**Systematic Communication Modelling with Latent Semantic Analysis**

Contents

[Abstract: 3](#_Toc85433956)

[Chapter 1 Introduction 3](#_Toc85433957)

[Chapter 2 Literature Review 4](#_Toc85433958)

[Chapter 3 Methodology 8](#_Toc85433959)

[Chapter 4 Implementation 16](#_Toc85433960)

[Chapter 5 Results: 37](#_Toc85433961)

[Bibliography 43](#_Toc85433962)

List of Figures

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)

# Abstract:

Traditional Information retrieval systems are designed to retrieve the information from the source by giving a query or group of strings we are searching for. By matching those strings with the source data to retrieve the most relevant document from the internet or any other databases. The way similarity of the document is usually calculated by matching the search string with the information in the documents but retrieving the information by a statistical method can be most relevant and effective. Like searching for a research paper of a specific category like scientific research papers, psychology research papers with a statistical method will be helpful. To retrieve specific categorized information statistically, an application or a method should be developed, which could differentiate the different types of content available in different documents semantically. Likewise in this project developing a statistical method that could differentiate scientific articles from psychology articles using Latent Semantic Analysis.

# 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 or a method needs to be developed to analyse the content available in the document and differentiate it from other documents statistically.

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 sub lexical 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.

Research Questions:

What machine learning algorithm is effective to analyse the huge corpus of data in a document?

What statistical test best fits to evaluate the difference between the scientific articles and psychology articles?

# 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, Latent Semantic Analysis (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. Unsupervised learning techniques can be used to build higher-level representations of word meaning due to statistical redundancies observed in the language(Kintsch, 2001).

Choosing the corpus size plays an important role in obtaining test results. Large, heterogeneous corpora are a risk because they may include more noise or too much specific information from a single area, lowering the accuracy of the resulting models. However, there is no consensus on the predicted corpus size or what constitutes a large or small corpus(Dascalu and McNamara, 2017).

The author's study not only highlights the advantages of employing larger corpora to create LSA and LDA models but also paves the way for future research. The requirement to study LSA and LDA models constructed on large and small corpora that are counterbalanced in terms of domains covered and linguistic complexity is one of the most important of them. Future research should also look into the possibility of LDA models contributing to our knowledge of cognitive processing(Dascalu and McNamara, 2017).

##### 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.

Researchers have recently advanced beyond LSA models of semantic representations to investigate topic-based models of semanticity, the most prevalent of which is Latent Dirichlet Allocation(Blei et al., 2003). Because each word in the semantic vector space model can be represented as a context-free vector, LSA can be thought of as a cognitive statement of meaning. Although individual dimensions of concepts do not have a distinct meaning, the total representation provided by LSA can be thought of as a map of meanings. Positive correlations between LSA similarity scores and human recall using word association lists further support a semantic proximity effect in which LSA resembles human memory(Kintsch, 2001).

To embed words in a vectorial space, corpus-based semantic representations take advantage of statistical features of textual structure. Term pairs with comparable meanings tend to cluster together in this space. These approaches are based on the notion that words with similar meanings are more likely to appear in comparable contexts. This theory is known as the distributional hypothesis, and it provides a useful framework for comprehending and computing semantic relationships between words(Altszyler et al., 2017).

To learn correct word embeddings in small text corpora, the author compares the capabilities of Skip-gram and LSA. To accomplish so, we used nested subsamples of a medium-sized corpus to assess the model's capacity to represent semantic categories (such as drinks, countries, tools, and garments). When models are trained with medium-sized datasets ( 10 million words), Word2vec embeddings beat LSAs. When the corpus size is lowered, however, Word2vec performance suffers significantly, making LSA the more appropriate technique. This discovery adds to the debate over prediction-based vs. counter-based models. They assume that the performance drop of 10 Word2vec in small corpora is because prediction-based models require a large amount of training data to fit their large number of parameters.

We used LSA and Skip-gram capabilities to extract significant semantic word correlations in dream reports as a case study. Even in cases of series with a small number of dreams and a low frequency of target words, we discovered that LSA can accurately identify semantic word relations. This is a step forward in the application of word embeddings to dream content analysis. This branch of study investigates things like "what do we dream about?" and "how do gender, cultural background, and waking life events impact the content of our dreams?" They recommend that LSA be used to investigate word connections in dream reports, potentially shedding new light on an ancient topic of psychology research(Altszyler et al., 2017).

There is a constant proliferation of novel notions and scales in psychological measurement and theory. Psychological science is grappling with arbitrary measurement, construct dilution, and separation between research groups due to the sometimes-redundant nature of new scales. We introduce the Semantic Scale Network, an easy-to-use online tool, to overcome these challenges. The goal of this application is to use latent semantic analysis to find semantic overlap between scales automatically. Authors and reviewers can insert items for a new scale into the app and receive semantic overlap quantifications with related scales in the app's corpus. Unlike typical scale overlap evaluations, the programme can allow expert scale redundancy judgments without access to empirical data or understanding of all possible associated scales. They explain the Semantic Scale Network and discuss best practices for interpreting its outputs after a brief introduction to metrics of semantic similarity in texts(Rosenbusch et al., 2020).

Using two criteria, classify a network-similarity algorithm. First, at what network level does it operate? Second, what kind of comparison is it employing? We identify three levels for the first criterion: micro, meso, and macro. At the micro-level, a method extracts features from nodes or egonets;1 at the mezzo-level, it extracts features from communities; and at the macro-level, it extracts features from the global/network level, as their names suggest. We have three sorts of criteria for the second criterion: vector-based, classifier-based, and matching-based(Soundarajan et al., 2014).

