**Interim Progress Report**

**on**

**Systematic Communication Modelling with Latent Semantic Analysis**

Contents

[Chapter 1 Introduction 3](#_Toc79484840)

[Chapter 2 Literature Review 4](#_Toc79484841)

[Chapter 3 Methodology 5](#_Toc79484842)

[Chapter 4 Implementation 7](#_Toc79484843)

[Chapter 5 Ethical, Legal and Social issues 28](#_Toc79484844)

[Bibliography 30](#_Toc79484845)

List of Figures

[Figure 1 sample data from covid-psychology.txt 8](#_Toc79475143)

[Figure 2 Importing libraries 8](#_Toc79475144)

[Figure 3 code for loading data at article level from experiment 4 – “covid-science-large.txt” dataset. 9](#_Toc79475145)

[Figure 4 code for loading data at 2-3 paragraph-level from experiment 4 – “covid-science-large.txt” and “covid-psychology-large.txt” dataset. 9](#_Toc79475146)

[Figure 5 code for Pre-processing. 10](#_Toc79475147)

[Figure 6 Code for modelling LSA and generating top 5 words at article level from “covid-science-large.txt” dataset. 11](#_Toc79475148)

[Figure 7 Code for modelling LSA and generating top 5 words at article level from “covid-psychology-large.txt” dataset. 11](#_Toc79475149)

[Figure 8 Code for modelling LSA and generating top 5 words at 2-3 paragraph-level from “covid-science-large.txt” dataset. 12](#_Toc79475150)

[Figure 9 Code for modelling LSA and generating top 5 words at 2-3 paragraph-level from “covid-psychology-large.txt” 12](#_Toc79475151)

[Figure 10 Code for modelling LSA and generating top 5 words at paragraph-level(all paragraphs) from “covid-science-large.txt” dataset. 12](#_Toc79475152)

[Figure 11 Code for modelling LSA and generating top 5 words at paragraph-level(all paragraphs) from “covid-psychology-large.txt” dataset. 12](#_Toc79475153)

List of Tables

[Table 1 List of words from LSA model(LSA results) 13](#_Toc79475185)

[Table 2 Wordcloud results 17](#_Toc79475186)

[Table 3 Cosine similarity results 22](#_Toc79475187)

[Table 4 WordNet results 24](#_Toc79475188)

# Chapter 1 Introduction

With people spending increasing time online for studying and researching looking for resources like scientific research papers, newspapers, psychology research papers etc. It is hard and time consuming to find resources that we want from the ocean of content available on the internet. But, If different documents available on the internet are tagged according to the content available within the document, It will at least decrease the complexity of searching for a specific type of document to some extent. To achieve this first an application need to be developed to analyse the content available in the document semantically.

LSA is a natural language processing technique based on linear algebra that tries to capture and code the semantics of words and documents at different levels of document like paragraph level, document level, etc. (Pilato and Vassallo, 2015). There are a lot of algorithms to extract lexical information from text corpora. In this area, the focus is to identify what can be extracted from the distribution of words, excluding sublexical information, perceptual grounding and intimate constraint(Dennis et al., 2003). LSA is a well-known algorithm for computational modelling of the construction of semantic representations of the corpus. LSA is also widely used in information retrieval systems(Wenli, 2016). From(Mohamed and Watada, 2010), an LSA approach to document classification for knowledge application acquisition has been proposed, but the classification of documents is done by generating a set of words with high weights from the document and matching them to a predefined dictionary of categorised words obtained from training the documents. The scope of using a predefined dictionary to the model might not be the ultimate or the best solution to categorise the documents effectively.

The project is to develop an application to model systematic communication by extraction and representation of the contextual usage of words through the statistical computation applied to large corpora of text for both semantic and syntactic analysis. For my project, I will find the statistical difference between law and science, science, non-science and psychology papers from a corpus of texts, by extracting the insights from the data using Latent Semantic Analysis (LSA) and then by using statistical tests to compare the data extracted from each model and then find the insights that could differentiate those documents with statistical measure.

To avoid this limitation of using pre-defined words from the training data to classify the corpus of text, a statistical method needs to be developed to derive the semantic distance between the weighted words. Knowing that there are many algorithms to analyse the document semantically, knowing the best algorithm to produce weighted words to analyse the document is a problem. There are many statistical methods to test the results, like the chi-square test, t-test, Mantel test. Choosing one that suites to the project results is a problem. So, the research question raises here.

What machine learning algorithm is effective to analyse the document corpus?

What test best fits to test the results statistically?

# Chapter 2 Literature Review

Many of the first and most ambitious systemic approaches were proposed by (Von Bertalanffy, 1967), He was interested in mathematical laws that characterized systems in different fields. who as early 1920s tried to develop a general theory of systems to identify and formalize abstract principles that apply to most systems, be they sociological, physical, psychological and biological. And proved that a small number of systems of differential equations can explain natural and social phenomena in different fields.

Later (Ashby, 1961) carried out the further mathematical formulation of system behaviour. Explained the importance of feedback control, input, output, prediction and stability while working on aircraft fire control. He generalizes these concepts to biology and social sciences. And in (Brooks-Gordon and Freeman, 2006) mentioned different computational models of the construction of syntactic structures include the Pooled Adjacent Context Model (PAC), the Syntagmatic Paradigmatic model (SP) and computational models of the construction of semantic representations including, LSA, Hyperspace Analogue to Language (HAL), Sparse Random Context Representation (SRCR) and Word Association Space(WAS). These models are used to elucidate the similarity or difference, of word and passage meaning through the statistical analysis of large text corpora.

##### 2.1 LSA for Semantic analysis

In Latent semantic analysis, the term latent means hidden in the data which cannot be directly measured. These features are essential to the data but are not essential features of the data set. Latent Semantic analysis is a natural processing technique as well as an unsupervised learning technique. The latent semantic analysis aims to create representations of the text data in terms of topics or latent features. Which can produce dimensionality of the original text-based dataset(Landauer, 2014).

Traditional information retrieval system models mainly consist of probability models or space models or vector models. Its basic principle being the keyword retrieval system. It is like matching the words that need to be searched with the words from the index of the document. But the existence of synonyms and polysemy made these systems complex to retrieve the information(Wenli, 2016). Later (Dumais et al., 1988) proposed latent semantic analysis to improve access to textual information.

