis of Topics with Mann-Kendall Test

Discovering the trending topics in the corpus is a challenging and technical task. We have used the same strategy proposed by Sharma et al. [52] for trend analysis of the research articles in data. They used MK test as discussed in 3.4 to identify the significant trends based on the time series of the latent topics. Each document is a mixture of 14 topics and the same topic will be dominant in similar documents. Hence, we have divided all the articles into 14 segments based on the most dominant or representative topic to cluster documents. The normalized frequency of publication ($\#nf_{P_y}$) for each topic cluster is calculated by dividing the count of similar articles per year ($\#sa_y$) by the total number of articles published in the same year ($\#ta_y$) as shown in equation 4.1. For example, if similar articles clustered in 2010 are 10 and a total number of articles published in 2010 is 50, then the normalized frequency of publication is 0.2.

$$\#nf_{P_y} = \frac{\#sa_y}{\#ta_y} \tag{4.1}$$

4.2 ScispaCy for Named Entity Recognition and Entity Linking

ScispaCy [39] is an open-source framework in Python for processing biomedical, scientific, and clinical text. It is built over spaCy [55], a robust python library for general domain text. Figure 4.3 represent the text processing pipeline of spaCy. The pipeline consists of the steps tokenization, POS tagging, dependency parsing, and named entity recognition. Additional pipeline components can also be added to the existing pipeline for more specific tasks, such as entity linking.

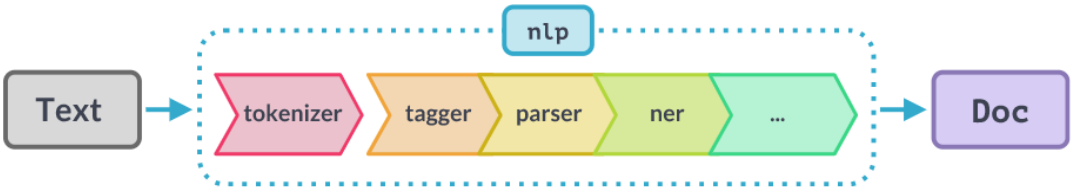


Figure 4.3: The schematic diagram of spaCy NLP pipeline[55]

4.2.1 Named Entity Recognition with ScispaCy

ScispaCy provides different NER models for more specific entity recognition tasks. The model **en_ner_bionlp13cg_md** is used to identify the genes or gene products (proteins) and cancer in the text and **en_ner_bc5cdr_md** model is used to identify diseases other than cancer in the text. Mostly genes or proteins are mentioned in the titles of the article; that is why we have concatenated the titles and corresponding abstracts of all the articles before starting the NER process for each method discussed in the following sections. Figure 4.4 demonstrates an illustration of NER on an abstract of the dataset. Entities in the text are categorized into four proper categories, GENE_OR_GENE_PRODUCT, CELL, CANCER, and CELLULAR_COMPONENT. The colors in the image are associated with different categories.

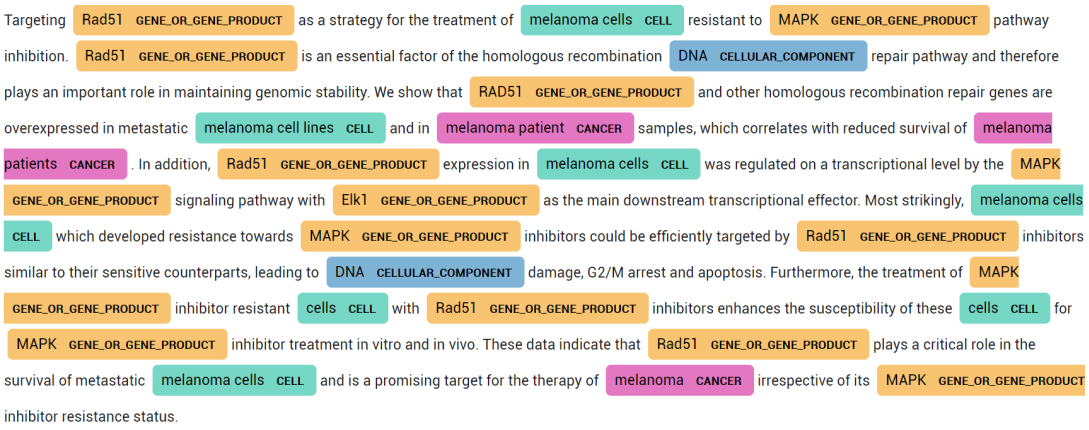


Figure 4.4: Example of NER on biomedical text

4.2.2 Entity Linking with ScispaCy

An additional pipeline component of scispaCy, *EntityLinker* is used to link the recognised named entities to Unified Medical Language System (UMLS) knowledge base [10]. UMLS is a collection of vocabularies associated with the biomedical and clinical domain containing 3 million concepts. EntityLinker pipeline performs string overlap-based search on entities to match them with UMLS concepts. The entities which are not found in UMLS knowledge base are dropped. For this project, we are only interested in genes or proteins associated with humans; thus, we have filtered genes related to humans only.

4.3 Hybridization of Co-occurrence Statistics Approach with Information Extraction Methods for Retrieval of PPIs & GIs

An integration of information extraction methods and co-occurrence statistics is used to discover the potential PPIs and GIs at a sentence level incited by Bunescu et al. [12]. The intuition of the approach is that if two genes/proteins are mentioned together in the abstract sentences frequently, then there might be an association between them. They used information-theoretic measure PMI [14] to rank the strength of an association between proteins. PMI is used to quantify the log-likelihood of the words' co-occurrence in Natural Language Processing (NLP). For two genes, g_1 and g_2 the PMI measure is computed based on the following quantities:

- N : the total number sentences containing gene pairs.
- $P(g_1, g_2) = \frac{n_{12}}{N}$: probability of co-occurrence of g_1 and g_2 in the same sentence; n_{12} is the number of sentences mentioning both g_1 and g_2.
- $P(g_1, g) = \frac{n_1}{N}$: probability that g_1 co-occur with any other gene g in the same sentence; n_1 is the number sentences mentioning g_1 and g.
- $P(g_2, g) = \frac{n_2}{N}$: probability that g_2 co-occur with any other gene g in the same sentence; n_2 is the number sentences mentioning g_2 and g.

Then PMI of g_1 and g_2 is defined as,

$$PMI(g_1, g_2) = log \frac{P(g_1, g_2)}{P(g_1, g)P(g_2, g)} = logN \frac{n_{12}}{n_1 n_2}$$

Following the strategy of Bunescu et al. [12], we have omitted the *log* operator for simplicity as PMI will only be used to rank the association strength between two genes/proteins. Moreover, we have not differentiated between genes and their products (proteins) as we are trying to find both GIs and PPIs. Thus we will mention both genes and gene products (proteins) as genes in the following sections for brevity.

4.3.1 PPIs & GIs Retrieval Model

The workflow of PPIs and GIs retrieval model from research papers is demonstrated in figure 4.5. In the information extraction steps, documents are segmented into sentences before starting the NER process. All sentences containing less than two genes were dropped. Then we generated a dictionary of all genes in the sentences and linked these entities with UMLS knowledge base to filter genes associated with humans. Based on these filtered genes, a gene-gene co-occurrence matrix M is computed as discussed in 3.1.3, keeping the context window size equal to the length of the sentence. Each entry $M_{i,j}$ of the matrix M represents the number of times gene, g_i and g_j appears in the same sentence. Based on the co-occurrence matrix M, we computed the matrix of PMI scores of all genes and generated a function to retrieve associated gene pairs automatically. Given a minimum threshold for co-occurrence frequency, the model returns a list of associated genes for query genes along with their PMI values.

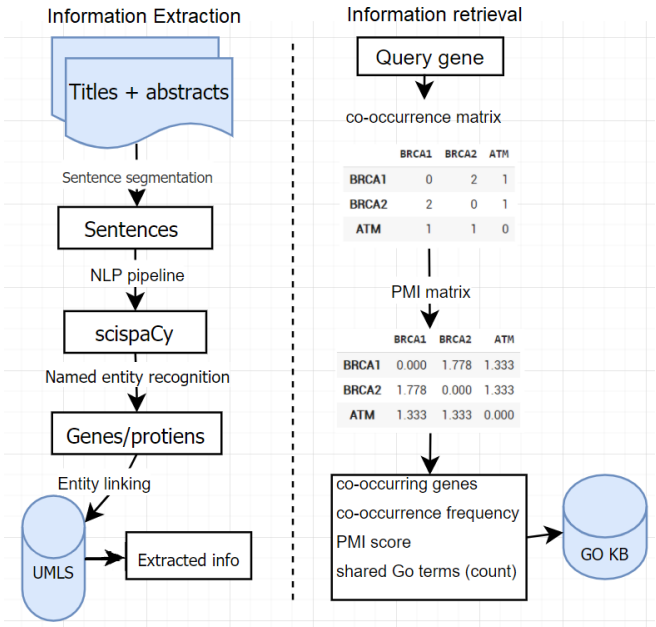


Figure 4.5: The schematic diagram of PPIs & GIs retrieval model workflow

Inspired by Min et al. [23] and Sun et al. [57] we have used another approach to predict an association between genes and rank the association strength based on Gene Ontology (GO) [34] annotations. GO is a popular biological ontology to classify the set of genes based on three functionalities; biological process, cellular component, and molecular function called GO terms. We have performed GO enrichment analysis on all genes identified in the data. Enrichment analysis determines which of the GO terms are over-represented in the input set of genes and provides a list of significant GO annotations. Based on these annotations, we can count how many times a pair of genes share the same biological functionality and

denote this count as *shared GO terms*. A higher value of shared GO terms between two genes indicates a stronger association between them.