In applications like transfer learning and change detection, quantifying the difference between two networks is crucial. For this goal, several network similarity algorithms have been proposed. The problem's origins can be traced back to the problem of [determining graph isomorphism], for which no polynomial-time algorithm exists. The task of network similarity is significantly broader(Soundarajan et al., 2014).

##### 2.2 Statistical test

Latent Semantic Analysis can be used to generate principal components of a text corpus (Slomovitz, 2017). But to compare these principal components with other text corpora, by considering each principal components elements as populations we need a statistical test. A statistical test that could differentiate different types of corpuses need to be defined.

Many statistical tests could be used to compare two populations. Some of them are for paired or matched observation, nominal data - McNemar’s Test, For Ordinal (Ordered categories) - Wilcoxon test, For Quantitative data (Discrete or Non-Nominal) – Wilcoxon, For Quantitative (Normal\*) – Paired t-test.

For comparing two or more groups there are two categories of tests parametric tests and non-parametric tests. In Parametric tests, they are Paired t-test, Unpaired t-test, Pearson correlation, One-way analysis of variance while their equivalent non-parametric tests are Wilcoxon rank-sum test, Mann-Whitney U test, Spearman correlation test, Kruskal Wallis Test(Campbell, 2021).

When sampling from multivariate normal populations, Kullback's statistic for verifying equality of many correlation matrices can be considered a modified likely hood ratio statistic, according to the article(Gupta et al., 2013). Expand data in terms of other random variables and then reverse the expansion term by term to get the asymptotic null distribution of data in series involving independent chi-square variables. An example is also provided to demonstrate the technique to be followed when using the statistic to assess the equivalence of correlation matrices.

The author of this paper(Gupta et al., 2013) looked at the difficulty of assessing how similar two networks are when node-correspondences are unknown. In real-world applications such as transfer learning and change detection, this issue arises regularly. There are a plethora of network similarity approaches to choose from, and it's difficult to know which one to use. The first empirical examination of the relationships between various network similarity approaches is presented. We provide (1) a method for identifying groups of comparable network similarity methods and (2) a method for calculating consensus among a collection of network-similarity methods. By applying our approaches to a variety of real datasets across several areas, we compare and contrast twenty network-similarity strategies. Our findings show that (1) multiple network-similarity approaches are surprisingly well correlated, (2) certain complex network-similarity methods may be approximated by a much simpler method, and (3) a few network similarity methods provide rankings that are extremely near to the consensus ranking.

It is desirable to be able to make statistical comparisons between correlation coefficients recorded on the same individuals in a variety of scenarios in psychological research. For example, an experimenter might want to see if two predictors have the same correlation with a criterion variable. In a different scenario, the investigator might want to see if a whole matrix of correlations has remained stable over time. The current paper examines the literature on such tests, identifies specific statistics to avoid, and describes a number of procedures that can be applied securely with medium to large samples. A number of numerical examples are offered to help illustrate the point(Steiger, 1980).

The former problem is handled in traditional statistical analyses by multivariate approaches that allow one to pay to the correlations among predictor variables; partial regression is a well-known solution to this problem. Path analysis is an interpretive method for conceptually distinguishing causal linkages from spurious associations resulting from chance correlations between variables. However, the second difficulty, namely that autocorrelation in the variables violates the principles of parametric analysis, confounds traditional parametric techniques(Diniz-Filho et al., 2013).

The Mantel test is a method for overcoming some of the difficulties that come with describing species-environment connections. Mantel's test is a regression in which the variables are distance or dissimilarity matrices that sum up pairwise similarities across sample sites(Diniz-Filho et al., 2013).

# 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 analysing 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.

Before collecting data, researchers must first determine whether scales are redundant to the new instrument, as they will need participant data to quantify shared variation. Despite the greatest attempts, it is difficult to be aware of every scale that might be important to one's research despite best efforts. Relevant scales are frequently published under various names or in unrelated areas and hence may go unnoticed by researchers. Second, researchers must collect data from the same test subjects for all scales, which might be difficult if there are too many related scales for each participant to complete. Third, the line between strong convergent validity and redundancy does not exist(Kabir, 2016).

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 scientific research papers and the other with psychology research papers.

##### 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).

Data cleaning, also known as data cleansing or scrubbing, is the process of discovering and removing mistakes and inconsistencies from data to improve its quality. Single data collections, such as files and databases, might have data quality issues, such as misspellings during data entry, missing information, or other erroneous data. The necessity for data cleansing increases dramatically when various data sources must be connected, such as in data warehouses, federated database systems, or worldwide web-based information systems. This is due to the fact that the sources frequently provide redundant data in various forms. Consolidation of disparate data formats and the deletion of redundant information is required in order to enable access to correct and consistent data(Broeck et al., 2005).

Text pre-processing is a technique for cleaning text data and preparing it for use in a model. Text data comprises noise in the form of emotions, punctuation, and text in a different case, among other things. When it comes to Human Language, there are many different ways to communicate the same thing, and this is only the beginning of the challenge. Machines cannot understand words; they require numbers; thus, we must convert text to numbers efficiently.

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 Remove Punctuations, remove words and digits containing digits, Remove, stop words, Rephrase Text, Stemming and Lemmatization, Lower case, Remove White spaces need to be performed.