Latent Semantic Analysis(Landauer et al., 1998), uses free text as an input, with the paragraph as a relevant contextual unit. An analysis of word frequency within the paragraphs represents the meaning and relationship of words as vectors which helps to compare words semantically even at high order co-occurrences(Kontostathis and Pottenger, 2002). (Foltz, 1996) rigorously tested LSA and it is regarded as a very successful theory and better than Hyperspace Analog to Language(Meena and Raj, 2014) which is a method for lexical analysis and a model of semantic analysis. HAL uses a sliding window of co-occurrences words as a measure of context and avoiding higher-order co-occurrences.

(Anwar, 2006) conducted some experiments with latent semantic analysis for tagging and ambiguity correction for words from the large corpus. And provided details about how LSA does better compared to other statistical word tagging and ambiguity correction methods. And the size and the meaning of the corpus plays an important role while using latent semantic analysis(Dascalu and McNamara, 2017). An application(Rosenbusch et al., 2020) has been produced addressing the arbitrary measurements and disconnection between research groups in psychology by detecting semantic overlap between the scales with semantic analysis. He also explained that These similarity or dissimilarity measurements can be shown by cosine similarity and can be represented a space by word network form.

# Chapter 3 Methodology

##### 3.1 Data collection

This is the first and important step of the experiment. The data collected needs to be research specific data. This is the process of assembling, quantifying information on target variables and analyzing insights(Morgan and Harmon, 2001). For this project, the data collected need to be both quantitative and qualitative data, for statistical insights and context-specific insights.

In quantitative terms, the datasets collected need to be significantly different and of different sizes to compare the results.

Ex: Covid science dataset of multiple sizes and covid non-science dataset of multiple sizes

In qualitative terms, the datasets collected need to be semantically differentiable.

Ex: One data set with research papers and the other with newspapers.

##### 3.2 Data cleaning

The data collected from multiple sources consists of a lot of noise in it like unwanted data, symbols etc. So, To perform experiments on the data collected the data need to be cleaned. Data cleaning is a process of detecting, identifying, and correcting data inconsistency, duplicity, redundancy(Broeck et al., 2005).

There are many methods developed for data cleaning. But before applying those methods, the data should be overseen once to detect, identify the issues in the data and skim any unwanted data. Different types of data have different techniques for data cleaning. Since our data is textual data, These techniques like tokenizing, removing stem words, removing stop words need to be performed. Where tokenizing is a process of converting the words in the large textual data to tokens.

##### 3.3 LSA

Latent Semantic Analysis(LSA) is a natural processing technique as well as an unsupervised learning technique, that tries to capture and code the semantics of words and documents at different levels.

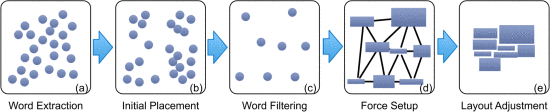
After collecting and cleaning the data the data need to be modelled using LSA. Which generates a weighted matrix by semantically comparing each word to all other words in the corpus.

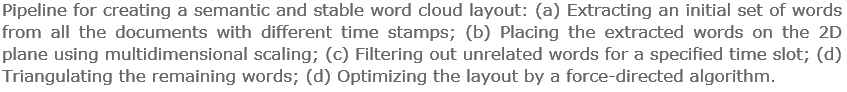
This weighted matrix can be achieved by importing the Latent Semantic Indexing(LSI) model from genism. models library and set the data to the model.

The LSI model first generates the document term matrix and then performs singular value decomposition on the matrix. From this, the weighted matrix of the words is generated at multiple levels of documents.

##### 3.4 WordCloud

WordCloud is an application program interface that is used for an easy and clear representation of the words which are related to each other in a corpus. WordCloud is a process of coupling the trend chart with the words with dynamic word clouds to visualize the material content evolutions in a set of documents(Cui et al., 2010).





##### 3.5 Wordnet

The term wordnet means a word network. Wordnet is a process of illustrating the relation(similarity or dissimilarity) between the words from different documents, graphically in a network. The network consists of nodes and edges. Nodes represent the words and the edges represent the similarity or dis-similarity metrics between those words in the corpus. The similarity or dissimilarity among words can be calculated by cosine similarity metrics. NetworkX is a python package that is used to generate wordnet.

##### 3.6 Statistic test

Many different procedures have been developed to compare the correspondence of one set of distances with another set. Important among them are the (1) matrix correlation techniques, (2) network maching techniques of Spielmen, (3) matrix dilation and rotation techniques, (4) smallest-space techniques of Lingoes and Guttman. These all methods suffer from a difficulty in assessing the statistical significance. Where the problem is that a set of all pairwise distances between k units (In this case semantics) cannot be independent.

The comparison of genetic divergence or genetic distances, estimated by pairwise FST and related statistics, with geographical distances by Mantel test is one of the most popular approaches to evaluate spatial processes driving population structure (Diniz-Filho et al., 2013). With the similar approach we apply mantel test for semantic word population and derive the geographical distance in the space using mantel test.

In order to evaluate the significance form different population. We need to try permutate with different populations(top words) generated from the corpus of texts.

# Chapter 4 Implementation

To understand the statistical compare the results from the algorithm on different datasets, I have conducted 4 different experiments with 4 different sizes of datasets.

Where Experiment 1: LSA at the article level, 2-3 paragraph level(200-words), paragraph level(Whole document) for both “law-text-all.txt” and “non-law-all.txt” datasets.

Experiment 2: LSA at the article level, 2-3 paragraph level(200-words), paragraph level(Whole document) for three datasets, “covid-science-small.txt”, “covid-non-science-small.txt” and “covid-psychology-small.txt” containing 10 papers on each dataset.

Experiment 3: LSA at the article level, 2-3 paragraph level(200-words), paragraph level(Whole document) for three datasets, “covid-science.txt”, “covid-non-science.txt” and “covid-psychology.txt” containing 10 papers on each dataset.

Experiment 3: LSA at the article level, 2-3 paragraph level(200-words), paragraph level(Whole document) for three datasets, “covid-science.txt”, “covid-non-science.txt” and “covid-psychology.txt” containing 20 papers on each dataset.

Experiment 4: LSA at the article level, 2-3 paragraph level(200-words), paragraph level(Whole document) for three datasets, “covid-science-large.txt” and “covid-psychology-large.txt” containing 40 papers on each dataset.