Figure 4.6 is an illustration of the model output of top 7 associated genes ranked based on higher PMI value for a query *MSH6* with a minimum co-occurrence threshold of 5. Columns CUI(G1) and CUI(G2) are unique concept identifiers from UMLS knowledge base. The pairs are sorted in descending order based on the count of the shared GO terms.

CUI(G1)	G1	CUI(G2)	G2	PMI	shared GO-terms
C0879393	MSH6	C4522159	MSH2	4.797539	194
C0879393	MSH6	C0879389	MLH1	4.134187	179
C0879393	MSH6	C3711796	ATM	0.204186	125
C0879393	MSH6	C1705526	TP53	0.251721	119
C0879393	MSH6	C3811684	NBN	0.595561	117
C0879393	MSH6	C0219474	BAX	2.672789	105
C0879393	MSH6	C0879392	MSH3	4.750497	104

Figure 4.6: Top 7 predicted genes associated with *MSH6*

4.3.2 Evaluation of the predicted Results

The most common metrics to evaluate the performance of information extraction tasks are precision (P), recall (R), and F1-score. Precision is the fraction of correctly predicted true associations to the total number of predicted associations. Recall, also known as sensitivity, is the fraction of correctly predicted association to the total number of true associations. F1-score is the harmonic mean of precision and recall.

Benchmark datasets are necessary to evaluate the performance of the model's predictions, and a negative control set is required to compute the precision and recall [41]. To the best of our knowledge, we do not have any reference data to verify the predicted interactions of genes studied in this data. However, we can find the precision as the fraction of predicted association that found a match in an external database containing information about PPIs similar to Min et al. [23] strategy. They have matched their results with external databases, and GO terms [34]. Hence for evaluation of our results, we have matched our results with external PPI database STRING[1] [58].

4.4 Association Rule Mining for Literature-based Discovery of Disease Candidate Genes

A hybridization of information extraction methods and association rule mining is used to discover the association between disease and genes in the articles inspired by Cellier et al. [13]. We have replicated their study in our project to discover the association between genes and disease. This method aims to automatically discover association rules of the form $disease_A \rightarrow gene_B$ from genes and disease mentioned in the articles at the sentence level. These rules mean that if $disease_A$ and $gene_B$ are discussed together frequently in several sentences of the abstracts, then it is highly likely that there is an association between them.

[1]STRING is a popular biological database and an online tool containing experimentally verified and predicted protein-protein interactions.

4.4.1 Association Rule Mining process

The Association rule mining process for the disease-genes association is completed in four phases as demonstrated in the figure 4.7.

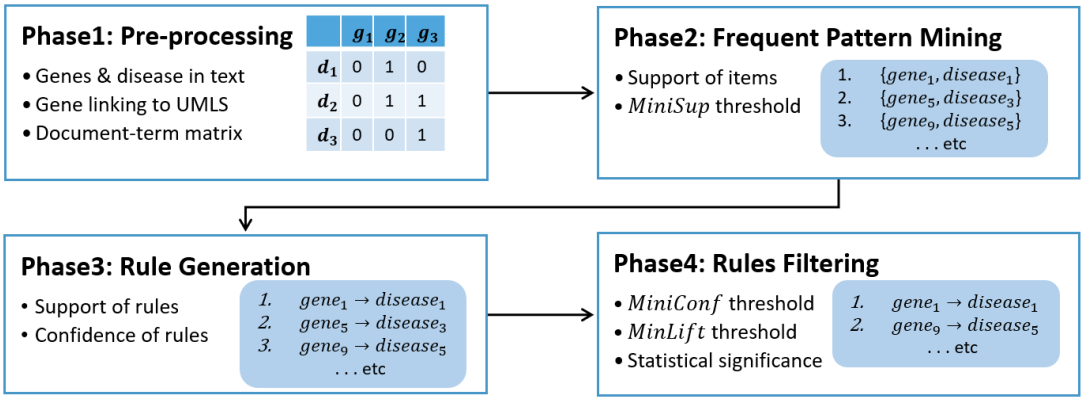


Figure 4.7: Association rule mining process for discovering disease-gene association

Preprocessing

In the preprocessing phase, we have segmented the abstracts into sentences and used scispaCy NER models to extract the disease and genes mentioned in sentences. Authors commonly mention genes, proteins, and disease with their abbreviations. For example, *acute myeloid leukemia* is often mentioned as AML in the articles. This can affect the rule mining process as the system will identify them as two different entities. To counter this problem, we have used scispaCy's *AbbreviationDetector* pipeline to detect abbreviations of the disease and cancer mentioned in the text and replaced them with their full form. All genes found in the sentences are linked with UMLS knowledge base for entity disambiguation. Moreover, we have joined disease names containing more than one word with a hyphen (-) to make a single token. Finally, genes and diseases found in each sentence are separated into different sets and named documents. We have only kept documents that contain both genes and disease, yielding a database of 24170 documents. Documents are encoded into binary vectors, and a document-term matrix is generated as discussed in 3.1.1.

Frequent Pattern Mining and Rule Generation

Alves et al. [5] suggest that the apriori algorithm performs better on sparse datasets where the frequent patterns are short. Since each processed document contains only a few disease/gene names resulting in a sparse dataset; thus we have used the apriori algorithm for frequent pattern mining and rule generation as discussed in 3.6.1. The choice of $minSup$ and $minConf$ thresholds is a well-known problem in association rule mining. A very high value of $minSup$ yields a few association rules; conversely, a small value results in very many association rules. Since the occurrence of gene name with a disease in research papers is a rare event, thus we have used smaller values of $minSup$ threshold to find association rules as suggested in [13]. We have run the apriori algorithm for a maximum of two iterations and mined association rules of length 2 when there is precisely one gene/disease in both antecedent and consequent of rules.

Ranking and Pruning of Association Rules

The limitation of the association rule mining algorithm is that it generates many redundant and spurious rules. Thresholds on confidence and lift are used to prune the association rules. The problem with these measures is that they are primarily influenced by the support of the consequent and antecedent of the rule. To counter this problem and control the false discovery rate (FDR), we have used the chi-square test [42] to measure the statistical significance of the association rules. The null hypothesis H_0 of the test is that the two variables are independent, and the alternative hypothesis H_a is that variables are dependent. Alvarez et al. [4] proposed a method to compute chi-square (χ^2) statistics of association rules as a function of support, confidence, and lift of the association rule. According to their formula:

$$\chi^2 = n * (lift - 1)^2 \left[\frac{supp * conf}{(conf - sup)(lift - conf)} \right]$$

In this formula, n represents the total number of observations (transactions). Based on the χ^2-test statistics, we have calculated the *p-value* of all association rules at a significance level of 0.05. If the *p-value* is less than 0.05, then null hypothesis H_0 is rejected in favor of the alternate hypothesis. Thus we will say that there exists a relationship between the antecedent and consequent of the rule.

4.5 Implementation

All the experiments are performed in google colab. Tesla V100-SXM2-16GB GPU, provided by google colab was used for named entity recognition and entity linking.

Programming Language & Softwares

- The project is implemented in programming language Python 3.6.9.
- Biopython [15] library is used to retrieve data from PUbMed [43].
- General Machine learning and NLP tools used for the project are MLxtend[44], NLTK[7], Gensim[46], SpaCy [55] and ScispaCy [39].
- Network graphs are generated with Cytoscape [56] and Cytoscape's StrinApp [17] is used to perform GO enrichment analysis.

5 Results

In this section, we will present results obtained on the basis of experimentation as discussed in the methodology section 4. The experimentation involves LDA topic modeling and information extraction methods of text mining on biomedical textual data.

5.1 LDA Topic Modeling & Trend Analysis

In this section we will share the results of LDA topic modeling on the *titles* and *abstracts* of the articles. Figure 5.1 represents the top 12 most probable words related to each latent topic generated by LDA.

Topic1	MGMT, methylation, colorectal, status, MLH1, cases, temozolomide, crc, samples, methods, glioblastoma, prognostic
Topic2	PCNA, dnapkcs, ku70, ku, functional, WRN, ku80, infection, mitotic, host, antigen, subunit
Topic3	PARP1, brain, acute, neurons, PARP, polyadpribose, neuronal, blood, diseases, xp, skin, mrna
Topic4	p53, uv, mitochondrial, skin, ATM, fibroblasts, pol, oxygen, ultraviolet, uvinduced, transcription, uvb
Topic5	biological, understanding, systems, structural, structures, molecules, interactions, research, number, pol, various, diseases
Topic6	ap, extracts, adduct, abasic, mitochondrial, mtdna, bases, crosslinks, topoisomerase, sensitivity, incision, removal
Topic7	XRCC1, CI, controls, breast, allele, XPD, cases, population, among, individuals, snps, subjects
Topic8	transcription, meiotic, FA, conserved, regions, drosophila, rearrangements, domains, cerevisiae, functional, essential, breakpoints
Topic9	immune, vdj, nonhomologous, ATR, immunoglobulin, csr, lymphocytes, AID, switch, somatic, ATM, DNAPK
Topic10	bone, wound, matrix, days, healing, culture, methods, cartilage, melanoma, microm, proliferation, folate
Topic11	changes, proliferation, muscle, transcription, reaction, expressed, mrna, uracil, cdna, tissues, migration, revealed
Topic12	breast, BRCA1, BRCA2, PARP, prostate, combination, sensitivity, therapeutic, drugs, ERCC1, advanced, inhibitor
Topic13	nonhomologous, dsb, hr, endjoining, efficient, crisprcas9, efficiency, reca, rad51, ligase, nucleases, nuclease
Topic14	mmr, MLH1, MSH2, colorectal, microsatellite, cases, hnpcc, deletions, nonpolyposis, sporadic, colon, MSH6

Figure 5.1: Top 12 most probable words related to 14 topics of LDA model

Representation of LDA topics as word clouds provides a better visualization of the terms related to each topic. Figure 5.2 represents the word clouds of all the topics. The font size of each term in the figure expresses their relative weight in the topics.

