Relationships Between Countries United Nation General Assembly Speeches

Edward Celella

School of Computer Science

University of Birmingham

Abstract

The aim of this paper is to discover if there is a relationship between the topics in a countries United Nations General Debate speech, and three main factors: The economical and social development of a country, the geographical location of a country, and the year the speech was given. By employing latent semantic analysis to a document vector (constructed from a data set of speeches given by countries at the General Debate from 1970 to 2015), a correlation was found for all three factors. In particular, the year in which a speech was made had the largest factor on a speeches topic. When removing this variable from the data set, by sampling the speeches from one year, a correlation was found for the other two stated factors as well. Principle component analysis was also applied to the latent semantic vectors, however as both use singular valued decomposition, this yielded no useful information.

Keywords: United, Nation, General, Debate, Speeches, Topic, Relations, Principle Component Analysis, Latent Semantic Analysis

2020 MSC: Intelligent Data Analysis (Extended), 06-20233

Email address: emc918@student.bham.ac.uk (Edward Celella)

Contents

1. Introduction

The General Assembly of the United Nations (UN) is a forum consisting of all member countries, and is "one of six main organs of the UN" [4]. One of the most important meetings of the General Assembly is the annual general debate, held at the UN headquarters in New York. The meeting is the main point of political discussion for the UN, in which all members have the opportunity to convey their views on the current major global topics.

1.1. Research Questions

This study aims to find relationships between the speeches given by each country at the general debate. Specifically, the following questions will be investigated:

1. Is there a correlation between the topics discussed in a nations speech and its social and economical development?
2. Is there a relationship between the topics discussed in a nations speech and its geographical location?
3. Is there a correlation between the topics of a nations speech and the year the speech was made?

As stated the general debate is a platform for each country to express their interests and views on problems. Due to this, the hypothesis presented by this study is that there will be a strong correlation between the content of a speech, and the economical/social development of a country. This is because countries with a larger economy and standard of living, do not have to deal with same daily problems as countries without these benefits. One example of such a topic is absolute poverty.

In addition to this a strong correlation is expected to be seen between a countries speech and geographical location. This relationship is expected as the culture of a country effects the views of citizens on all topics, and countries who are closer together have similar cultures, as societies develop based on their environment [3].

Lastly, a correlation is also expected between the year a speech was made, and the topics it contains. This prediction is merely due to the fact that over the span of 45 years, the importance of certain topics changes based on the geopolitical landscape.

In order to answer these questions a variety of statistical analysis techniques will be employed in order to analyse the data. The two main tools which will be utilised are latent semantic analysis (LSA), and principle component analysis (PCA).

1.2. Data Sets

The data set used for this assignment was provided by the United Nations [6]. It contains each speech given at the general debate from the years 1970-2015. Each speeches text is given in full, along with the meeting session number, year delivered, and the ISO alpha-3 code of the country which delivered the speech. Overall there are 7507 speeches, given by 199 different countries, over a span of 45 years.

To answer the research questions stated, further data sets were required in order to provide additional information. Specifically, a data set containing the human development index of each country was obtained [1], as well as a data set containing the continent in which each country is situated [2].

2. Data Pre-processing

As outlined in section 1.2, the primary piece of data being analysed is the full speech given by each country, meaning that all the data is the form of text. Therefore, the text must be transformed into a numerical representation of itself, in order to apply the outlined statistical methods in section 1.1.

2.1. Stop Word Removal and Stemming

Before any transformation is applied to the data, the vocabulary of each speech was reduced in an attempt to reduce the dimensionality of the data. This was achieved in two ways.

Firstly, a stop word list was produced. This list contains common words which have no significance to the topic of a speech. Every instance of words in this list was removed from each speech. The list itself contained common words in the English vocabulary, as well as country names. The country names were added to this list as a majority of speeches started with representatives thanking countries by name, and introducing their country.

Secondly, the remaining text in the speeches was then stemmed. Stemming was utilised to remove word inflections, resulting in the base of the word being produced. Meaning similar words become the same token.

Punctuation, numbers, and special characters were also removed as they provide no useful information regarding topic.

2.2. Term Frequency-inverse Document Frequency (TF-IDF)

The resulting text after the processing described in section 2.1, then required transforming into a numerical representation. To achieve this, the TF-IDF technique was utilised.