##### 4.1 Data

For this project, 5 different datasets with large corpora of text were collected to perform LSA on them. The first one is law-text-all.txt which consists of 3 articles with legal judgements. In second dataset non-law-all.txt, consists of 4 articles with general conversations. Each article is separated by the word NEW\_ARTICLE.

In third dataset covid-science.txt, consists of covid related scientific research papers collected from multiple sources (ex. IEEE Xplorer, Google Scholar). This dataset is then partitioned into multiple sizes, 10 papers(covid-science-small), 20 papers(covid-science), 40 papers(covid-science-large) in order to test LSA performance at multiple levels. Each article is separated by the word “NEW\_PAPER”.

The fourth dataset is covid-non-science.txt, which consists of covid related newspaper articles collected from different news article providers like the Daily Mail and others. Then similar to the third dataset, the covid-non-science dataset is also partitioned into multiple sizes, 10 papers(covid-non-science-small), 20 papers(covid-non-science). Each article is separated by the word “NEW NEWSPAPER”.

The fifth dataset is covid-psychology.txt, which consists of covid related psychology research papers collected from multiple sources(ex. IEEE Xplorer, Google Scholar). Then similar to the third dataset, the covid-psychology dataset is also partitioned into multiple sizes, 10 papers(covid-psychology-small), 20 papers(covid-psychology), 40 papers(covid-psychology-large). Each article is separated by the word “NEW PAPER”.

After the data is collected from multiple sources, load the dataset. While loading the data set, the data is divided into multiple levels of documents like article level or paragraph level to perform LSA at multiple levels of the document and extract insights from data at multiple levels to understand how LSA performs to the data at multiple levels of the document.

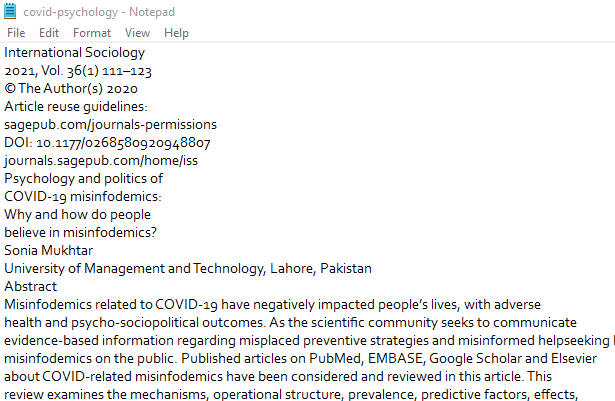


Figure 1 sample data from covid-psychology.txt

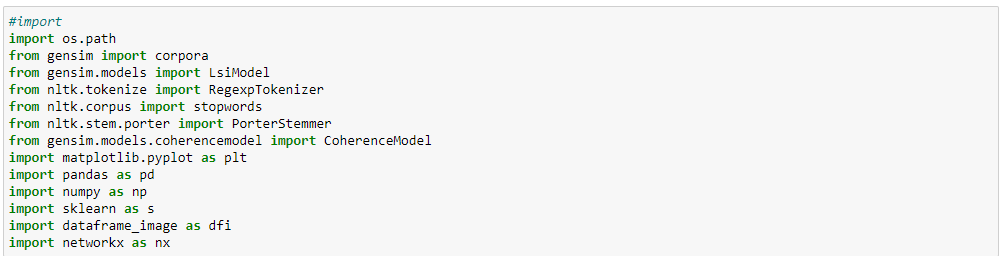


Figure 2 Importing libraries



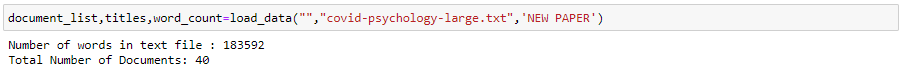


Figure 3 code for loading data at article level from experiment 4 – “covid-science-large.txt” dataset.



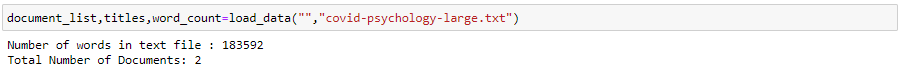


Figure 4 code for loading data at 2-3 paragraph-level from experiment 4 – “covid-science-large.txt” and “covid-psychology-large.txt” dataset.

##### 4.2 Data Cleaning

The data collected from the different sources contain a lot of noise and might lead to deviated results when experimenting with the data, So to remove the noise multiple preprocessing techniques were applied to the data like tokenizing, removing stopwords, and stemming.

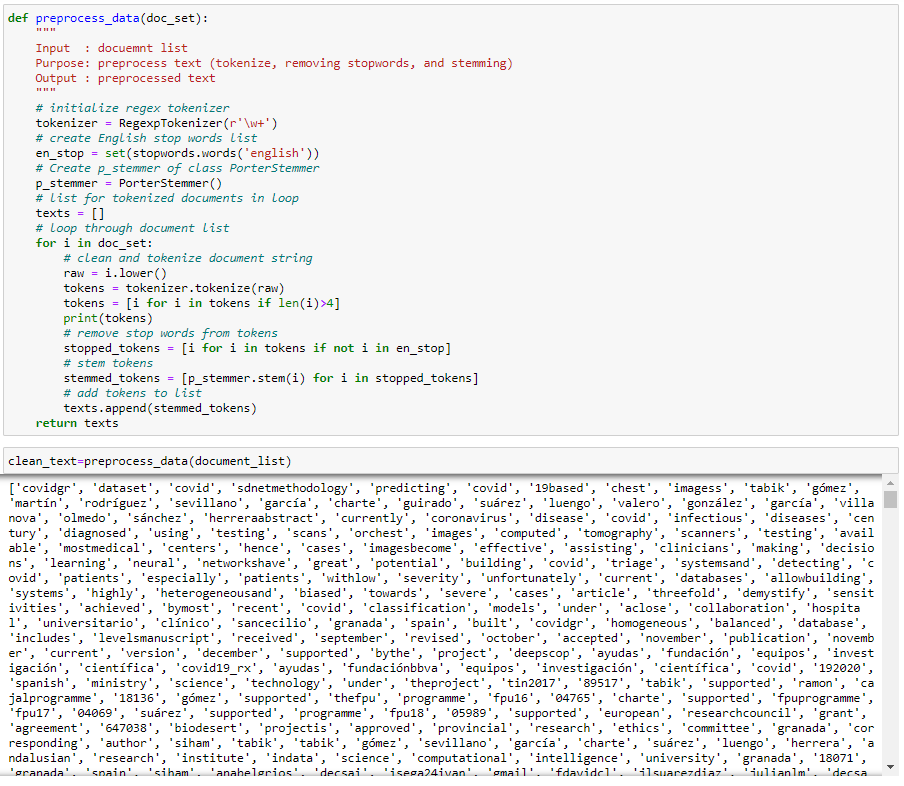


Figure 5 code for Pre-processing.