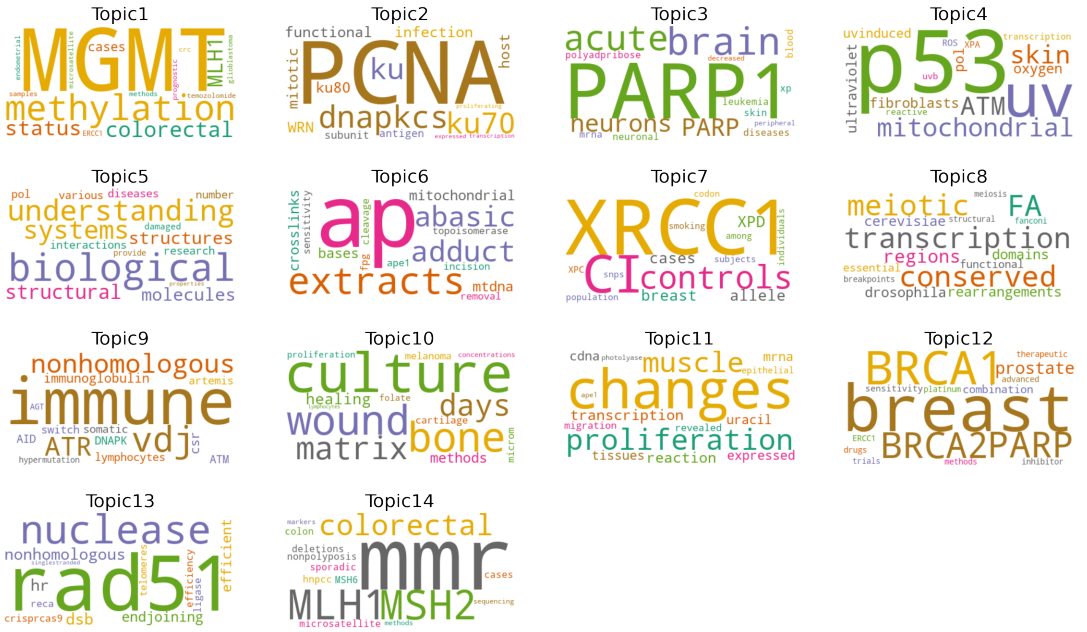


Figure 5.2: Word clouds of the LDA topics

We have used GO terms to annotate the genetic components occurring in the topic keywords to understand the biological functions of these components. Based on GO annotations and other top 15 keywords of each topic, we have assigned labels to each latent topic. Table 5.1 represents the assigned labels to each topic.

Topic ID	Topic Label
Topic-1	MGMT methylation in glioblastoma & colorectal cancer
Topic-2	PCNA & antigens(DNA-PKcs, KU70, KU80, WRN)
Topic-3	Leukemia & regulation of neuron death(PARP, PARP1)
Topic-4	UV radiations & skin cancer
Topic-5	Bioloical system, structures & interactions
Topic-6	DNA adducts & abasic
Topic-7	Nucleotide excision repair(XRCC1, XPD, XPC) & SNP
Topic-8	Fanconi anemia & DNA transription
Topic-9	Adaptive immune system & DNA repair markers
Topic-10	Cartilage wound healing & melanoma
Topic-11	DNA proliferation & transcription
Topic-12	Breast cancer & prostate cancer inhibitors
Topic-13	NHEJ & Homologous repair (HR) & double-strand break repair
Topic-14	Microsatellite instability & DNA mismatch repair

Table 5.1: Labels of topics based on top keywords of LDA and GO terms

5.1.1 Trend Analysis of the Articles with LDA Topic Models

Using the results of LDA topic models, we have clustered the documents based on the most dominant topic. Figure 5.3 shows the percentage of documents falling within each cluster. We can clearly see in the figure that topic-2, topic-3, topic-9 and topic-14 shares the biggest number of documents as the dominant topic. Figure 5.4 represents the stacked counts plot of documents falling within a topic over the past 20 years.

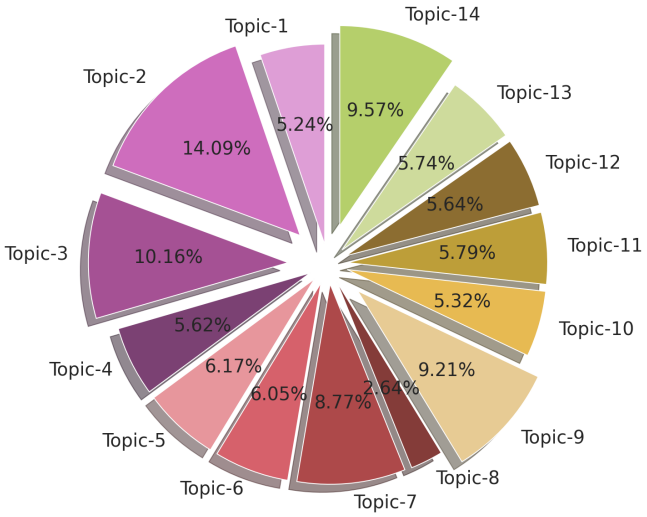


Figure 5.3: Documents clusters based on the most dominant topic

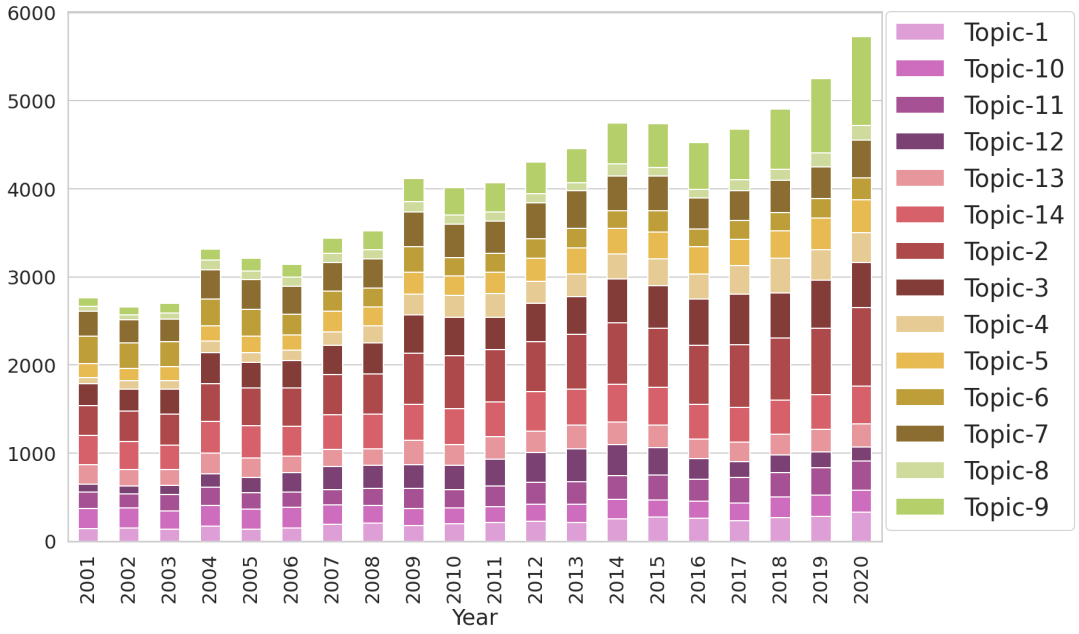


Figure 5.4: Trending topics based on the most dominant topic of articles over past 20 years

The time-series plots in figure 5.5 represent the trending topics of the articles in the period $2001 - 2020$. The x-axis represents the publishing year, and the y-axis represents the normalized frequency of publications of each topic cluster. Based on these time-series plots, we can infer that topic-2, topic-4, and topic-9 depict a clear upward trend, while topic-6, topic-7, topic-10, topic-13, and topic-14 represent a definite downward trend over the past 20 years.