TF-IDF is a commonly used tool to convert text to a numerically represented version of itself. It operates by producing a vector, in which the importance of words in a given corpus are represented. The importance of words are determined by the number of times a word appears, offset by the amount of documents each word appears in. This means if a word appears regularly in a document, however it appears irregularly throughout the corpus, then this word is a good measure of similarity. The converse of this is true, as if a word appears regularly throughout all documents then it is not a good measure of similarity.

In an effort to reduce the size of the vocabulary, only the top 5000 words were kept, which appeared in less than 80% of the documents, and more than 20%.

The result of this produced a vector for each document which contained the relative score each word in the vocabulary obtained for that document.

3. Labelling

In order to evaluate the results obtained after statistical analysis, each document required labels which could be used to answer the questions described in section 1.1.

The original data set already contained the year each speech was made, and so each speech was given label indicating the decade in which in was delivered.

Next, each speech was labelled with the ISO continent code, which was obtained using the data set described in section 1.1. The continent will be used

to analyse the geographical location of a countries effect on a speech, as a majority of continents share similar cultures.

Finally, in order to analyse the social and economical development of a countries effect on its speech, the human development index was utilised. The human development index (HDI) is a rating issued by the United Nations, which takes into account a range of variables including the average salary, education, and general life quality of a country [5]. Due to this it is a widely used indicator for a countries social and economic status. These scores can be broken down into four main categories [7]:

- Very High - A HDI of above or equal to 0.8.
- High - A HDI of above or equal to 0.7.
- Medium - A HDI of above or equal to 0.55.
- Low - A HDI of lower than 0.55.

Using the data set outlined in section 1.2, and the banding system described above, each speech was tagged with its countries corresponding HDI bracket.

3.1. Pre-processed Data Set

The data set after applying the labels described in this section, produced a dataset in the format outlined in table 1.

Table 1: Final data set.

Index	Year	Country	Continent	Decade	HDI	abil	...	zone
1	1989	MDV	AS	1980s	HI	0.028	...	0.114
2	1989	FIN	EU	1980s	VH	0.000	...	0.026
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
7507	2001	KWT	AS	2000s	VH	0.875	...	0.012

4. Latent Semantic Analysis (LSA)

4.1. Description of Analysis Method

LSA is a natural language processing technique which finds relationships between words in order to decipher topics. The general idea behind the method is that words that appear in the same document are related. Thus, by producing sets of words that appear regularly together, topics can be automatically detected.

LSA achieves this by applying singular valued decomposition to a term matrix (e.g. TF-IDF). In short, it calculates the co-variance between terms, and rotated the axis in order to reduce the co-variance. This essentially collapses two terms together, forming a relationship between them. This is done repeatedly, forming sets of words collapsing terms together with high variance.

The similarity between these generated topics and the original documents can then be deciphered using the cosine angle between the topic and the document.

4.2. Application

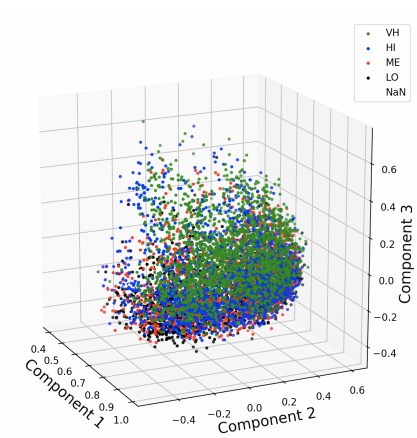
LSA was applied to the TF-IDF vector described in section 3.1. In order to reduce dimensionality as much as possible, the top 10 topics with the most variance were kept. These topics only account for 16% of the variance in the data, however this was a necessary reduction in order to evaluate the results. Table 2 shows an example of the terms used to generate topics.

Table 2: Topics generated by LSA.

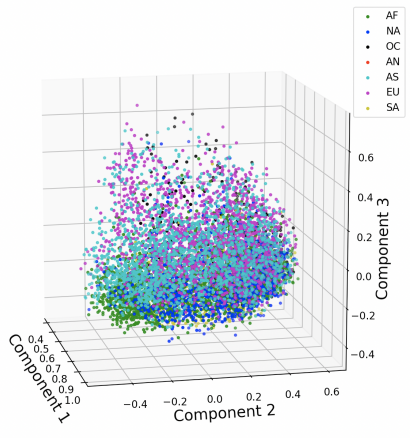
Topic	Term 1	Term 2	Term 3	Term 4	Term 5	...
1	nuclear	confer	war	problem	weapon	...
2	sustain	challeng	climat	terror	millennium	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