##### 4.3 LSA

Preparing corpus by creating the term dictionary for corpus and generating document term matrix to perform LSA using Latent Semantic Indexing model(lsimodel) from python “genism.models” library. And then generating the top 5 words from the corpus.

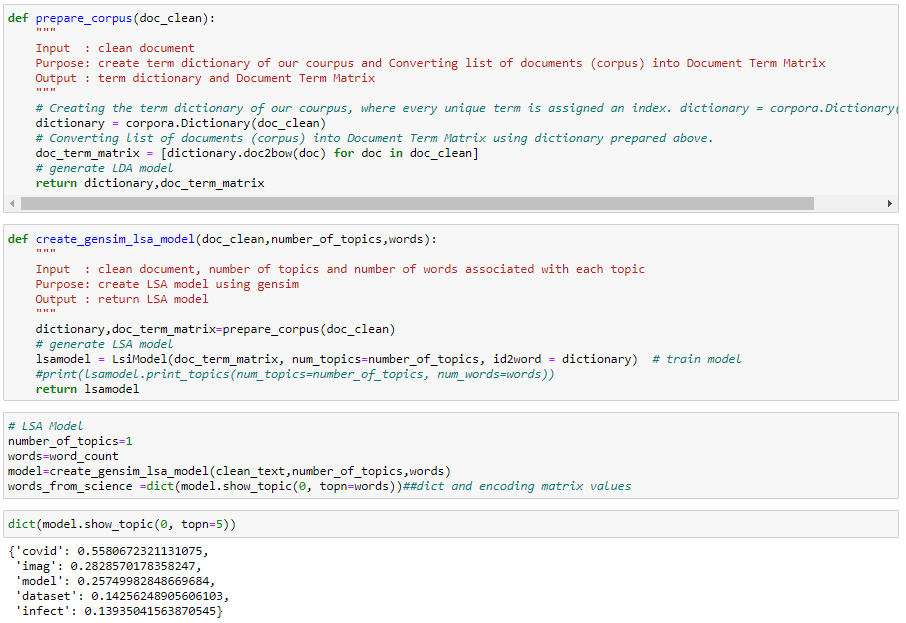


Figure 6 Code for modelling LSA and generating top 5 words at article level from “covid-science-large.txt” dataset.

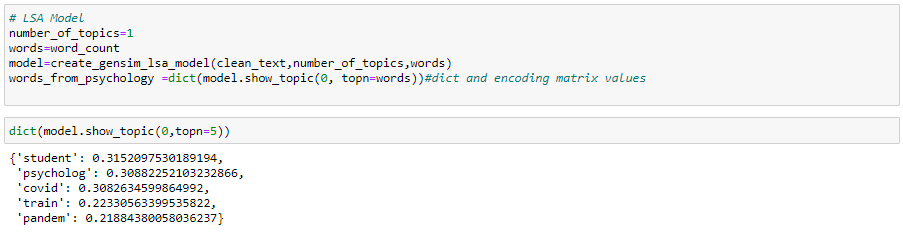


Figure 7 Code for modelling LSA and generating top 5 words at article level from “covid-psychology-large.txt” dataset.

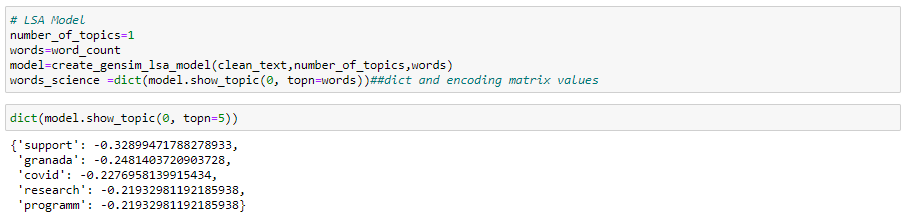


Figure 8 Code for modelling LSA and generating top 5 words at 2-3 paragraph-level from “covid-science-large.txt” dataset.

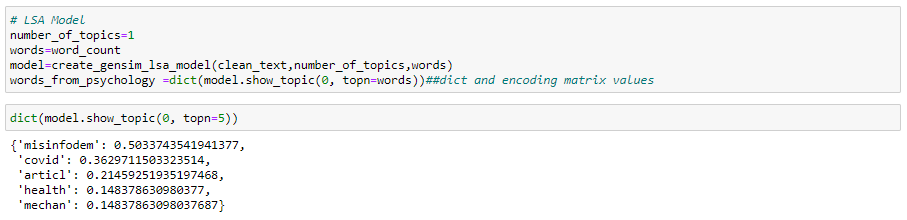


Figure 9 Code for modelling LSA and generating top 5 words at 2-3 paragraph-level from “covid-psychology-large.txt”

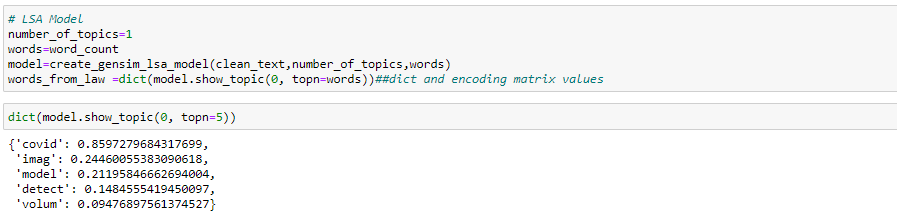


Figure 10 Code for modelling LSA and generating top 5 words at paragraph-level(all paragraphs) from “covid-science-large.txt” dataset.