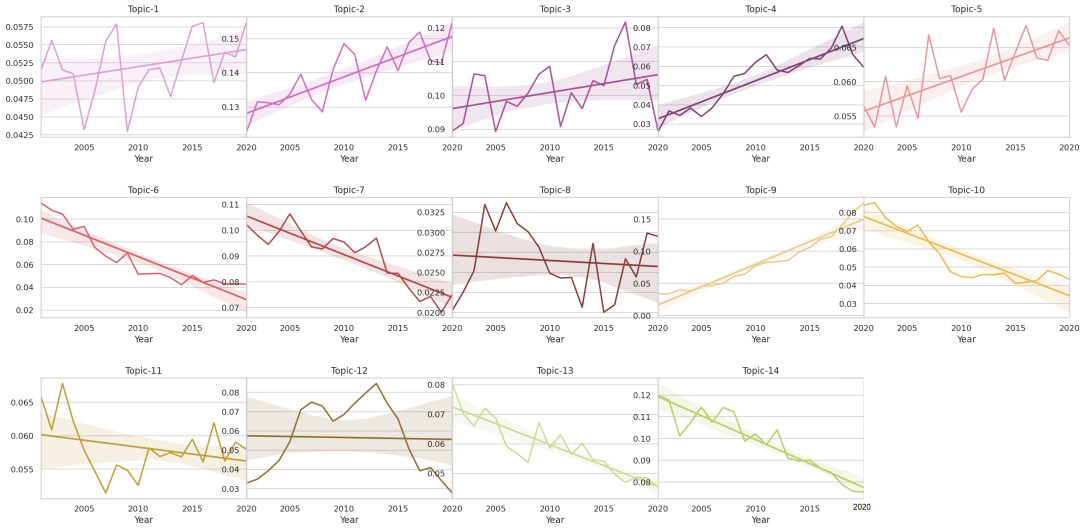


Figure 5.5: Trend analysis of LDA topics on the articles in period 2001-2020

Mann-Kendall Test Results

Table 5.2 shows the results of MK test on the time-series of LDA topic clusters. The hypothesis is tested at a significance level $\alpha = 0.05$. Thus for any p-value less than 0.05, we have rejected the null hypothesis (no trend in data) in favor of the alternate hypothesis. Column Z_{MK} shows the values of standardized statistics. A higher absolute value of Z_{MK} and a lower p-value represent a strong upward or downward trend in the time-series. Based on MK test results, we can say that topic-9 shows the strongest upward trend and topic-14 shows the strongest downward trends.

Topic ID	Trend	p-value	Z_{MK}
Topic-1	no trend	0.0979	1.655
Topic-2	increasing	0.0002	3.731
Topic-3	no trend	0.1442	1.459
Topic-4	increasing	4.77E-06	4.575
Topic-5	increasing	0.0026	3.0173
Topic-6	decreasing	2.49E-07	-5.159
Topic-7	decreasing	2.14E-05	-4.250
Topic-8	no trend	0.5376	-0.616
Topic-9	increasing	1.95E-09	6.0022
Topic-10	decreasing	6.59E-05	-3.990
Topic-11	no trend	0.6265	-0.487
Topic-12	no trend	0.9225	0.0973
Topic-13	decreasing	3.49E-06	-4.639
Topic-14	decreasing	6.91E-07	-4.964

Table 5.2: Mann-Kendall test results on time-series of LDA topic clusters

Some of the inferences that we have made based on topic modeling and trend analysis are:

- *Topic-9*, adaptive immune response & somatic recombination is the top trending topic in the domain of DNA repair pathways.

- *Topic-2*, DNA repair marker PCNA and antigens[1] (KU70, KU80, WRN, DNAPKcs) are the most discussed genetic markers with immune response.
- *Topic-12*, breast cancer & prostate cancer inhibitors remained trending topic in the period 2010 – 2015.
- *Topics-1*, *Topics-3*, *Topics-8* and *Topics-11* shows no statistically significant changes of the trend.

5.2 Retrieval of PPIs and GIs with Information Extraction Methods & Co-occurrence Statistics

In this section, we will show the results of PPIs and GIs predicted by using an integration of classical co-occurrence statistics with information extraction methods. Figure 5.6 represents the top 10 most discussed genes and proteins in DNA repair pathways dataset.

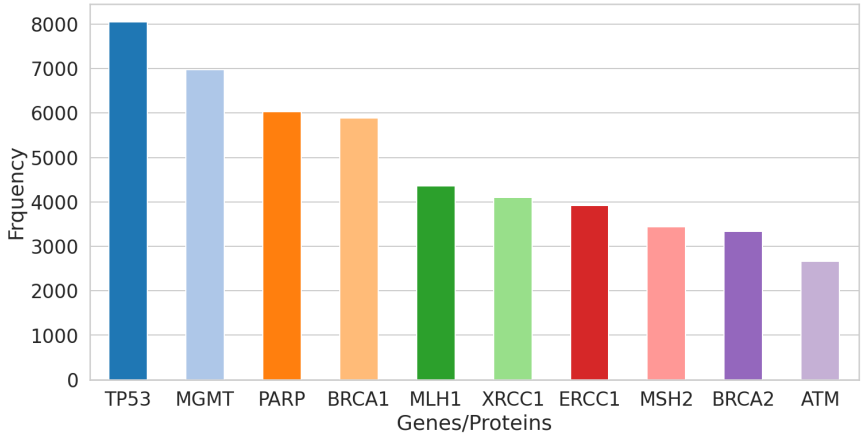


Figure 5.6: Top 10 genes/proteins mentioned in the articles

As discussed in 3.1.3 we do not have gold standard labels to compute recall of the predicted results. However, we modified the precision metric by matching our results with external PPI database STRING [58] and label them as positive. STRING reports 6069 interactions for the provided 295 genes/proteins. We queried our co-occurrence model for all 295 genes/proteins and extracted top 50 associated protein pairs based on PMI value for each query term. Protein pairs that do not share any biological functionality (biological process, cellular component, molecular functions) based on GO terms are dropped. Table 5.3 shows the evaluation results of our model.

min. threshold	**Co-occurrence model**	**PPI STRING evidence**
5	$\frac{\text{Positive}}{\text{Total}}$	$\frac{508}{781} = 0.65$
	$\frac{\text{Negative}}{\text{Total}}$	$\frac{273}{781} = 0.35$
10	$\frac{\text{Positive}}{\text{Total}}$	$\frac{276}{400} = 0.69$
	$\frac{\text{Negative}}{\text{Total}}$	$\frac{276}{400} = 0.69$

Table 5.3: Evaluation of the co-occurrence statistical model

[1]Foreign substance which induces an immune response in the body.

Figure 5.7 represents the distribution of shared GO-terms amongst the potential interacting pair of genes predicted by our co-occurrence model. We can see that most of the genes in each pair shares around 50 or more GO-terms which implies that these genes are involved in the same biological processes, cellular components, or molecular functions, and hence a strong association between them can be inferred.

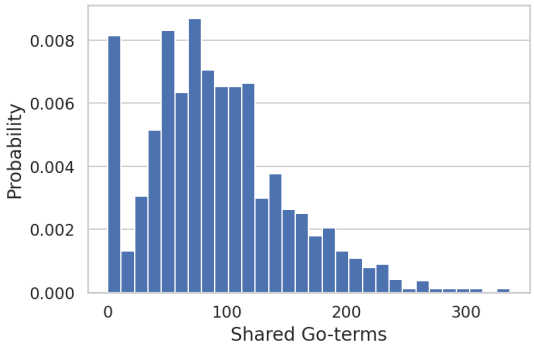


Figure 5.7: Distribution of the shared GO terms in predicted pairs of associated genes

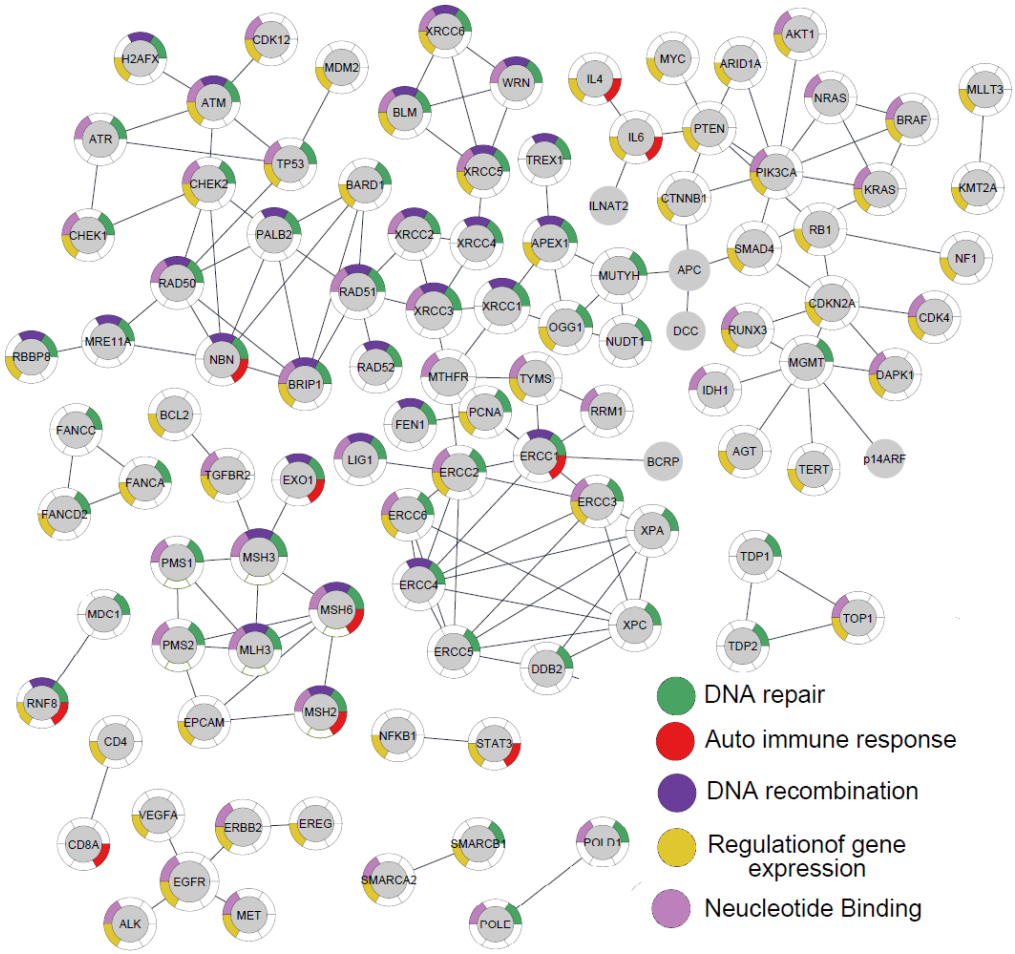


Figure 5.8: An example network graph of PPIs and GIs predicted by model with minimum co-occurrence threshold of 10 and PMI measure greater than 5

Figure 5.8 represents network graph of PPIs/GIs with minimum co-occurrence threshold of 10 and PMI measure greater than 5. The purpose of sharing this network is to provide an illustration of GO terms annotation of genes. Different colors around the circles are associated with multiple biological processes in which these genes are involved. For instance, most of the genes in the network are associated with the DNA repair process. The red color represents the proteins that are involved in immune defense response. We have found 32 biomarkers in this dataset that are involved in both DNA damage response and immune defence.