Here is the sample example for data pre-processing

Data before pre-processing:

['REGINA v SECRETARY OF STATE FOR THE HOMEDEPARTMENT, Ex parte HINDLEY[COURT OF APPEAL][2000] 1 QB 152HEARING-DATES: 5, 6, 7, October 5 November 19985 November 1998CATCHWORDS:Prisons - Prisoners\' rights - Release on licence - Mandatory life sentence prisoner - Tariff element of determinate length provisionally fixed but not communicated to prisoner’]

Applying these pre-processing techniques:

|  |  |  |
| --- | --- | --- |
| **Pre-processing technique** | **Function** | **purpose** |
| Regex tokenizer | RegexpTokenizer(r'\w+') | Used to extract the tokens from string by using regular expression. |
| Stop word list | set(stopwords.words('english')) | Stop words are frequently used in Text Mining and Natural Language Processing (NLP) to exclude terms that are so widely used that they contain little meaningful information. Ex: “a”, “the”, “is”, etc. |
| Porter Stemmer | PorterStemmer() | The Porter stemmer is a method for removing frequent morphological and inflexional endings from English words. Ex: (play, playing) 🡪 play |

Data after pre-processing:

['regina', 'secretary', 'state', 'homedepartment', 'parte', 'hindley', 'court', 'appeal', '2000', '152’,’hearing', 'dates', 'october', 'november', '19985', 'november', '1998’, ’catchwords', 'prisons', 'prisoners', 'rights', 'release', 'licence', 'mandatory', 'life', 'sentence', 'prisoner', 'tariff', 'element', 'determinate', 'length', 'provisionally', 'fixed', 'communicated', 'prisoner']

##### 3.3 Latent Semantic Analysis (LSA)

The most widely used techniques for determining document proximity can be split into three categories: Adaptive Strategies (AS) such as probabilistic, neural networks, and genetic algorithms, as well as Vector Space Models (VSM) and Latent Semantic Indexing (LSI)(Ibrahimov et al., 2002). For this project, we focus on LSI.

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. model’s 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.

Steps involved in LSA:

Diagram

Description automatically generated

Figure 1 Hierarchical structure of lsa (Databricks Academy, 2019)

Function for LSA

from gensim.models import LsiModel

model = LsiModel(common\_corpus, id2word=common\_dictionary)

The dilemma of how to locate relevant documents from search phrases prompted, the development of Latent Semantic Analysis. When we compare words to identify relevant papers, the basic challenge that occurs is to compare the meanings or concepts behind the words. But LSA tries to tackle this problem by putting both words and documents in a "concept" space and comparing them there.

Chart, bubble chart

Description automatically generated

Figure 2 LSA sorting words to their concept space (Bhagwant, 2011)

##### 3.4 Word Cloud

Word Cloud 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. Word Cloud 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).

Although word clouds are more versatile than traditional time series views when it comes to displaying intricate word associations, they aren't usually built for side-by-side comparison. For example, the size and position of a word in distinct word clouds normally vary, and words regularly appear and depart over time. Exploring temporal trends of documents with varied time stamps is difficult using traditional word clouds. We present a versatile method for creating a word cloud layout tailored to such papers. To suit diverse user requirements, our technique can structure layouts according to multiple semantic-coherence criteria (Cui et al., 2010).

Pipeline for creating a word cloud. A group of words is retrieved from a collection of documents to start the pipeline (see Figure a). Then, based on their properties, we put all extracted words on a 2D plane (see Figure b). filter out any unimportant or unrelated words from the 2D plane (see Figure c) and generates a triangular mesh using Oelau-nay triangulation" on the remaining points, each of which is at the centre of a word. The font size of each word is determined by the frequency of that word at that moment (see Figure d). Finally, to alter point placements and get a suitable layout, we use an adapted force-directed algorithm (see Figure e)(Cui et al., 2010).

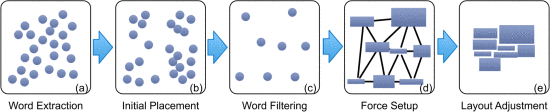


Figure 3 The pipeline for creating a semantic and stable word cloud layout (Cui et al., 2010)

Steps:

1. Initially set of words with different timestamps are generated from the documents;
2. Multidimensional scaling is used to place the extracted words on a 2D plane;
3. Unrelated words are filtered out at this step;
4. The remaining words are set up in a network;
5. Optimising the arrangement using a force-directed method (Cui et al., 2010).

Word cloud API:

from wordcloud import WordCloud

WordCloud([font\_path, width, height,…]) #Word cloud object for generating and drawing

**Text

Description automatically generated**

Figure 4 Word cloud representation

##### 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.

NetworkX is a Python library that allows you to explore and analyse networks and network methods. Simple graphs, directed graphs, graphs with parallel edges, and graphs with self-loops are all represented by the core package's data structures. NetworkX graphs can have any (hashable) Python object as a node, and edges can have any data as an edge; this flexibility made NetworkX excellent for expressing networks in a variety of scientific domains. Many graph algorithms are implemented in addition to the basic data structures for determining network attributes and structure metrics, such as shortest paths, betweenness centrality, clustering, and degree distribution, among others. NetworkX includes generators for several classic graphs and common graph models, including the Erdoes-Renyi, Small World, and Barabasi-Albert models, as well as the ability to read and write many graph formats for easy interchange with existing data. NetworkX is a strong tool for scientific computations because of the Python programming language's ease of use and flexibility, as well as its connection to the SciPy tools. We share some of our recent work on coupled oscillator synchronisation to show how NetworkX facilitates research in the field of computational networks(Hagberg et al., 2008).

Function for Networkx:

Import networkx as nx

nx.draw(G,pos,alpha=1,width,with\_labels = True,node\_size, edge\_color,node\_color,..)

Diagram

Description automatically generated

Figure 5 5-word Networkx representation from a document

Where:

The nodes represent the words from the documents and the edges represent the weight of similarity between the words in the document.