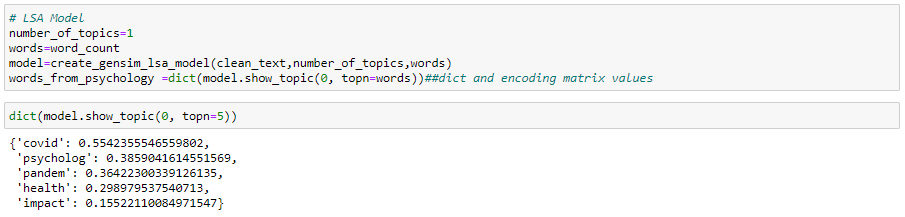


Figure 11 Code for modelling LSA and generating top 5 words at paragraph-level(all paragraphs) from “covid-psychology-large.txt” dataset.

**The below table shows the top 5 words and their weights in the document generated by “lsimodel”.**

Table 1 List of words from LSA model(LSA results)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment 1: Document at various levels for short dataset (law: non-law)** | | | | |
| **No.of Documents** | **law-texts-all** | | **non-law-all** | |
| **Top words** | **weights** | **Top words** | **weights** |
| 1. **Article level (4|4)** | **'tariff'** | 0.3329384971656329, | **'judg'** | 0.3207357836732813 |
| **'secretari'** | 0.3314772021127373 | **'evid'** | 0.3092444069593179 |
| **'prison'** | 0.2491591199567464 | **'west'** | 0.2804028170023648 |
| **'life'** | 0.2361170851923661 | **'trial'** | 0.2772485745197205 |
| **'sentenc'** | 0.2295229637034024 | **'case'** | 0.2281040582305726 |
|  |  |  |  |  |
| 1. **2-3 paragraphs** | **'tariff'** | 0.2852867613391 | **'life'** | 0.424625074653 |
| **'life'** | 0.245096194271 | **'prison'** | 0.366156436155 |
| **'prison'** | 0.2450961942716 | **'sentenc'** | 0.291549439087 |
| **'would'** | 0.2419339607783 | **'hindley'** | 0.257685215145 |
| **'sentenc'** | 0.2041150688308 | **'posit'** | 0.233080800590 |
|  |  |  |  |  |
| 1. **Paragraph level (450|3038)** | **'tariff'** | -0.372423678930 | **'hindley'** | -0.63764826639 |
| **'secretari'** | -0.325658507266 | **'myra'** | -0.58632670557 |
| **'prison'** | -0.262324361300 | **'west'** | -0.23136661906 |
| **'state'** | -0.246189129218 | **'case'** | -0.18241034492 |
| **'life'** | -0.234305014747 | **'trial'** | -0.14821551842 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment 2: Document at various levels for a small dataset (COVID-science: non-science: Covid-psychology)** | | | | | | |
| **No.of Documents** | **covid-science-small** | | **Covid-non-science-small** | | **Covid-psychology-small** | |
| **Top words** | **weights** | **Top words** | **weights** | **Top words** | **weights** |
| **Article level (10|10)** | **'covid'** | 0.46802988644393423, | **'test'** | 0.6105961140643029, | **'covid'** | 0.41992720259455585, |
| **'model'** | 0.3319825262829552, | **'peopl'** | 0.31072871004100155, | **'psycholog'** | 0.28694986504872455, |
| **'imag'** | 0.32671571923604975 | **'covid'** | 0.25119870827829194, | **'pandem'** | 0.24145037935803, |
| **'class'** | 0.20559425407282925, | **'symptom'** | 0.2302237165624363, | **'impact'** | 0.21712822111271726, |
| **'dataset'** | 0.15776417661741118 | **'fever'** | 0.17811639545747093 | **'health'** | 0.19835033050070103 |
|  |  |  |  |  |  |  |
| **2-3 paragraphs** | **'support'** | -0.3289947178827894, | **'market'** | 0.2986573559242644, | **'misinfodem'** | 0.5033743541941375, |
| **'granada'** | -0.24814037209037246, | **'covid'** | 0.21979867961934094, | **'covid'** | 0.36297115033235156, |
| **'covid'** | -0.2276958139915434, | **'friday'** | 0.18596179123202228, | **'articl'** | 0.21459251935197457, |
| **'programm'** | -0.21932981192185963, | **'infect'** | 0.15212490284470373, | **'commun'** | 0.14837863098037715, |
| **'research'** | -0.21932981192185963 | **'copper'** | 0.14653245307956103 | **'health'** | 0.14837863098037676 |
|  |  |  |  |  |  |  |
| **Paragraph level (7401|229)** | **'covid'** | 0.7920442330385795, | **'trial'** | 0.4673333939365011, | **'pandem'** | 0.3492162493102458, |
| **'model'** | 0.2953962153046564, | **'vaccin'** | -0.3789496907239388, | **'polici'** | 0.33206567897289235, |
| **'imag'** | 0.2745133138045724, | **'stage'** | -0.33285475306510043, | **'health'** | 0.29289264374192664, |
| **'detect'** | 0.16365565281326466, | **'expect'** | -0.2332697529258982, | **'covid'** | 0.26562150610270957, |
| **'class'** | 0.12952267447918445 | **'underway'** | -0.20015564175913142 | **'impact'** | 0.2463281943440908 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment 3: Document at various levels for large dataset (20 papers COVID-science: non-science: Covid-psychology)** | | | | | | |
| **No.of Documents** | **covid-science.txt** | | **Covid-non-science.txt** | | **Covid-psychology.txt** | |
| **Top words** | **weights** | **Top words** | **weights** | **Top words** | **weights** |
| **Article level (20|20)** | **'covid'** | 0.5195000201940321, | **'test'** | 0.5525012615032745, | **'covid'** | 0.3733564956224059, |
| **'imag'** | 0.3819419924959906, | **'peopl'** | 0.32711032908806575, | **'psycholog'** | 0.26785991090419736, |
| **'model'** | 0.24414357176752507, | **'covid'** | 0.32350304714824757, | **'social'** | 0.2517797817165504, |
| **'dataset'** | 0.17846130209695354, | **'symptom'** | 0.20147297742697395, | **'epidem'** | 0.19223792506781157, |
| **'class'** | 0.15297059818817663 | **'travel'** | 0.15909033106510667 | **'health'** | 0.18194229083665558 |
|  |  |  |  |  |  |  |
| **2-3 paragraphs** | **'support'** | 0.32899471788278933, | **'market'** | 0.2986573559242641, | **'misinfodem'** | -0.5033743541941365, |
| **'granada'** | 0.24814037209037237, | **'covid'** | 0.2197986796193419, | **'covid'** | -0.3629711503323515, |
| **'covid'** | 0.22769581399154337, | **'friday'** | 0.18596179123202256, | **'articl'** | -0.2145925193519742, |
| **'programm'** | 0.21932981192185955, | **'infect'** | 'infect': 0.15212490284470295, | **'commun'** | -0.14837863098037748, |
| **'research'** | 0.21932981192185955 | **'reuter'** | 0.1465324530795613 | **'mechan'** | -0.14837863098037712 |
|  |  |  |  |  |  |  |
| **Paragraph level (14379|387)** | **'covid'** | 0.8353445537976227, | **'trial'** | -0.46149246082032147, | **'covid'** | 0.44528765261781605, |
| **'imag'** | 0.3551934971949819, | **'vaccin'** | -0.38111970491120056, | **'pandem'** | 0.3598378349771223, |
| **'detect'** | 0.14217453078608513, | **'stage'** | -0.3279128249788164, | **'health'** | 0.33461199438050226, |
| **'model'** | 0.1267002478578127, | **'expect'** | -0.23283573850743006, | **'polici'** | 0.2627823706235591, |
| **'dataset'** | 0.1093369812928102 | **'dose'** | -0.19988003539528476 | **'impact'** | .2152347635291778 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment 4: Document at various levels for large dataset (40 papers COVID-science: Covid-psychology)** | | | | |
|  | | | | |
| **No.of Documents** | **covid-science-large.txt** | | **Covid-psychology-large.txt** | |
| **Top words** | **weights** | **Top words** | **weights** |
| **Article level (40|40)** | **'covid'** | 0.5580672321131127, | **'student'** | 0.31520975301891685, |
| **'imag'** | 0.2828570178358147, | **'psycholog'** | 0.3088225210323284, |
| **'model'** | 0.2574998284866968, | **'covid'** | 0.3082634599864977, |
| **'dataset'** | 0.1425624890560568, | **'train'** | 0.22330563399536046, |
| **'infect'** | 0.13935041563870734 | **'pandem'** | 0.21884380058036093 |
|  |  |  |  |  |
| **2-3 paragraphs** | **'support'** | -0.32899471788278967, | **'misinfodem'** | 0.5033743541941371, |
| **'granada'** | -0.24814037209037254, | **'covid'** | 0.3629711503323513, |
| **'covid'** | -0.22769581399154254, | **'articl'** | 0.21459251935197415, |
| **'programm'** | -0.21932981192185977, | **'mechan'** | 0.14837863098037712, |
| **'research'** | -0.21932981192185977 | **'commun'** | 0.14837863098037704 |
|  |  |  |  |  |
| **Paragraph level (274941|183592)** | **'covid'** | 0.8597188592513563, | **'covid'** | 0.5542346968279828, |
| **'imag'** | 0.24459808570501076, | **'psycholog'** | 0.38587594842118456, |
| **'model'** | 0.21196118253925922, | **'pandem'** | 0.36424816089766787, |
| **'detect'** | 0.14846685076222263, | **'health'** | 0.2988945427299109, |
| **'volum'** | 0.09486963239930264 | **'impact'** | 0.15508253422230434 |