To verify some other associations between genes other than PPI we have ranked all the extracted associated gene pairs based on their PMI measure. Figure 5.9 represents the network graph of PPI interactions with a minimum co-occurrence threshold of 5. We filtered the top 50 links in the network based on PMI measure, and these links are colored red in the network. Then we randomly chose two small sub-networks circled in the figure 5.9 to evaluate manually. STRING reports an interaction between *TREX1* and *cGAS* but there is no interaction reported between *IFN1* and other terms. Similarly, the STRING reports no interaction for the other three genes circled in the network. Thus, we manually checked for an association between these terms using the internet and found a few articles reporting an association between these genes. Table 5.4 provides the details of the association between these protein-coding genes and reference of the research papers.

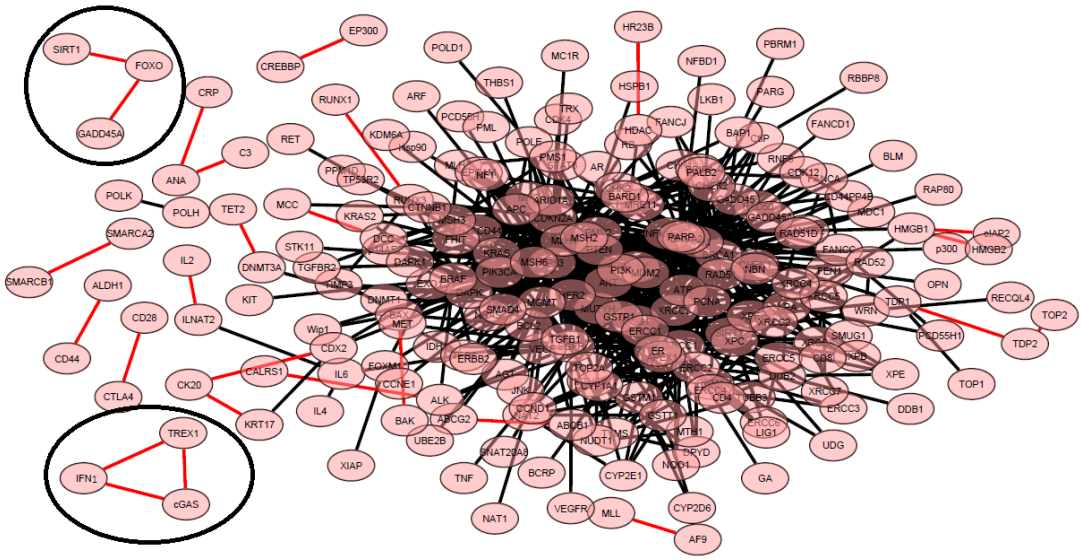


Figure 5.9: Network graph of all genes in the articles with minimum co-occurrence threshold of 5

Year	Refrence	Association	Description
2012	Thakur et al. [59]	$FOXO \rightarrow GADD45A$ $FOXO \rightarrow SIRT1$	*SIRT1 plays a role in the regulation of GADD45A expression by FOXO.*
2015	Ma et al. [31]	$cGAS \rightarrow IFN1$	*Induction of cGAS by IFN1 plays a role in +ve regulation of IFN1*

Table 5.4: References to the association of genes discovered by co-occurrence statistics model

5.3 Association Rule Mining Results for Disease-Gene Association

In this section, we will show the association rule mining results to discover disease candidate genes discussed in the research papers. We have run the apriori algorithm for only two iterations to mine frequent patterns of length two from the set of documents as discussed in 4.4. The normalized $minSup$ threshold of 0.00041 is used to mine frequent patterns from 24170 sets of disease-genes transactions and discovered 3656 frequent patterns. This means that only those entity pairs are considered frequent that occur in at least 10 sentences together. Table 5.5 represents the number of pruned association rules based on different settings of $minConf$. As the database is not very large and transactions comprise only disease and gene names appearing in the sentences of abstracts that is why we have chosen a small value of $minConf$ threshold (0.1) to initially filter the association rules. Moreover, to control the false discovery rate and further pruning of association rules, we are using hypothesis testing with χ^2-test statistic.

$minConf$	0.1	0.2	0.3	0.4	0.5
Association rules	2090	1292	909	684	512

Table 5.5: Total number of association rules with different configurations of $minConf$

In figure 5.10, the left image represents the scatter plot of support versus confidence of association rules, and the right image represents the scatter plot of support versus lift. The color map in both images represents the *p-value* of the association rules. We can notice that only a few association rules with higher support have higher confidence. The aforementioned makes sense because entities with high support values are likely to appear in many documents, reducing the rule's chance of having higher confidence. It is also evident that association rules with low support and confidence appear to have high *p-value*, thus are considered false positives.

On the other hand, association rules with low support have a high lift. This makes sense as lift is the fraction of support of association rules with supports of both antecedent and consequent of the rule. Thus lift measure assigns a high value to rare events. Thus we have ranked association rules based on the *p-values* at a significance level of 0.1. It means that there is only 10% chance of accepting a rule if there is no association.

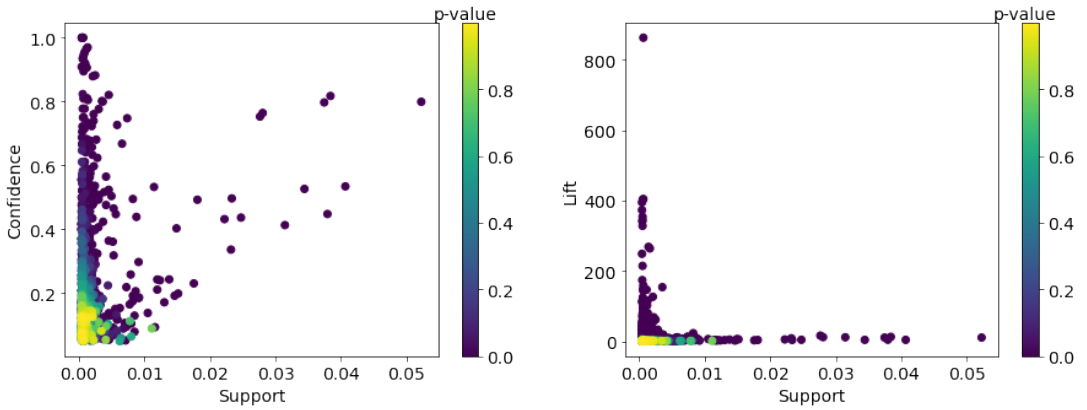


Figure 5.10: Scatter plot of support vs confidence & lift for 2090 association rules

Figures 5.11 and 5.12 represents the top 8 most frequently discussed cancers and autoimmune diseases in the abstracts respectively. From these figures, it is evident that different type of cancers are more frequently discussed than autoimmune diseases in this dataset.

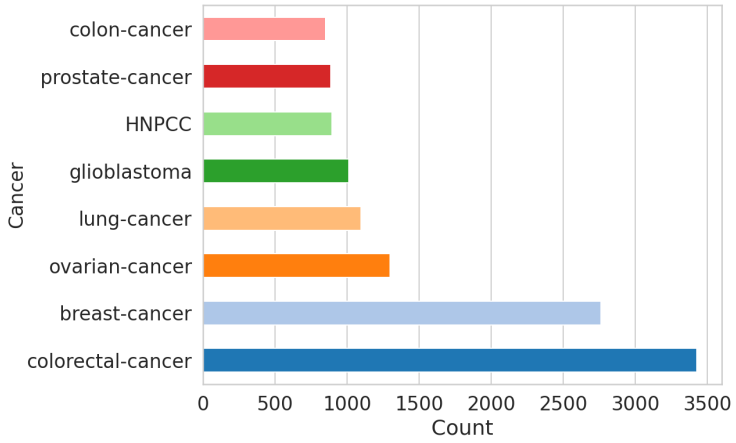


Figure 5.11: Most frequently discussed cancers in the dataset

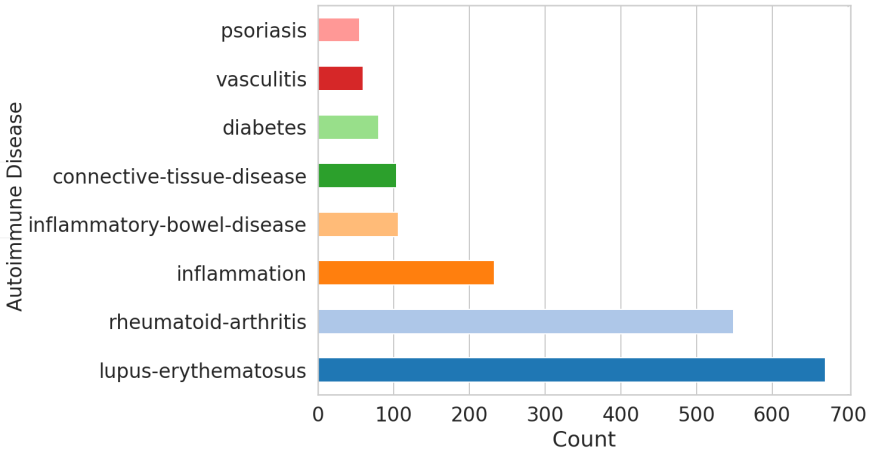


Figure 5.12: Most frequently discussed autoimmune diseases in the dataset

Figure 5.13 represents a visualization of autoimmune diseases found in the discovered association rules that are also discussed with genes at a sentence level. In the figure the x-axis represents the genes, the y-axis represents the autoimmune diseases, and each cell in the matrix represents the *p-value* of the association between disease and genes. All of the *p-values* are less than or equal to 0.05. Thus there is only 5% chance of false discovery. Some of the inferences that can be made based on these results are:

- *ATM*, *MLH1*, *MSH2*, *MSH6* DNA repair markers are potentially associated with autoimmune diseases.
- Gene *TREX1* is associated with multiple autoimmune diseases.