4.3. Results

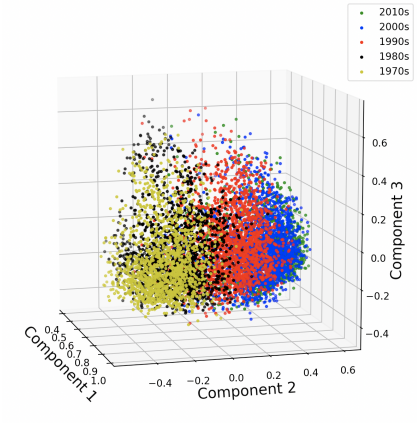
Figures 1a and 1b show each document plotted using the cosine similarity between the first 3 topics, with each point coloured by its HDI and continent location respectivly. As shown there is no correlation between these attributes and the topics of the given speech. However, when tagging each point by the decade in which it was given, a clear correlation can be seen. This is shown in figure 1c and 1d. This therefore shows that the year in which a speech was given has a drastic effect on a speeches topic.



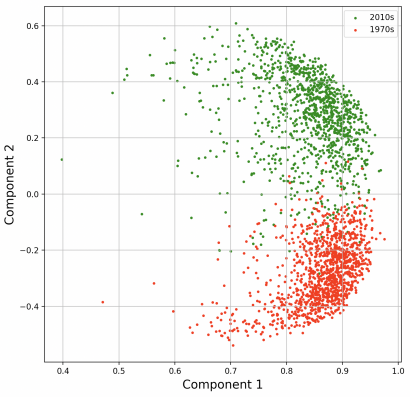
(a) LSA components labelled with HDI tags.



(b) LSA components labelled with continent tags.



(c) LSA components labelled with decade tags.



(d) Speeches made in the 2010s compared with speeches made in the 1970s

Figure 1: LSA analysis of speeches from all years.

Due to these results, the pre-processing method described in section 2.1, was applied to a subset of the data only containing speeches made in 2015. LSA was then applied to this subset in order to decipher if the effect of the year a speech was made overpowered any correlation in terms of HDI and geographical location. The results, shown below, do show correlation between a speeches topic and both the HDI of a country, as well as a countries geographical location. In order to showcase the clusters more clearly, only subsets of the tags were utilised.

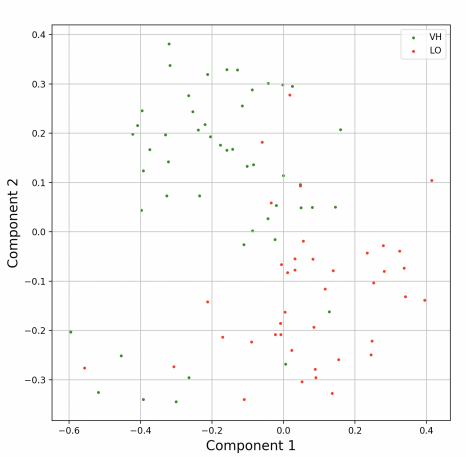


Figure 2: LSA analysis of 2015 speeches from very high and low HDI countries.

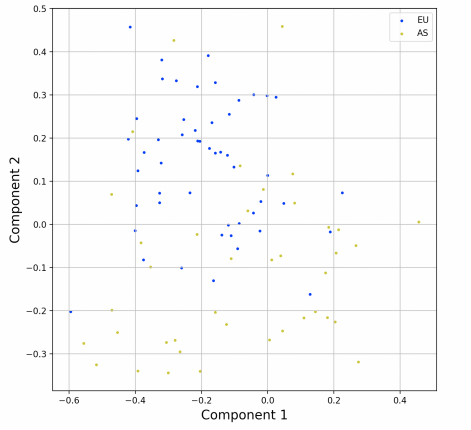


Figure 3: LSA analysis of 2015 speeches from Europe and Asia.

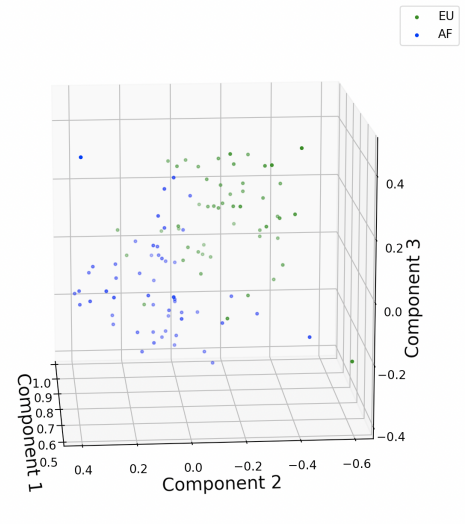


Figure 4: LSA analysis of 2015 speeches from Europe and Africa.