##### 4.4 WordCloud

In this step, generating the wordcloud for the top 20 words with high weights from the LSA model by using the wordcloud application program interface(API) and saving the wordcloud to an image.

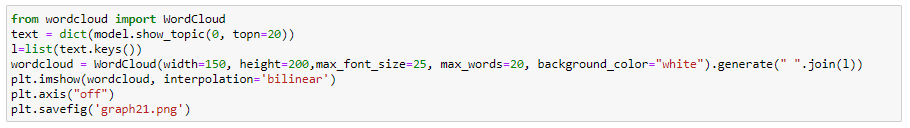


Figure 4.4.1 Code for generating WordCloud

The below table shows WordCloud for the top 20 words from “lsimodel” from 4 experiments.

Table 2 Wordcloud results

|  |  |  |
| --- | --- | --- |
| **Experiment 1: Document at various levels for short dataset (law: non-law)** | | |
| **No.of Documents** | **law-texts-all.txt** | **non-law-all.txt** |
| 1. **Article level (4|4)** | **Text  Description automatically generated** | **Text  Description automatically generated** |
| 1. **2-3 paragraphs** | Text  Description automatically generated | Text  Description automatically generated |
| 1. **Paragraph level (450|3038)** | Text  Description automatically generated | Text  Description automatically generated |

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment 2: Document at various levels for a small dataset (COVID-science: non-science: Covid-psychology)** | | | |
| **No.of Documents** | **covid-science-small.txt** | **Covid-non-science-small.txt** | **Covid-psychology-small.txt** |
| **Article level (10|10)** | Text  Description automatically generated | Text  Description automatically generated |  |
| **2-3 paragraphs** | Text  Description automatically generated | Text  Description automatically generated |  |
| **Paragraph level (7401|229)** | Text  Description automatically generated | Text  Description automatically generated with medium confidence |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment 3: Document at various levels for large dataset (20 papers COVID-science: non-science: Covid-psychology)** | | | |
| **No.of Documents** | **covid-science.txt** | **Covid-non-science.txt** | **Covid-psychology.txt** |
| **Article level (20|20)** | Text  Description automatically generated | Text  Description automatically generated with medium confidence |  |
| **2-3 paragraphs** | Text, qr code  Description automatically generated | Text  Description automatically generated |  |
| **Paragraph level (14379|387)** | Text  Description automatically generated | Text  Description automatically generated |  |

|  |  |  |
| --- | --- | --- |
| **Experiment 4: Document at various levels for large dataset (40 papers COVID-science: Covid-psychology)** | | |
| **No.of Documents** | **covid-science-large.txt** | **Covid-psychology-large.txt** |
| **Article level (40|40)** |  |  |
| **2-3 paragraphs** |  |  |
| **Paragraph level (274941|183592)** |  |  |

##### 4.5 Wordnet

Cosine-similarity is a distance metric used to generate the difference or distance between the words. To generate the distance matrix for the words we use pairwise-cosine-similarity from sklearn. metrics class and to graphically represent and show them in a space we use networkx library.