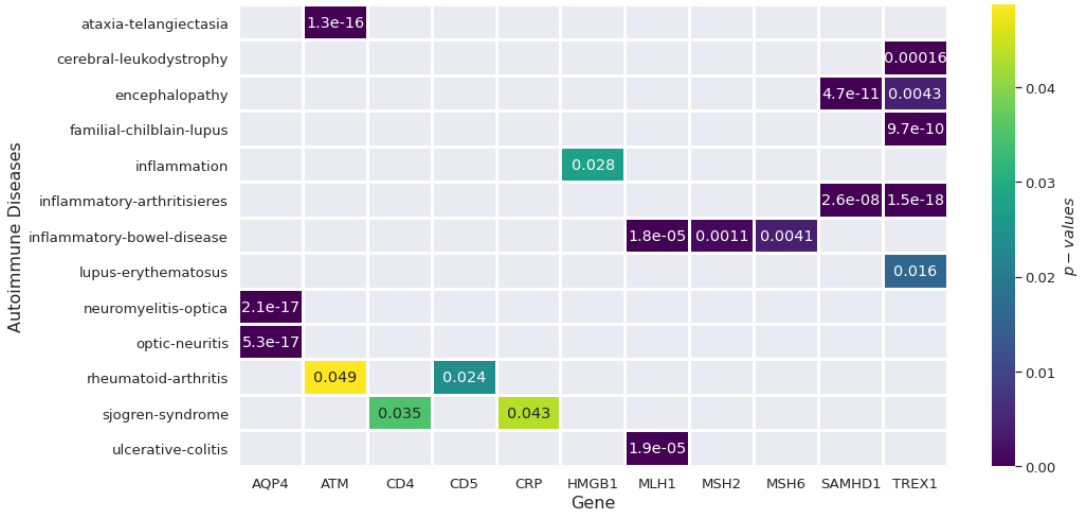


Figure 5.13: Autoimmune disease-gene association analysis at a significance level of 0.05

Figure 5.14 represents a visualization of the most frequent cancer types discovered in the association rules. In the figure, the x-axis represents the genes, the y-axis represents the cancer types, and each cell in the matrix represents the *p-value* of the association between disease and genes. All of the *p-values* are less than or equal to 0.1. Thus there is only a 10% chance of false discovery. Some of the inferences that we can make based on these results are:

- Colon and colorectal cancers share the same DNA repair markers mostly
- Breast cancer and ovarian cancer share the same DNA repair markers

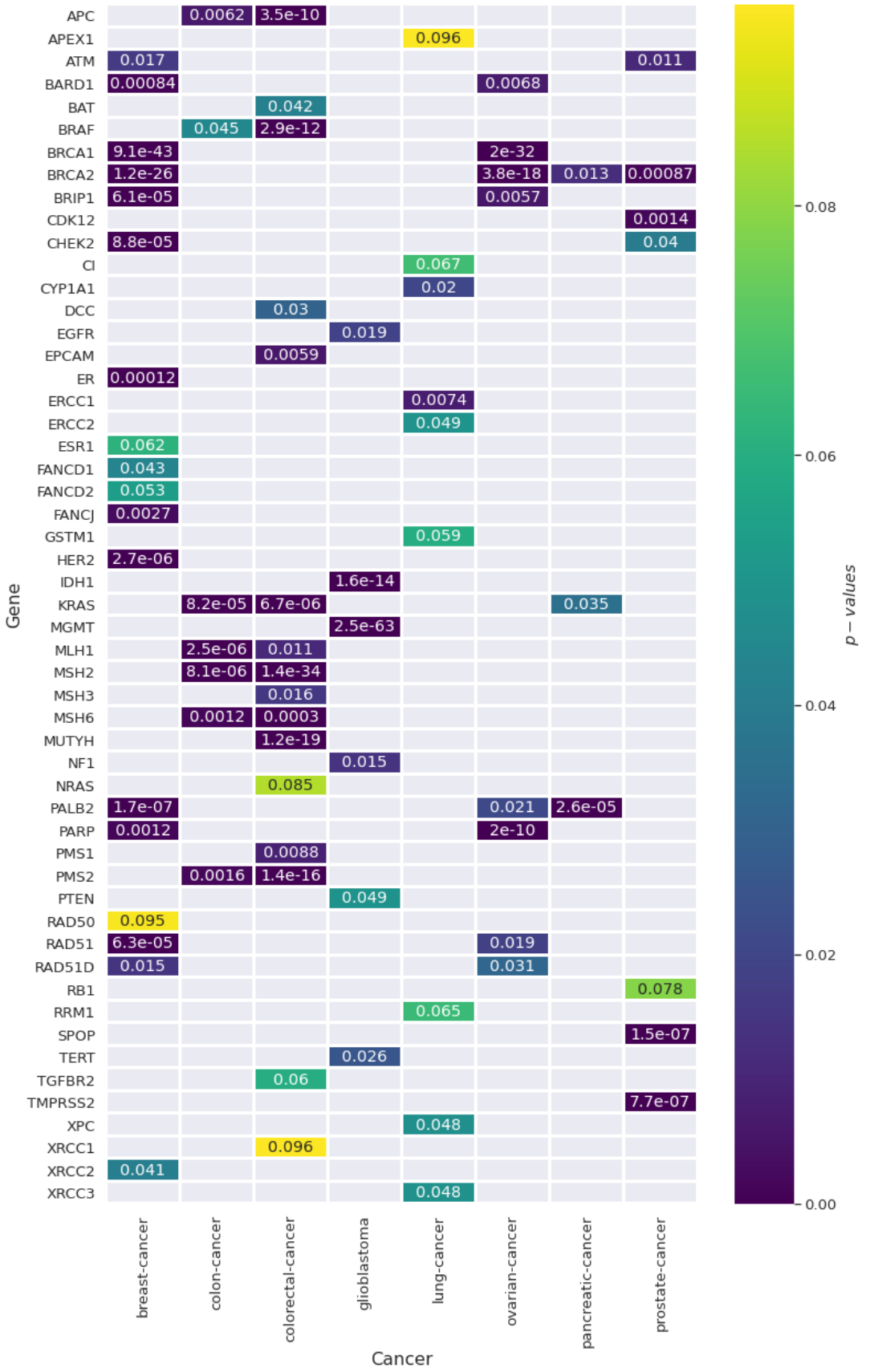


Figure 5.14: Cancer-gene association analysis at a significance level of 0.1

6 Discussion

In this section, we will draw a conclusive analysis of methods and results used to achieve the goals of this thesis project. The limitations of the methods used to achieve the aims of the thesis will be examined. Finally, we will discuss the scope of this work and how it can be helpful for domain experts in biomedicine and clinical researchers.

6.1 Results

Examination of Literature and Trend Analysis with Topic Modeling

Biomedical literature contains complex biological terminologies, and top-weighted keywords of LDA topics involved many genes and protein names that might be associated with multiple biological processes or molecular functions. Thus, to verify and understand the semantic relations between the keywords of each latent topic, we required a domain expert help or external biological databases. To cope up with this issue, we have used GO terms annotation of genes and proteins appearing in the top mentioned keywords of each topic. For instance, MGMT, methylation, temozolomide, glioblastoma, MLH1, and colorectal are present in the top 15 terms of topic-1. By using these key terms, we found that MGMT methylation is a biomarker of response to temozolomide[1] in glioblastoma. Similarly, MGMT is a biomarker of colorectal cancer, and MLH1 is a DNA repair marker associated with colorectal cancer.

In topic-9, genetic markers (ATM, ATR, AID, DNAPK), lymphocytes, immune, and immunoglobulin are present in the top 12 keywords. Based on the GO terms annotation of genetic components, we found that these proteins are DNA repair markers that are critical factors in the production of immunoglobulins molecules, and variable diversity joining (VDJ) and class switch repair (CSR) are somatic recombination mechanisms in developing lymphocytes. On the whole, genetic markers of immunoglobulin (DNA repair markers) and somatic recombination are parts of one big process called the adaptive immune system. We have gathered such information for all latent topics of LDA, and conclude that most of these topics represent a strong semantic coherence.

[1] A drug used to treat brain tumors such as glioblastoma

In figure 5.3 we can see that most of the documents shared topic-2 as a dominant topic that is associated with antigens which are molecular structures, and their presence in the body activates an immune response. According to MK results 5.2 topic-2 exhibits a strong increasing trend. Similarly, topic-9 (adaptive immune response) has the highest value of standardized MK test statistics and hence proved to be one of the most trending topics. Thus based on the trend analysis results, we can infer that the blend of immune defense and DNA repair process is the most trending research topic in this domain.