##### 3.6 Statistical test – Mantel test

To compare the correspondence of one set of distances with another, a variety of approaches have been created. Among these are (1) matrix correlation techniques, (2) Spielmen's network matching techniques, (3) matrix dilation and rotation techniques, and (4) Lingoes and Guttman's smallest-space approaches. All of these strategies have trouble determining statistical significance. The issue is that all pairwise distances between k units (in this case semantics) cannot be independent.

One of the most common techniques for evaluating spatial dynamics driving population structure is to compare genetic divergence or genetic distances, estimated by pairwise FST and associated statistics, with geographical distances, as measured by the Mantel test (Diniz-Filho et al., 2013). Using a similar method, we apply the mantel test to the semantic word population and use the mantel test to calculate the geographical distance in the space. To assess the importance of various populations. We must experiment with various populations (top words) derived from the corpus of texts.

When comparing one corpus to another, we need to see if the shape of the global representation of the document collection is stable. The matrix of pairwise similarities across the papers is one way to depict this global form. The Mantel test is a technique for comparing two similarity matrices. This test is a popular way to evaluate the relationships between two distance matrices, or more broadly, two similarity or proximity matrices. It involves using a statistic r to assess the significance of the measure of connection between the elements in two -matrices and then comparing it to the distribution of the values discovered by randomly reallocating the order of the elements in one of the matrices. For two similarity matrices, we used the statistic r, which is the sum of the products of the standardised similarities(Besançon and Rajman, 2002).

“Mantel's test is based on a regression analysis, in which the variables are distance dissimilarity (or similarity) matrices summarizing all pairwise sample combinations”(Diniz-Filho et al., 2013).

Assuming, 2D data consisting of group of top similarity words (Z(s1),Z(s1),…, Z(sN). If the pixel values are correlated or spatially dependent, spatial coordinates that are near to one another will be paired with intensity values that are likewise close to one another. To statistically model the clustering process.

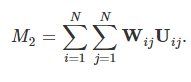
Wij=∥si−sj∥ Describes the spatial proximity of si and sj

And

Uij=|Z(si)−Z(sj)| quantifies the closeness in range Z(si) and Z(sj)

Where i and j represent to integer positions on sampling lattice

Mantel test statistic (Khademi et al., 2009)



Example for evaluating Mantel test:

Considering multiple samples data in two different spaces (Biological and Environmental space). Where nodes represent the samples and the edges represent the relation between the samples in terms of their wights, by considering different parametric axis (Species abundance in biological space and mean sea surface temperature, tidal current and depth in environmental space).

Diagram

Description automatically generated

Figure 6 mantel test for samples at different spaces

Then obtain the pair-wise distance matrices in both space

Function for pairwise distance matrix

sklearn.metrics.pairwise.nan\_euclidean\_distances(matrix, matrix)

To check the similarity or significance between the two pairwise distance matrices, need to calculate the correlation between the two matrices. And describe the similarity odr relation in terms of Z, r, and p metrics

Using formulae:

Text

Description automatically generated with low confidence

Diagram

Description automatically generated

Mantel test statistics(z, r, and p-values) can be obtained by

Import mantel

mantel.test(X, Y, perms, method, tail, …)

where X, Y are the pairwise distance matrices

# Chapter 4 Implementation

6 different experiments

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 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 Xplore, 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 like 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 Xplore, 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.

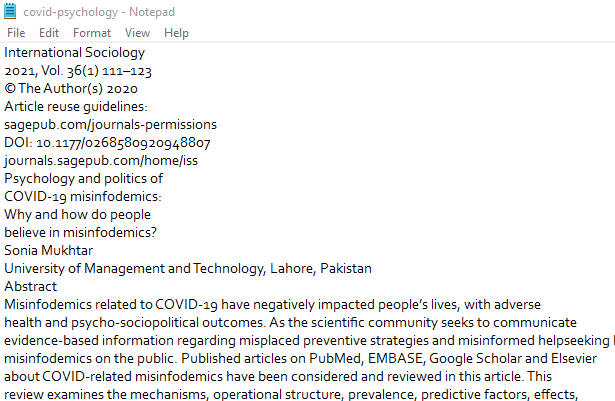


Figure sample data from covid-psychology.txt

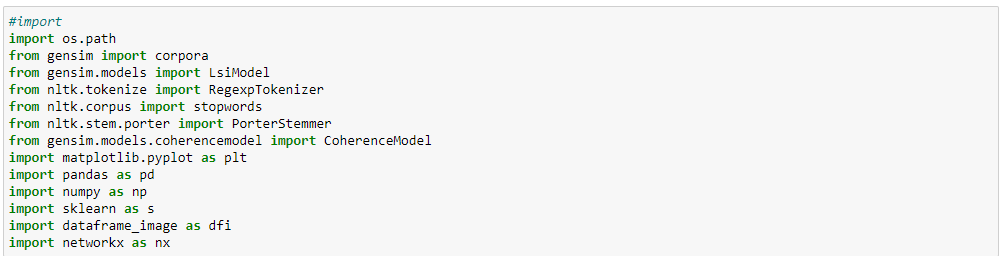


Figure Importing libraries



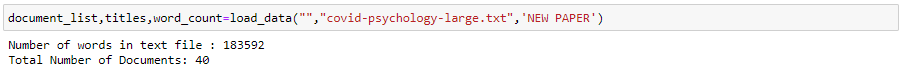


Figure code for loading data at article level from experiment 4 – “covid-science-large.txt” dataset.