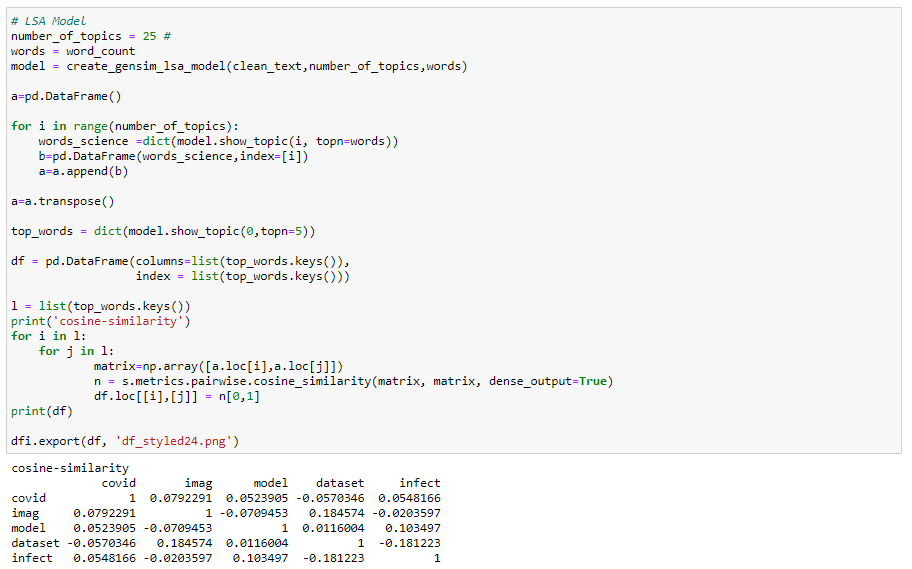


Figure 4.5.1 Code to generate cosine similarity matrix.

|  |  |  |
| --- | --- | --- |
| **Experiment 1: Document at various levels for short dataset (law: non-law)** | | |
| **No.of Documents** | **law-texts-all.txt** | **non-law-all.txt** |
| 1. **Article level (4|4)** |  |  |
| 1. **2-3 paragraphs** |  |  |
| 1. **Paragraph level (450|3038)** |  |  |

Table 3 Cosine similarity results

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment 2: Document at various levels for a small dataset (COVID-science: non-science: Covid-psychology)** | | | |
| **No.of Documents** | **covid-science-small.txt** | **Covid-non-science-small.txt** | **Covid-psychology-small.txt** |
| **Article level (10|10)** |  |  |  |
| **2-3 paragraphs** |  |  |  |
| **Paragraph level (7401|229)** |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment 3: Document at various levels for large dataset (20 papers COVID-science: non-science: Covid-psychology)** | | | |
| **No.of Documents** | **covid-science.txt** | **Covid-non-science.txt** | **Covid-psychology.txt** |
| **Article level (20|20)** |  |  |  |
| **2-3 paragraphs** |  |  |  |
| **Paragraph level (14379|387)** |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Experiment 4: Document at various levels for large dataset (40 papers COVID-science: Covid-psychology)** | | |
| **No.of Document** | **covid-science-large.txt** | **Covid-psychology-large.txt** |
| **Article level (40|40)** |  |  |
| **2-3 paragraphs** |  |  |
| **Paragraph level (274941|183592)** |  |  |

The below table shows the NetworkX graph for the top 5 words from “lsimodel” from 4 experiments.

Table 4 WordNet results

|  |  |  |
| --- | --- | --- |
| **Experiment 1: Document at various levels for short dataset (law: non-law)** | | |
| **No.of Documents** | **law-texts-all.txt** | **non-law-all.txt** |
| 1. **Article level (4|4)** |  |  |
| 1. **2-3 paragraphs** |  |  |
| 1. **Paragraph level (450|3038)** |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment 2: Document at various levels for a small dataset (COVID-science: non-science: Covid-psychology)** | | | |
| **No.of Documents** | **covid-science-small.txt** | **Covid-non-science-small.txt** | **Covid-psychology-small.txt** |
| **Article level (10|10)** |  |  |  |
| **2-3 paragraphs** |  |  |  |
| **Paragraph level (7401|229)** |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Experiment 3: Document at various levels for large dataset (20 papers COVID-science: non-science: Covid-psychology)** | | | |
| **No.of Documents** | **covid-science.txt** | **Covid-non-science.txt** | **Covid-psychology.txt** |
| **Article level (20|20)** |  |  |  |
| **2-3 paragraphs** |  |  |  |
| **Paragraph level (14379|387)** |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Experiment 4: Document at various levels for large dataset (40 papers COVID-science: Covid-psychology)** | | |
| **No.of Document** | **covid-science-large.txt** | **Covid-psychology-large.txt** |
| **Article level (40|40)** |  |  |
| **2-3 paragraphs** |  |  |
| **Paragraph level (274941|183592)** |  |  |

##### 4.6 statistic test

Function for mantel test

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Here we generate statistical difference between the top word by permutating with different populations using mantel test.

Results generated from mantel test

Mantel-Test for Law and Non-Law datasets (4 articles)

|  |  |  |  |
| --- | --- | --- | --- |
| Mantel-Test | r | p-value | z-score |
| Article level | 0.45342 | 0.0127 | 2.935799 |
| 2-3 Paragraphs | -0.21510 | 0.714285 | -0.68497 |
| Paragraph level | -0.29385 | 0.878 | -1.220183 |

Mantel test for science – psychology (40 articles)

|  |  |  |  |
| --- | --- | --- | --- |
| Mantel-Test | r | p-value | z-score |
| Article level | -0.40041 | 0.8811 | -1.21426 |
| 2-3 Paragraphs | -0.12443 | 0.6547 | -0.52135 |
| Paragraph level | -0.35870 | 0.8187 | -1.06277 |

Mantel test for science – psychology (10 articles)

|  |  |  |  |
| --- | --- | --- | --- |
| Mantel-Test | r | p-value | z-score |
| Article level | -0.31881 | 0.87864 | -1.20322 |
| 2-3 Paragraphs | -0.12443 | 0.6625 | -0.527201 |
| Paragraph level | -0.32813 | 0.8417 | -1.00066 |

Mantel test for science – non-science(10 articles)

|  |  |  |  |
| --- | --- | --- | --- |
| Mantel-Test | r | p-value | z-score |
| Article level | 0.16347 | 0.3104 | 0.5929 |
| 2-3 Paragraphs | -0.2556 | 1.0 | -1.1348 |
| Paragraph level | -0.04745 | 0.5642 | -0.17695 |

# Chapter 5 Ethical, Legal and Social issues

##### 5.1 Ethical Issues

No matter what technology has served humanity with endless possibilities in life. Most of the times Technology is mostly considered unbiased and ethically untroubled. However, Machine Learning or Natural Language Processing algorithms became part of human daily life. These algorithms are developed and implemented by human beings, raising questions with ethical implications(Bates, 2019).