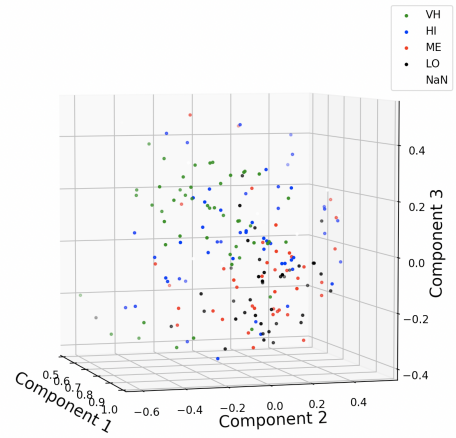


Figure 5: LSA analysis of 2015 speeches tagged with HDI brackets.

5. Principle Component Analysis (PCA)

5.1. Description of Analysis Method

PCA is another technique which utilises singular value decomposition. It can be considered to be a more generic form of LSA, whereby it to rotates the axis, in order to reduce co-variance to 0. Thus, giving a lower dimensional explanation of the data, whilst preserving information.

5.2. Application

PCA is not useful when working with document vectors such as TF-IDF. This is due to the large dimensionality of the data. Because of this PCA was applied to the vectors generated by LSA, to identify if the number of these vectors could be reduced. However, as both are forms of singular valued decomposition, the expected results of this were minimal. Furthermore it was also expected that PCA would have no impact on the correlation of the data.

Applying PCA to the 2015 subset resulted in no reduction in dimensionality. In addition to this, PCA was run on a larger subset of the LSA components, but again this resulted in minimal reduction and so was disregarded.

When applying PCA to the entire data set (with the top 10 LSA components), it managed to retain 99% of the variance in 9 components, instead of the provided 10. Figure 6 shows the variance of each component.

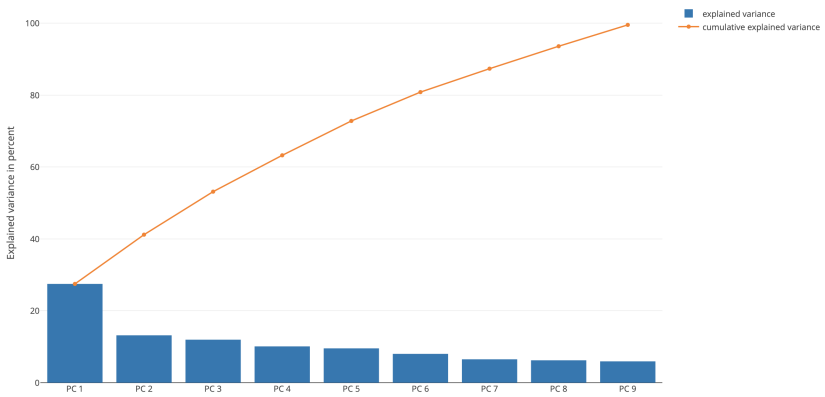
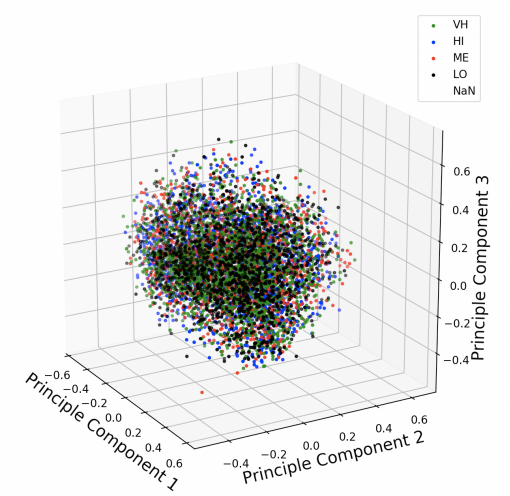


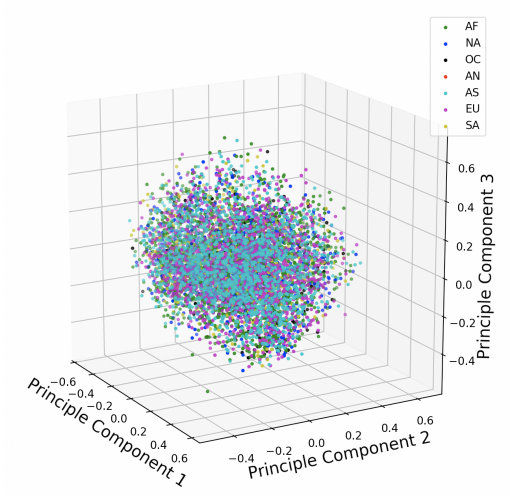
Figure 6: Scree plot of principle component variance.