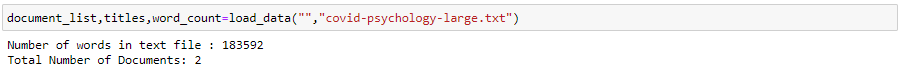


Figure code for loading data at 2-3 paragraph-level from experiment 4 – “covid-science-large.txt” and “covid-psychology-large.txt” dataset.

##### 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 pre-processing techniques were applied to the data like tokenizing, removing stop words, and stemming.

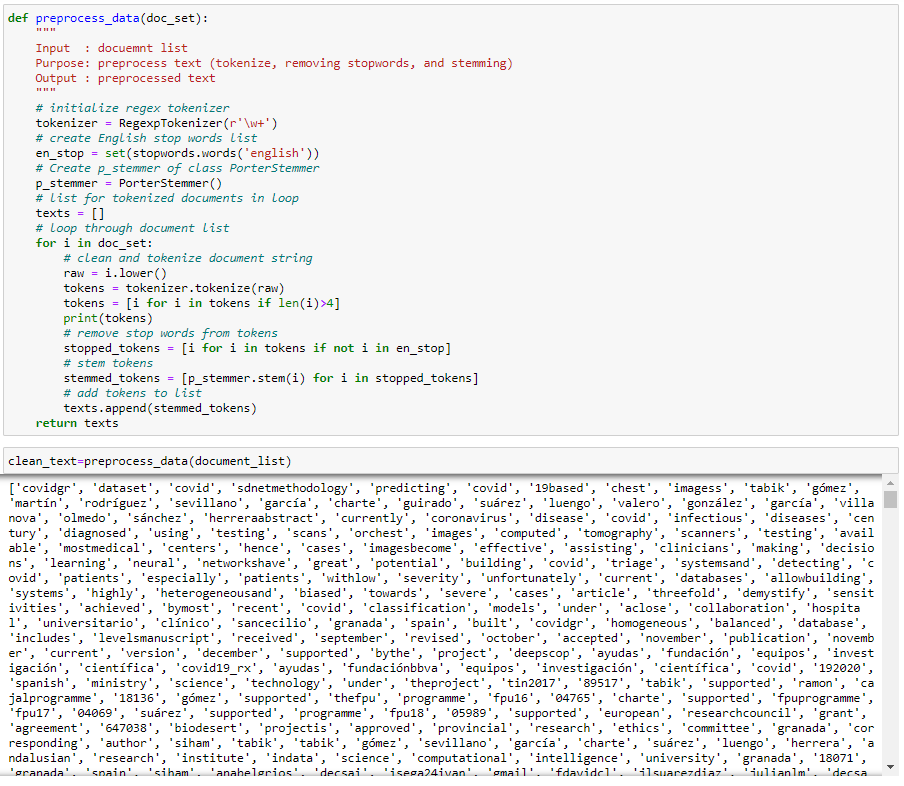


Figure code for Pre-processing.

##### 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.

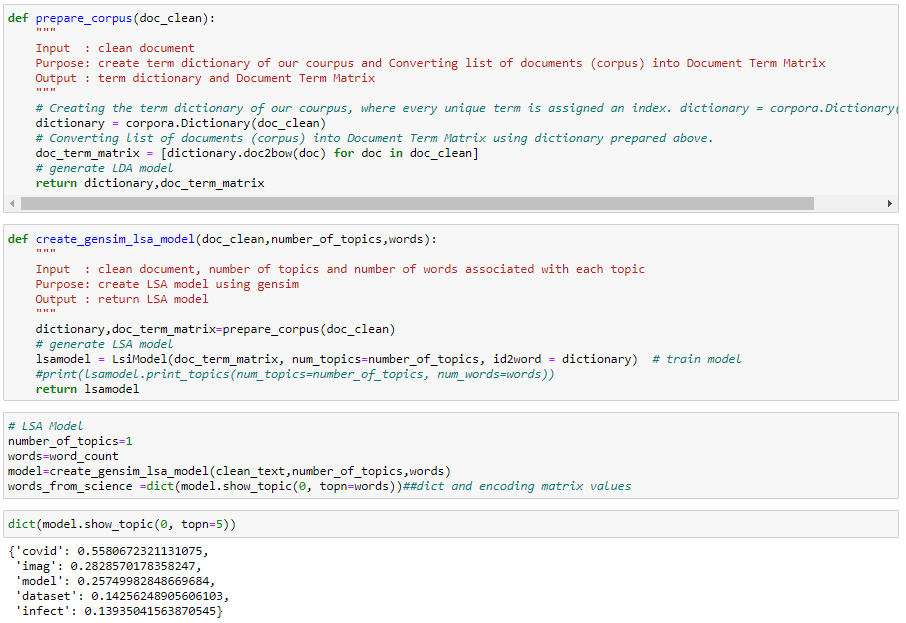


Figure Code for modelling LSA and generating top 5 words at article level from “covid-science-large.txt” dataset.

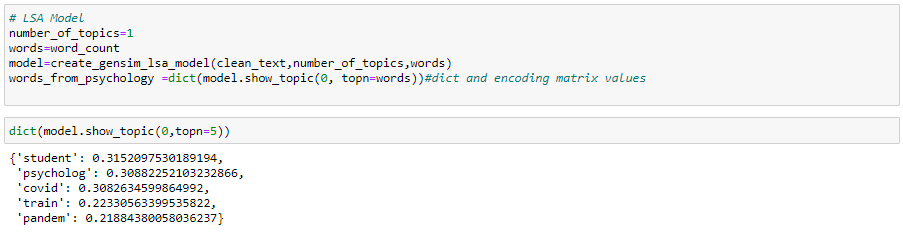


Figure Code for modelling LSA and generating top 5 words at article level from “covid-psychology-large.txt” dataset.