Latent Semantic Analysis(LSA) is applied to collect key terms based on their contextual similarity. This algorithm often overlooks peoples minds over social disruptions of upcoming technologies and later provides technical directions of how relevant human values could be translated into the design to avoid the occurrence of ramifications(Kwon and Park, 2018). The data collected from newspapers, research papers concerns personal data exploitation.

While words extraction from the huge corpus of real-time data might sometimes lead to extraction of personal information.

Considering these ethical issues, to specify ethical zones from the ethical tool kit

Ethical OS – Risk Zone 6: Data Control and monetization

Ethical OS – Machine Ethics and Algorithmic Biases

##### 5.2 Legal Issues

Often the data collected for this experiment are collected from research papers, newspapers to semantically analyse that corpus of text. The data collected need to be authentic and look for any copyrights.

* Copyright Infringement and Receipt of Work – data collected need to be copyright free.
* The Data Protection Act(2018) – Following data protection principles, considering processing and use of data and safe holding of accurate data for no longer than absolutely necessary.
* GDPR Regulations – Consideration of personal data from the newspapers or specific experimenting details from research papers.

##### 5.3 Social Issues

Machine learning algorithms in information retrieval systems are associated with several social and ethical issues to their use, nature and design. LSA like a clearly defined methodology with the developers would lead to social concerns.

##### Project Plan

Week 1 – 25/05/21 -Start

Week 24 – 3/12/21 – end(Last day of submission)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| week | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| Activity |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Proposal finished |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Literature review |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data collection |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Develop App(experiments) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data Analysis/ Statistical Test |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Write up |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Submit final draft |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Make final changes |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Submission |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# Bibliography

Anwar, A., 2006. Experiments with latent semantic analysis for word tagging, in: Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing, CSECS’06. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, pp. 275–278.

Ashby, W.R., 1961. An introduction to cybernetics. Chapman & Hall Ltd.

Bates, M.J., 2019. Understanding information retrieval systems: management, types, and standards. Auerbach Publications.

Broeck, J.V. den, Cunningham, S.A., Eeckels, R., Herbst, K., 2005. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLOS Medicine 2, e267. https://doi.org/10.1371/journal.pmed.0020267

Brooks-Gordon, B., Freeman, M., 2006. Law and Psychology: Current Legal Issues. Oxford University Press UK.

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M.X., Qu, H., 2010. Context preserving dynamic word cloud visualization, in: 2010 IEEE Pacific Visualization Symposium (PacificVis). Presented at the 2010 IEEE Pacific Visualization Symposium (PacificVis), pp. 121–128. https://doi.org/10.1109/PACIFICVIS.2010.5429600

Dascalu, M., McNamara, D., 2017. How Important Is Size? An Investigation of Corpus Size and Meaning in both Latent Semantic Analysis and Latent Dirichlet Allocation.

Dennis, S., Dennis@colorado, S., Edu, 2003. A comparison of statistical models for the extraction of lexical information from text corpora.

Diniz-Filho, J.A.F., Soares, T.N., Lima, J.S., Dobrovolski, R., Landeiro, V.L., de Campos Telles, M.P., Rangel, T.F., Bini, L.M., 2013. Mantel test in population genetics. Genet Mol Biol 36, 475–485. https://doi.org/10.1590/S1415-47572013000400002

Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, R., 1988. Using latent semantic analysis to improve access to textual information, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’88. Association for Computing Machinery, New York, NY, USA, pp. 281–285. https://doi.org/10.1145/57167.57214

Foltz, P.W., 1996. Latent semantic analysis for text-based research. Behavior Research Methods, Instruments, & Computers 28, 197–202. https://doi.org/10.3758/BF03204765

Kontostathis, A., Pottenger, W., 2002. Detecting Patterns in the LSI Term-Term Matrix. undefined.

Kwon, H., Park, Y., 2018. Proactive development of emerging technology in a socially responsible manner: Data-driven problem solving process using latent semantic analysis. Journal of Engineering and Technology Management 50, 45–60. https://doi.org/10.1016/j.jengtecman.2018.10.001

Landauer, T., Foltz, P., Laham, D., 1998. An Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T.K. (Ed.), 2014. Handbook of latent semantic analysis, 1. iss. in paperback. ed. Routledge, New York.

Meena, K., Raj, L., 2014. Evaluation of the descriptive type answers using hyperspace analog to language and self-organizing map, in: 2014 IEEE International Conference on Computational Intelligence and Computing Research. Presented at the 2014 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–5. https://doi.org/10.1109/ICCIC.2014.7238415

Mohamed, R., Watada, J., 2010. An evidential reasoning based LSA approach to document classification for knowledge acquisition, in: 2010 IEEE International Conference on Industrial Engineering and Engineering Management. Presented at the 2010 IEEE International Conference on Industrial Engineering and Engineering Management, pp. 1092–1096. https://doi.org/10.1109/IEEM.2010.5674188

Morgan, G., Harmon, R., 2001. Data Collection Techniques. Journal of the American Academy of Child and Adolescent Psychiatry 40, 973–6. https://doi.org/10.1097/00004583-200108000-00020

Pilato, G., Vassallo, G., 2015. TSVD as a Statistical Estimator in the Latent Semantic Analysis Paradigm. IEEE Transactions on Emerging Topics in Computing 3, 185–192. https://doi.org/10.1109/TETC.2014.2385594

Rosenbusch, H., Wanders, F., Pit, I.L., 2020. The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. Psychological Methods 25, 380–392. https://doi.org/10.1037/met0000244

Von Bertalanffy, L., 1967. General theory of systems : Application to psychology. Social Science Information 6, 125–136. https://doi.org/10.1177/053901846700600610

Wenli, C., 2016. Application Research on Latent Semantic Analysis for Information Retrieval, in: 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). Presented at the 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 118–121. https://doi.org/10.1109/ICMTMA.2016.37