5.3. Results

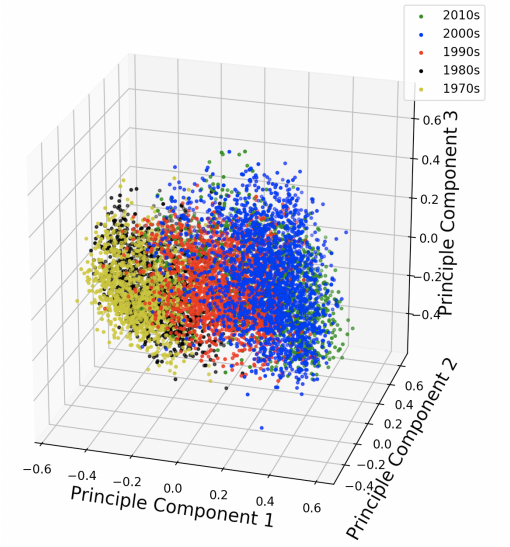
The results of PCA, as expected, had no impact on the results produced using LSA. Plotting the speeches in terms of their principle components yielded similar graphs, as shown in figure 7, except they were clustered around the center. This yields no additional analysis of the data.



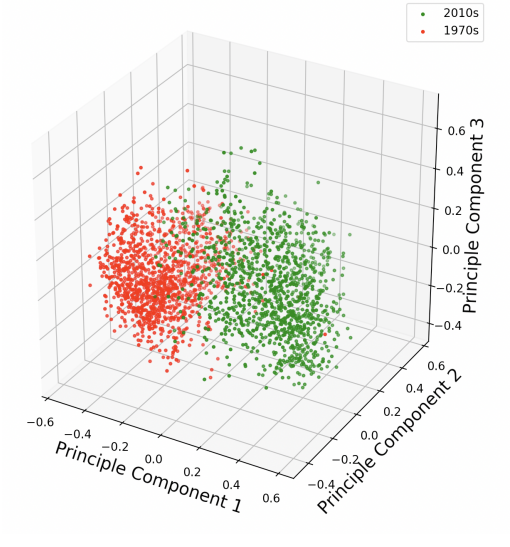
(a) PCA components labelled with HDI tags.



(b) PCA components labelled with continent tags.



(c) PCA components labelled with decade tags.



(d) Speeches made in the 2010s compared with speeches made in the 1970s

Figure 7: PCA analysis of speeches from all years.

6. Conclusion

The analysis methods of PCA and LSA are both useful tools for statistics. However, as this paper demonstrates, it is important to select methods which apply best to the given data. In this case LSA proved to be the effective tool, due to the fact that data was in text format.

The analysis given by LSA shows that the year a speech was made at the United Nations General Assembly, is the largest contributor to the topics raised in the speech. This makes logical sense as the geopolitical landscape changes over time, thus bringing to bear the importance of other topics. For example in 1970, the height of the cold war, the topics of that time were different to those in the 2015.

When analysing the speeches within a certain year, specifically 2015, LSA shows that there is infact a correlation between both a countries economical and social development, and its geographical location. This again was the expected outcome of the paper, due to the fact countries face different issues, and share similarities, based on these factors.

References

[1] Aqel, I. A., 2018. Human development index dataset.
URL https://github.com/iyadaqel/Human-Development-Index-dataset

[2] MaxMind, 2020. Iso 3166 country codes with associated continent.
URL https://dev.maxmind.com/geoip/legacy/codes/country$_c$*ontinent*/

[3] Peet, R., 2006. Modern geographical thought. Blackwell.

[4] United Nations, 2019. General assembly of the united nations.
URL https://www.un.org/en/ga/

[5] United Nations, 2019. Human development index.
URL http://hdr.undp.org/en/content/human-development-index-hdi

[6] United Nations, 2019. Un general debates.
URL https://www.kaggle.com/unitednations/un-general-debates

[7] World Population Review, 2020. Human development index (hdi) by country 2020.
URL http://worldpopulationreview.com/countries/hdi-by-country/