PPIs, GIs & Disease-gene Association Results

As discussed in 4.3 we do not have any gold standard references to calculate the recall of the results that is why we evaluated our PPI and GI retrieval model results with external biological database of PPIs as shown in table 5.3. The positive rates based on matching the results with the external database imply that genes/proteins that frequently co-occur in literature are not necessarily interacting; they might have some other kind of associations. For instance, genes/proteins belonging to the same families or shared pathways are mentioned together in the abstracts, but an interaction between them is not guaranteed. Moreover, we have not differentiated between genes and gene products (proteins) as we are interested in acquiring both PPIs and GIs from the text.

The purpose of the association rule mining task was to discover potential genes and biomarkers that are associated with cancer and autoimmune disease. This dataset is more concentrated with tumor suppressor genes (DNA repair markers), so it is more likely that genes and proteins discussed in this dataset are related to DNA damage response. To the best of our knowledge, we have not found any external database that emphasizes DNA repair markers associated with a disease. Therefore, we are unable to find standard gold references to match our results; that is why we have used *p-values* of association rules as confidence in the rule. The smaller the *p-value*, the stronger the association between terms.

The association rule mining results for discovering disease-gene association yield that DNA repair markers associated with cancer is a well-researched topic. In the figures 5.11 and 5.12, we can see that the count of types of cancers discussed in this dataset is relatively larger than autoimmune diseases. Thus it is sporadic to find DNA repair markers associated with autoimmune diseases as this is a topic of interest for many biologists and clinical researchers these days. Therefore we have discovered only a few DNA repair markers that are possibly associated with autoimmune diseases, such as *MLH1*, *MSH2*, *MSH6* and *ATM*.

6.2 Methods

LDA Topic Modeling and Trend Analysis

The first objective of this project is to examine the biomedical literature and analyze the research trends in DNA repair pathways data. We have used statistical machine learning method LDA topic modeling to achieve this goal. For effective preprocessing of data, we refer to Schofield et al. paper [49] that quantifies the performance of LDA topic modeling with different variations of prepossessing of data. They conclude that removing stopwords has a meager impact on the quality of topic models, and stemming the words can worsen the quality of topic models. This information helped us save time from time-consuming tasks of identifying domain-specific stopwords, stemming and lemmatization.

We have used the unigram language model (BoW representation) to feed data to the LDA model, where the probability of a sequence of words is broken into the frequency of a single word in the corpus. Usually, authors mention proteins/genes or drugs acronyms in the text

that comprise a single word; that is why the unigram language model was an appropriate choice for our work. Thus with the help of GO terms annotations and other top-weighted keywords in topics, we have assigned labels to each topic. However, the LDA model is unable to capture the sentence structure as the probability of occurrence of each term is independent of other terms. One of the limitations of LDA topic models is that the number of topics (K) should be chosen before the modeling process. Indeed it is hard to decide how many topics are enough to represent a dataset as evaluation of the LDA topic models is not a trivial task. A large value of K can lead to repeating terms in several topics, and a small value may be insufficient to capture the hidden structures in data. To choose an optimal value of K for this data, we have used the topic coherence measure (Cv) as described in 3.3. This measure scores the model topics based on the semantic coherence of the top-weighted terms of each topic.

For research trend analysis, we refer to Sharma et al. [52] work that utilizes multiple topic modeling techniques together with statistical hypothesis test Mann-Kendall to analyze the significant research trends in machine learning. This approach helps us identify and analyze the increasing, decreasing, or constant research trends in the data based on the topic modeling results. Since the research papers used for LDA topic modeling are densely concentrated in one domain of biomedicine, i.e., DNA repair pathways, and we have clustered documents based on the most dominant topic for trend analysis. For instance, a document d is a mixture of all 14 LDA topics, and the approach used by us cluster d based on the most dominant topic to which it belongs, but d might belong to another topic to a degree of 40% or 35%. This can make us indecisive of the trend analysis results.

Hybridization of Information Extraction Methods with Co-occurrence Statistics & Association Rule mining

To achieve the second aim of the thesis, we have first identified the entities genes, proteins, and disease names in the text for all of the biological information extraction tasks. Since NER is one of the most daunting tasks of biological information extraction due to the complex nature of terminologies, and sometimes models can identify drug names as genes. Thus for entity disambiguation, we referred to external biological knowledge base UMLS. The entity linking process helped us make sure that the entities we have extracted from the unstructured text are actual genes and proteins associated with humans. For biological information extraction tasks the principle of approaches employed is that if two terms co-occur in sentences of corpus multiple times, then there might be an association between them.

To retrieve and rank the possibly associated pairs of genes, we have used the information-theoretic measure PMI following the strategy of Bunescu et al. [12]. PMI assigns a high value to the rare events; this helps in capturing rare gene pairs that are possibly involved in an interaction and also helps to control false discovery rate (FDR). Other than PMI ranking, we have predicted an association between genes based on GO annotations of genes following the strategy of Sun et al. [57]. We inferred pair of genes that are not involved in the same biological process, cellular component, or molecular function together does not have any association or interaction. In contrast, those genes which shared many biological functionalities are strong contenders of biological interactions, such as PPI andGI. An advantage of this approach is that once we have the set of documents containing gene/protein names, it does not require a lot of time and high compute power to predict an association between gene pairs.

For disease-gene associations from the unstructured text, we have utilized association rule mining concepts. Here we have considered genes and disease appearing in one sentence as one transaction of items for the rule mining process. Since only a few disease or gene en-

tities co-occur in one sentence resulting in a very sparse dataset, we have chosen the apriori algorithm for frequent pattern mining as apriori performs better on sparse datasets. One of the limitations of using the apriori algorithm is that it repeatedly scans the whole database to generate candidates. This process is time-consuming and requires a huge memory to store frequent patterns of previous iterations. Thus we have run the apriori algorithm for only two iterations to frequent patterns of length two only. These patterns yield a disease-gene pair, a disease-disease pair, or a gene-gene pair that frequently occurs together in the sentences.

The choice of minimum thresholds of support and confidence to discover useful and meaningful association rules is a well-known problem. We have chosen a minimum support threshold of 10, which implies that a disease or gene is frequent if it occurs in at least ten sentences. Another deficiency of the association rule mining is that it generates many redundant and spurious rules even after employing $minSup$ and $minConf$ thresholds filtration. To counter this problem and control FDR we referred to Alvarez et al. [4] paper that proposes a simple way to compute χ^2 test statistics to check dependence between antecedent and consequent of association rules as a function of support, confidence and lift of association rules. Thus using χ^2 test statistics we have computed *p-value* of each association rules and discarded all those rules for which *p-value* is less than 0.1.

A drawback of both biological information extraction approaches used in this project is that they cannot identify the type of relationship between two terms. For example, in the sentence *ATM binds with BRCA1*, the term 'bind' indicates a relation between two proteins *ATM* and *BRCA1*. Nevertheless, our proposed systems cannot identify the term *bind* as we have not utilized verbs in POS while inferring an association between entities. One can achieve better results for PPI, GI and disease-gene associations if only those sentences are filtered which contains verbs such as *interact*, *bind*, *agent*, *candidate* etc together with genes, proteins and diseases as these verbs confirms an association between terms.

6.3 The work in a wider context

The topic modeling with LDA work can be applied to temporal data of any other domain's corpus to analyze the trending research topics. Moreover, bigram or trigram language models can be fed to the LDA model depending on the structure of the text to capture sentence structure in the corpus. This work can also be used to learn hidden semantic themes and summarize large textual data, research articles or news articles. This project's biological information extraction approaches can be replicated on other biomedical corpora to discover relations between biological entities.

7 Conclusion

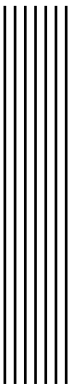
In this section, we will provide a conclusive analysis of the main goals of the project and answers to the research questions. The thesis project aims to achieve two primary goals given a large biomedical text corpus. The first objective is to provide an unsupervised machine learning approach to examine the literature and identify research trends, and the second objective is to extract potential relationships between biological entities reported in the text. Conclusion on the research questions discussed in introduction 1.3 is as follows:

1. Recommend an unsupervised machine learning approach to discover hidden themes and semantically coherent topics from unstructured biomedical text using?
   We have proposed a conventional statistical and machine learning model LDA to discover hidden themes and semantically coherent topics from the corpus. The LDA model seems to perform well on discovering semantically coherent topics but it cannot capture the sentence structure. The method's approach is discussed in section 4.1 and the results of the method are shown in 5.1.

2. How can we analyze and visualize the evolving research trends in unstructured text data over time?
   To analyze the research trends over time in textual data, we have used LDA topic modeling results to cluster documents based on the most dominant topic. After obtaining the topic clusters, we can perform trend analysis as discussed in section 4.1.3. We have employed statistical hypothesis test MK to identify significant increasing or decreasing trends in the research topics. The results of this approach are shown in section 5.1.

3. How can we leverage text mining methods to retrieve relationships between biological entities such as PPIs, GIs and disease-gene associations from biomedical literature?
   We have used hybridization of classical co-occurrence statistical methods with information extractions methods of text mining to discover potential PPIs and GIs studied in the literature. Other than this, our approach utilized GO terms annotations of genes and proteins to predict an association between genetic components studied in this dataset. The approach of the method is discussed in section 4.3. A hybridization of information extraction methods with the popular data mining method, association rule mining, is employed to discover disease-gene associations discussed in the literature. An association between biological entities (proteins, genes, disease) is inferred based on their

co-occurrence frequency on a sentence level in both approaches. We conclude that these approaches are simple and perform well to obtain new potential associations between biological entities reported in the text. However, these approaches are unable to capture the type of relations between entities as we have not used POS tags while inferring an association. We conclude that genes/proteins that frequently co-occur in literature are not necessarily interacting; they might have some other kind of associations. Hence using POS tags that confirm an interaction might increase the true positive rate.