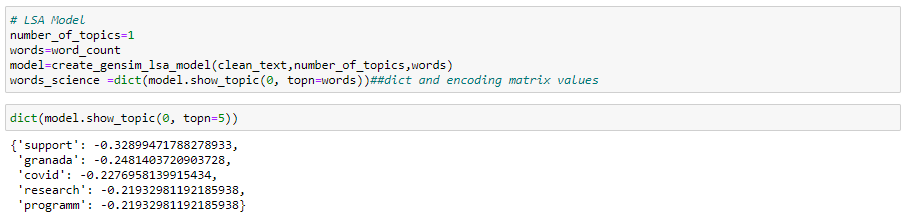


Figure Code for modelling LSA and generating top 5 words at 2-3 paragraph-level from “covid-science-large.txt” dataset.

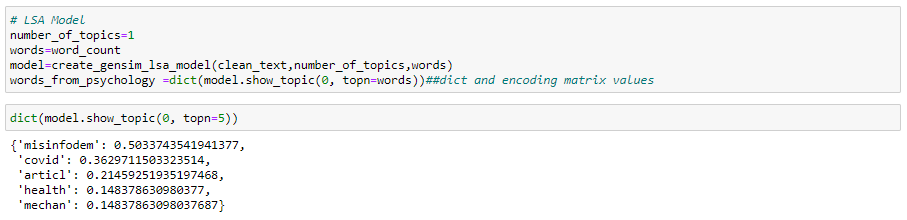


Figure Code for modelling LSA and generating top 5 words at 2-3 paragraph-level from “covid-psychology-large.txt”

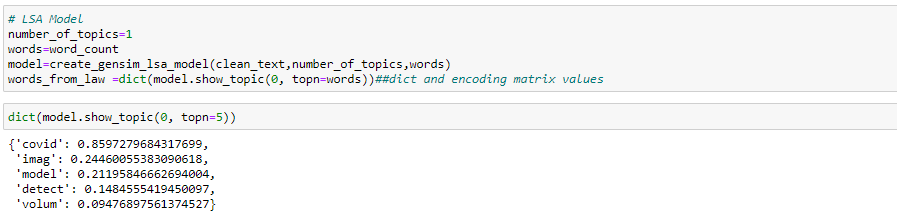


Figure Code for modelling LSA and generating top 5 words at paragraph-level(all paragraphs) from “covid-science-large.txt” dataset.

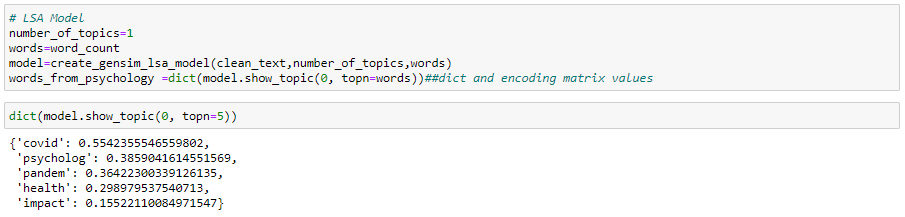


Figure Code for modelling LSA and generating top 5 words at paragraph-level(all paragraphs) from “covid-psychology-large.txt” dataset.

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

Table 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.

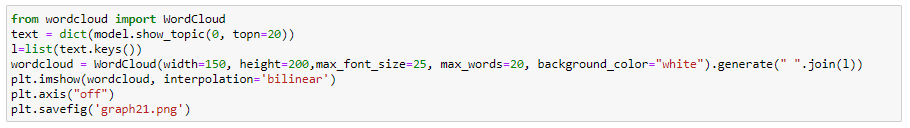


Figure 4.4.1 Code for generating WordCloud

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

Table 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.

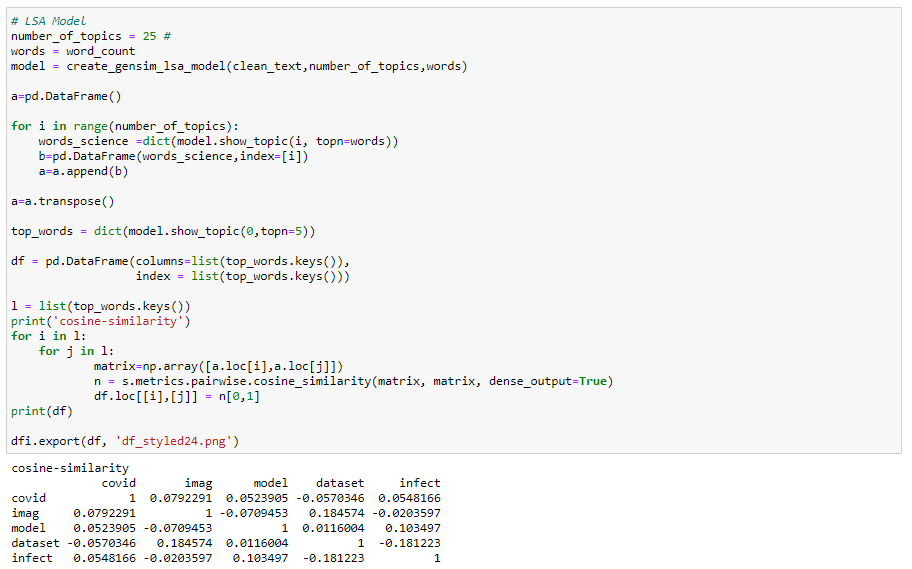


Figure 4.5.1 Code to generate cosine similarity matrix.

|  |  |  |
| --- | --- | --- |
| **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 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 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

# Chapter 5 Results:

Mantel test results

mantel (X’, Y’)

where X’, Y’ are pairwise distances between top words respectively.