Bibliography

1. Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M. & Homouz, D. Analyzing a co-occurrence gene-interaction network to identify disease-gene association. *BMC bioinformatics* **20,** 1–15 (2019).
2. Agrawal, R., Imieliński, T. & Swami, A. *Mining Association Rules between Sets of Items in Large Databases* in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (eds Buneman, P., Jajodia, S. & Kim, W.) (Association for Computing Machinery, Washington, D.C., USA, 1993), 207–216.
3. Alenlöv, J. *Apriori Algorithm* 2017. `https://www.ida.liu.se/~732A75/material/2021-Lecture6.pdf` (Apr. 19, 2021).
4. Alvarez, S. A. *Technical Report BC-CS-03-01: Chi-squared computation for association rules* tech. rep. 13 (Computer Science Department, Boston College, 2003).
5. Alves, R., Rodriguez-Baena, D. S. & Aguilar-Ruiz, J. S. Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics* **11,** 210–224 (2010).
6. Barnickel, T., Weston, J., Collobert, R., Mewes, H.-W. & Stümpflen, V. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PloS one* **4,** e6393 (2009).
7. Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit* (O'Reilly Media, Inc., 2009).
8. Blei, D. M. & Lafferty, J. D. Topic models. *Text mining: classification, clustering, and applications* **10,** 34 (2009).
9. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research* **3,** 993–1022 (2003).
10. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* **32,** D267–D270 (2004).
11. Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. *Dynamic Itemset Counting and Implication Rules for Market Basket Data* in *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data* (eds Peckman, J. M., Ram, S. & Franklin, M.) (Tucson, Arizona, USA, May 1997), 255–264.

12. Bunescu, R., Mooney, R., Ramani, A. & Marcotte, E. *Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline* in *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis* (eds Verspoor, K., Cohen, K. B., Goertzel, B. & Mani, I.) (New York, NY, June 2006), 49–56.

13. Cellier, P., Charnois, T., Plantevit, M., Rigotti, C., Cremilleux, B., Gandrillon, O., Klema, J. & Manguin, J.-L. Sequential pattern mining for discovering gene interactions and their contextual information from biomedical texts. *Journal of biomedical semantics* **6,** 1–12 (2015).

14. Church, K. & Hanks, P. Word association norms, mutual information, and lexicography. *Computational linguistics* **16,** 22–29 (1990).

15. Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

16. Dierk, S. The SMART retrieval system: Experiments in automatic document processing—Gerard Salton, Ed.(Englewood Cliffs, NJ: Prentice-Hall, 1971, 556 pp., 15.00). *IEEE Transactions on Professional Communication,* 17–17 (1972).

17. Doncheva, N. T., Morris, J. H., Gorodkin, J. & Jensen, L. J. Cytoscape StringApp: network analysis and visualization of proteomics data. *Journal of proteome research* **18,** 623–632 (2018).

18. Evangelopoulos, N., Zhang, X. & Prybutok, V. R. Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems* **21,** 70–86 (2012).

19. Firth, J. R. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957).

20. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence,* 721–741 (1984).

21. Gong, L. Application of biomedical text mining. *Artificial Intelligence: Emerging Trends and Applications,* 417 (2018).

22. Han, J., Pei, J. & Yin, Y. Mining frequent patterns without candidate generation. *ACM sigmod record* **29,** 1–12 (2000).

23. He, M., Wang, Y. & Li, W. PPI finder: a mining tool for human protein-protein interactions. *PloS one* **4,** e4554 (2009).

24. Hofmann, T. *Probabilistic latent semantic indexing* in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (eds Laskey, K. B. & Prade, H.) (1999), 50–57.

25. Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. Using literature-based discovery to identify disease candidate genes. *International journal of medical informatics* **74,** 289–298 (2005).

26. Jurafsky, D. & Martin, J. H. in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition 3rd edition* chap. Information Extraction (Prentice Hall, Inc., New Jersey, USA, 2020).

27. Kendall, M. G. *Rank Correlation Methods (4th Edition)* (Charles Griffin, 1975).

28. Kuhlmann, M. *Text clustering and topic modelling* `https://www.ida.liu.se/~732A92/commons/TM-2020-3.pdf` (May 1, 2021).

29. Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **5,** 1–22 (2016).

30. Lungu, I., Pirjan, A., *et al.* Research issues concerning algorithms used for optimizing the data mining process. *J. Inf. Syst. Oper. Manage.* **4,** 108–125 (2010).

31. Ma, F., Li, B., Liu, S.-y., Iyer, S. S., Yu, Y., Wu, A. & Cheng, G. Positive feedback regulation of type I IFN production by the IFN-inducible DNA sensor cGAS. *The Journal of Immunology* **194,** 1545–1554 (2015).

32. Mann, H. B. Nonparametric tests against trend. *Econometrica: Journal of the econometric society,* 245–259 (1945).

33. Manning, C. & Schutze, H. *Foundations of statistical natural language processing* (MIT press, 1999).

34. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research* **47,** D419–D426 (2019).

35. Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. *Optimizing semantic coherence in topic models* in *Proceedings of the 2011 conference on empirical methods in natural language processing* (2011), 262–272.

36. Muzaffar, A. W., Azam, F. & Qamar, U. A relation extraction framework for biomedical text using hybrid feature set. *Computational and mathematical methods in medicine* **2015** (2015).

37. Nasar, Z., Jaffry, S. W. & Malik, M. K. Information extraction from scientific articles: a survey. *Scientometrics* **117,** 1931–1990 (2018).

38. Nastasi, C., Mannarino, L. & D'Incalci, M. DNA Damage Response and Immune Defense. *International Journal of Molecular Sciences* **21,** 7504 (2020).

39. Neumann, M., King, D., Beltagy, I. & Ammar, W. *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing* in *Proceedings of the 18th BioNLP Workshop and Shared Task* (Association for Computational Linguistics, Florence, Italy, 2019), 319–327. https://www.aclweb.org/anthology/W19-5034 (Feb. 28, 2021).

40. Ohta, T., Pyysalo, S., Kim, J.-D. & Tsujii, J. A re-evaluation of biomedical named entity–term relations. *Journal of bioinformatics and computational biology* **8,** 917–928 (2010).

41. Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T. & Iliopoulos, I. Protein–protein interaction predictions using text mining methods. *Methods* **74,** 47–53 (2015).

42. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50,** 157–175 (1900).

43. *PubMed* https://pubmed.ncbi.nlm.nih.gov/ (Feb. 25, 2021).

44. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *The Journal of Open Source Software* **3.** http://joss.theoj.org/papers/10.21105/joss.00638 (Apr. 2018).

45. Ray, S. & Craven, M. *Representing sentence structure in hidden Markov models for information extraction* in *International Joint Conference on Artificial Intelligence* **17** (Citeseer, 2001), 1273–1279.

46. Rehurek, R. & Sojka, P. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* **3** (2011).

47. Röder, M., Both, A. & Hinneburg, A. *Exploring the Space of Topic Coherence Measures* in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (eds Gabrilovich, E. & Tang, J.) (Association for Computing Machinery, New York, NY, USA, 2015), 399–408.

48. Salton, G. & Yang, C.-S. On the specification of term values in automatic indexing. *Journal of documentation* (1973).

49. Schofield, A., Magnusson, M., Thompson, L. & Mimno, D. Understanding Text Pre-Processing for Latent Dirichlet Allocation.

50. Schonlau, M. & Guenther, N. Text mining using n-grams. *Schonlau, M., Guenther, N. Sucholutsky, I. Text mining using n-gram variables. The Stata Journal* **17,** 866–881 (2017).

51. Schütze, H., Manning, C. D. & Raghavan, P. *Introduction to information retrieval* (Cambridge University Press Cambridge, 2008).

52. Sharma, D., Kumar, B. & Chand, S. A Trend Analysis of Machine Learning Research with Topic Models and Mann-Kendall Test. *International Journal of Intelligent Systems and Applications* **11,** 70–82 (2019).

53. Shatnawi, M. Review of recent protein-protein interaction techniques. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology* **12,** 99–121 (2015).

54. Singh, M. *Word embedding* `https://medium.com/data-science-group-iitr/word-embedding-2d05d270b285`. (Mar. 05, 2021).

55. *Spacy Language Processing Pipeline* `https://spacy.io/usage/processing-pipelines` (Feb. 28, 2021).

56. Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics* **47,** 8–13 (2014).

57. Sun, K., Gonçalves, J. P., Larminie, C. & Pržulj, N. Predicting disease associations via biological network analysis. *BMC bioinformatics* **15,** 1–13 (2014).

58. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47,** D607–D613. `http://string.embl.de/` (Mar. 15, 2021) (2019).

59. Thakur, B. K., Lippka, Y., Dittrich, T., Chandra, P., Skokowa, J. & Welte, K. NAMPT pathway is involved in the FOXO3a-mediated regulation of GADD45A expression. *Biochemical and biophysical research communications* **420,** 714–720 (2012).

60. Yoav, G. & Graeme, H. Neural Network Methods in Natural Language Processing. *Morgan & Claypool: San Rafael, SR, USA* (2017).

61. Zhao, Q. & Bhowmick, S. S. Association rule mining: A survey. *Nanyang Technological University, Singapore,* 135 (2003).