X Y

Law dataset Nonlaw dataset

Covid science Psychology (40 articles)

Covid science Psychology (10 articles)

Covid science Psychology (20 articles)

Results generated from mantel test with 10 words network and 20-word network

10-word network: By comparing Top 5 words from each dataset

20-word network: By comparing Top 10 words from each dataset

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article Level | 0.45342 | 0.0127 | 2.935799 | 0.12178 | 0.1489 | 1.06275 |
| 2-3 Paragraphs | -0.21510 | 0.714285 | -0.68497 | 0.06725 | 0.3282 | 0.40405 |
| Paragraph level | -0.29385 | 0.878 | -1.22018 | -0.31741 | 0.9493 | -1.71181 |

Mantel test for science – psychology (10 articles dataset)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article Level | -0.31881 | 0.87864 | -1.20322 | -0.37001 | 0.96258 | -1.69272 |
| 2-3 Paragraphs | -0.12443 | 0.6625 | -0.52720 | -0.16750 | 0.88753 | -1.14787 |
| Paragraph level | -0.32813 | 0.8417 | -1.00066 | -0.07063 | 0.61801 | -0.32868 |

Mantel test for science – psychology (15 articles dataset)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article Level | -0.55832 | 0.9722 | -1.76688 | -0.41870 | 0.9848 | -1.91222 |
| 2-3 Paragraphs | -0.12443 | 0.6549 | -0.51143 | -0.15851 | 0.8535 | -1.04880 |
| Paragraph level | -0.36722 | 0.8156 | -1.09169 | -0.47758 | 0.9934 | -2.17452 |

Mantel test for science – psychology (20 articles dataset)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article Level | -0.45097 | 0.92318 | -1.40863 | -0.45208 | 0.99278 | -2.32042 |
| 2-3 Paragraphs | -0.12443 | 0.65294 | -0.51520 | -0.15851 | 0.85532 | -1.05703 |
| Paragraph level | -0.47229 | 0.92942 | -1.43762 | -0.26062 | 0.88279 | -1.19640 |

Mantel test for science – psychology (30 articles dataset)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article Level | -0.44085 | 0.9074 | -1.33729 | -0.13052 | 0.7165 | -0.58437 |
| 2-3 Paragraphs | -0.12443 | 0.6558 | -0.51814 | -0.16750 | 0.8834 | -1.15334 |
| Paragraph level | -0.42116 | 0.9493 | -1.29004 | -0.36792 | 0.9549 | -1.65607 |

Mantel test for science – psychology (40 articles dataset)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article Level | -0.10233 | 0.6703 | -0.45536 | 0.16347 | 0.3104 | 0.5929 |
| 2-3 Paragraphs | -0.16750 | 0.8881 | -1.14848 | -0.2556 | 1.0 | -1.1348 |
| Paragraph level | -0.29198 | 0.9005 | -1.29744 | -0.04745 | 0.5642 | -0.17695 |

Mantel test results vertical comparison

X Y

Article 2-3 p

Article paragraph

2-3 p paragraph

mantel (X”, Y”)

where X”, Y” are pairwise distances between top words respectively.

For each dataset Law dataset, non-Law dataset, Covid science dataset (40 articles and 20 articles), Covid psychology dataset (40 articles and 20 articles.

Results generated from mantel test with 10 words network and 20-word network

10-word network: By comparing Top 5 words from each dataset

20-word network: By comparing Top 10 words from each dataset

Law dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | -0.09916 | 0.5375 | -0.36904 | -0.16939 | 0.9614 | -1.58676 |
| Article ---> Paragraph | 0.30533 | 0.183333 | 1.0028 | 0.29004 | 0.1036 | 1.44017 |
| 2-3 P ---> Paragraph | 0.21549 | 0.14583 | 1.21009 | 0.19494 | 0.0558 | 1.68612 |

Non-Law dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | 0.08840 | 0.2401 | 0.72033 | 0.08802 | 0.2293 | 0.78101 |
| Article ---> Paragraph | 0.43517 | 0.01626 | 1.81199 | 0.55781 | 0.001 | 2.73972 |
| 2-3 P ---> Paragraph | -0.24308 | 0.7751 | -0.8898 | -0.11016 | 0.7182 | -0.56198 |

Covid science dataset (15 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | 0.27763 | 0.1765, | 1.06224 | 0.13258 | 0.2438 | 0.72441 |
| Article ---> Paragraph | 0.69915 | 0.01805 | 1.78310 | 0.49915 | 0.0257 | 1.66534 |
| 2-3 P ---> Paragraph | 0.22988 | 0.1795 | 1.02993 | 0.02976 | 0.4411 | 0.18913 |

Covid psychology dataset (15 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | -0.63556 | 0.9678 | -2.00749 | -0.28840 | 0.9284 | -1.47024 |
| Article ---> Paragraph | 0.37749 | 0.17301 | 0.94661 | 0.43386 | 0.0024 | 1.85060 |
| 2-3 P ---> Paragraph | -0.02156 | 0.4867 | -0.06129 | -0.22712 | 0.8713 | -1.17128 |

Covid science dataset (20 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | 0.28155 | 0.1844 | 1.04480 | 0.18832 | 0.1602 | 1.03448 |
| Article ---> Paragraph | 0.78462 | 0.02916 | 1.83809 | 0.48517 | 0.0043 | 1.67499 |
| 2-3 P ---> Paragraph | 0.26279 | 0.1414 | 1.17312 | -0.00810 | 0.5304 | -0.05190 |

Covid psychology dataset (20 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | -0.12292 | 0.6418 | -0.38989 | -0.22976 | 0.873 | -1.17834 |
| Article ---> Paragraph | 0.65826 | 0.0169 | 1.93129 | 0.69237 | 0.0004 | 2.99234 |
| 2-3 P ---> Paragraph | 0.02425 | 0.481 | 0.06582 | 0.05196 | 0.412 | 0.25059 |

Covid science dataset (30 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | 0.27763 | 0.1765 | 1.06224 | 0.13258 | 0.2438 | 0.72441 |
| Article ---> Paragraph | 0.69915 | 0.01805 | 1.78310 | 0.49915 | 0.0257 | 1.66534 |
| 2-3 P ---> Paragraph | 0.22988 | 0.1795 | 1.02993 | 0.02976 | 0.4411 | 0.18913 |

Covid psychology dataset (30 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | -0.63556 | 0.9678 | -0.38989 | -0.28840 | 0.9284 | -1.47024 |
| Article ---> Paragraph | 0.37749 | 0.17301 | 0.94661 | 0.43386 | 0.0024 | 1.85060 |
| 2-3 P ---> Paragraph | -0.02156 | 0.4867 | -0.06129 | -0.22712 | 0.8713 | -1.17128 |

Covid science dataset (40 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | 0.150793 | 0.3091 | 0.57597 | 0.10640 | 0.284 | 0.59211 |
| Article ---> Paragraph | 0.76665 | 0.0047 | 1.93097 | 0.53417 | 0.0091 | 1.80730 |
| 2-3 P ---> Paragraph | 0.18428 | 0.2865 | 0.67298 | 0.00396 | 0.5062 | 0.01858 |

Covid science dataset (40 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | 0.150793 | 0.3091 | 0.57597 | 0.10640 | 0.284 | 0.59211 |
| Article ---> Paragraph | 0.76665 | 0.0047 | 1.93097 | 0.53417 | 0.0091 | 1.80730 |
| 2-3 P ---> Paragraph | 0.18428 | 0.2865 | 0.67298 | 0.00396 | 0.5062 | 0.01858 |

Covid psychology dataset (40 articles)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mantel Test | 10-word network | | | 20-word network | | |
| r | p-value | Z-score | r | p-value | z-score |
| Article ---> 2-3 P | -0.48586 | 0.9388 | -1.62018 | -0.35639 | 0.9624 | -1.80561 |
| Article ---> Paragraph | 0.23856 | 0.30337 | 0.60114 | 0.51077 | 0.0013 | 1.96954 |
| 2-3 P ---> Paragraph | -0.25881 | 0.762 | -0.87218 | -0.22467 | 0.8612 | -1.11457 |

# Bibliography

Altszyler, E., Sigman, M., Ribeiro, S., Slezak, D.F., 2017. Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. Consciousness and Cognition 56, 178–187. https://doi.org/10.1016/j.concog.2017.09.004

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.

Besançon, R., Rajman, M., 2002. Evaluation of a Vector Space Similarity Measure in a Multilingual Framework.

Bhagwant, 2011. Latent Semantic Analysis (LSA) Tutorial. Personal Wiki. URL https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/ (accessed 10.18.21).

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.

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.

Campbell, M.J., 2021. Statistics at Square One. John Wiley & Sons.

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.

Databricks Academy, 2019. Introduction to Latent Semantic Analysis (1/5).

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

Gupta, A., Johnson, B., Nagar, D., 2013. Testing Equality of Several Correlation Matrices. Revista Colombiana de Estadística 36.

Hagberg, A., Swart, P., S Chult, D., 2008. Exploring network structure, dynamics, and function using networkx (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab. (LANL), Los Alamos, NM (United States).

Ibrahimov, O., Sethi, I., Dimitrova, N., 2002. The performance analysis of a Chi-square similarity measure for topic related clustering of noisy transcripts, in: Object Recognition Supported by User Interaction for Service Robots. Presented at the 16th International Conference on Pattern Recognition, IEEE Comput. Soc, Quebec City, Que., Canada, pp. 285–288. https://doi.org/10.1109/ICPR.2002.1047452

Kabir, S.M., 2016. METHODS OF DATA COLLECTION. pp. 201–275.

Khademi, A., Hosseinzadeh, D., Venetsanopoulos, A., Moody, A., 2009. Nonparametric statistical tests for exploration of correlation and nonstationarity in images, in: 2009 16th International Conference on Digital Signal Processing. Presented at the 2009 16th International Conference on Digital Signal Processing, pp. 1–6. https://doi.org/10.1109/ICDSP.2009.5201186

Kintsch, W., 2001. Predication. Cognitive Science 25, 173–202. https://doi.org/10.1207/s15516709cog2502\_1

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

Slomovitz, G., 2017. Latent Semantic Analysis (LSA) for syslog correlation, in: 2017 International Conference on Electronics, Communications and Computers (CONIELECOMP). Presented at the 2017 International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 1–4. https://doi.org/10.1109/CONIELECOMP.2017.7891819

Soundarajan, S., Eliassi-Rad, T., Gallagher, B., 2014. A Guide to Selecting a Network Similarity Method, in: Proceedings of the 2014 SIAM International Conference on Data Mining. Presented at the Proceedings of the 2014 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, pp. 1037–1045. https://doi.org/10.1137/1.9781611973440.118

Steiger, J.H., 1980. Tests for Comparing Elements of a Correlation Matrix 7.

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

-----------------------------------------------------Not Included---------------------------------------------------------------

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 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |