UN General Debates Analysis in R

Jason Kehl

MIS500 – Foundations of Data Analytics

Colorado State University – Global Campus

Dr. Dwight Davis

April 7, 2019

UN General Debates Analysis in R

Each year, the United Nations holds a session where the countries of the world may deliver a statement to the UN general assembly (United Nations, n.d.-a). Due to the social and political nature of this data set, it will likely be an interesting subject for an exercise in textual data analysis. I will analyze the UN General Debates data set utilizing techniques such as data cleansing, stop words, undesirable word filtering, data augmentation, data annotation, lemmatization, unsupervised machine learning, and visualization. The goal of the analysis is to answer several specific questions about the data set.

Data Acquisition

The United Nations organization provided the data from the General Debates and posted it on Kaggle (United Nations, 2016), so the initial data is readily available as a primary source. To facilitate the data flexibility desired, I obtained three secondary data sources for data augmentation and filtering: stop word data, country code data, and UN officials' names. I was conveniently able to find stop words and country code data directly in R packages. I obtained the UN presidents names from the UN website (United Nations, n.d.-b), as well as the names of former secretaries general (United Nations, n.d.-c).

Data Preparation Activities

I settled on 8 general data preparation steps. Steps 6-8 proved to be most difficult and required a large proportion of the data preparation time.

1. Contractions were removed using a manual replacement of common contractions using the gsub function from the base R library.
2. Special characters were removed using gsub to remove any non-text or digit.

3. “Undesirable” words were identified and compiled to form a set of words that were not useful for analysis, such as “united”, “nations”, “national”, and “council”. These were filtered out of the results whenever appropriate, since they pollute the results with non-meaningful words and phrases. Undesirable words and phrases are frequently self-referencing such as “United Nations” and “Security Council”.
4. The tidytext R package provided a useful means to filter out stop words.
5. An additional column was required to decode the year into a decade, such as 1970s, 1980s, etc. I accomplished this using manual decoding logic within the mutate R function.
6. The original UN data set only includes 3 letter codes for countries. To achieve a more easily readable display and more grouping options, I used the countrycode R package to join the iso3c code with the available country code data. This package provided a country name, region, and continent. Additionally, several countries did not appear in the data set because they are not modern countries, e.g. DDR, the Democratic Republic of Germany. The subset was small, so I manually decoded a subset to handle the extras.
7. The original text included frequent mentions of the UN Presidents and UN General Secretaries, so it was necessary to create a data set of their first and last names to filter them out. The data set was formed by copying and pasting a table of UN presidents to a csv file to import to R, then using regular expressions to separate the first and last name. I handed the UN General Secretaries similarly. I performed a manual data compilation for this data set because the list was short, and a table of information was not readily available.
8. To prepare the data set for analysis, I tokenized the text into individual words using the udpipe package (AMR, 2018). The package also included useful annotation data such as parts of speech and lemmatizations. The primary challenge for this phase was the size of the

UN data set and the processing time required for the udpipe annotation. The first attempt to annotate the entire data set took over 4 hours, which was not practical. To solve this, I employed a multiprocessing solution that split the data set into 8 parallel processes and brought the total annotation process closer to an hour. This was a huge success. I also exported the annotated conllu format data directly to file to save 90% on disk space and load time. The annotated csv file was 12 GB on disk, where the annotated conllu file was 1.2 GB.

Analysis Plan

R Studio is the tool to be used for all analysis activities. It is an effective organizational tool to document ideas, thoughts, conclusions, steps, and visualizations allowing a progression toward answering each question, and a clear path for understanding and reproducing the results. As a related note, GIT will be used as a revision control platform for this project.

To drive the analysis, I compiled a list of questions to answer during the analysis activities. I devised the following steps to facilitate a logical plan of analysis.

1. Clean and prepare the data
2. Explore each question individually. For each question, proceed with the following:
    a. Aggregate the data as appropriate for the problem statement and visualization.
    b. Begin exploring visualizations with the goal of answering the proposed question with as few visualizations as possible. In some cases, multiple visualizations will be generated if it adds interest to the topic.
    c. Note any conclusions and insight gained.
3. Machine Learning:
    a. Linear Discriminant Analysis

i. Groups the words into a specified number of themes using machine learning to contextually discover the correct theme for each word (Liske, 2018b).

ii. Create a term frequency, inverse document frequency (tf-idf) data set grouped by country and filter the data set to a known subset of countries.

iii. "Fit" the algorithm by selecting the number of topics to match the known number of countries. This allowed me to gauge whether this algorithm effectively identified the correct themes for the selected countries.

b. Rapid Automatic Keyword Extraction

i. Discovers two-word phrases based on the number of occurrences of each individual word and frequency of a co-occurrence word (AMR, 2018).

ii. Run the RAKE algorithm using noun-adjective pairings of lemmatized words.

Expectations

I expect that the top-used words and themes would change over time, and from region to region. Over-all sentiment changes over time should lead to interesting insights as well. Adding two-word phrases will likely clarify the general themes over single words and will show a clearer thematic picture. Unsupervised Machine Learning will add significant depth to the analysis by automatically categorizing the words into themes using an LDA algorithm. Data cleansing will certainly be a large part of this project, but if effective, will lead to more interesting visualizations and insights.

Analytics Questions and Analysis Observations

1. How much data preparation is required to see meaningful results?

The data cleansing process outlined in Data Preparation Activities were largely effective for most of the analysis techniques used. Data cleansing is an iterative process and can be tuned

for the data set. However, the RAKE technique used in this analysis still resulted in many proper names beyond those that were filtered, which showed that further data cleansing was necessary beyond what was planned for the analysis.

2. What are the top words by world region?

"Peace" is a highly utilized world across the globe. It could be that a phrase containing the word should be filtered, for instance if the word "peace" is found in a phrase referring to an organizational body of the UN. Further analysis would be required to ascertain the context, which is outside the scope of this analysis. "Development" is a common word for several south pacific nations.

region	word	count
Western Asia	peace	9597
Western Africa	peace	8371
Eastern Africa	peace	7329
Caribbean	development	5651
South America	peace	4849
Eastern Europe	peace	4613
Southern Asia	peace	4569
South-Eastern Asia	peace	4361
Southern Europe	peace	4015
Northern Africa	peace	3930
Middle Africa	peace	3912
Central America	peace	3784
Northern Europe	peace	3389
Western Europe	peace	3120
Eastern Asia	peace	2862
Southern Africa	africa	2092
Melanesia	development	1566
Northern America	peace	1213

Figure 1. Top words by world region (1970 – 2016)

3. What are the top words for the five Permanent Members of the UN Security Council (PMUNSC)?

Russia, China and France heavily refer to themselves. It might be good to filter out self-references. The most common words between countries are “peace” (across all), “nuclear” (US, Rus, China), “weapons” (US and Rus), and “development” (China, UK, France).

Some themes of interest can be guessed from this data, such as “human rights” for the US, “South Africa” for the UK, “European Union” for France, and “nuclear weapons” for the US, Russia, and China.

country	word	count
United States	peace	758
United States	nuclear	415
United States	human	374
United States	weapons	338
United States	rights	294

country	word	count
Russia	soviet	864
Russia	nuclear	719
Russia	union	685
Russia	peace	605
Russia	weapons	502

country	word	count
China	china	903
China	development	775
China	peace	728
China	nuclear	567
China	chinese	407

country	word	count
France	france	815
France	peace	456
France	development	343
France	europe	285
France	european	258

country	word	count
United Kingdom	peace	367
United Kingdom	africa	212
United Kingdom	conflict	201
United Kingdom	development	176
United Kingdom	south	176

Figure 2. Five top words for PMUNSC (1970 – 2016)

4. How have the top words changed over time for the PMUNSC?

“Peace” was heavily used by the US, China and Russia in the 1970s and declined for all countries since the 1990s. The word frequency of the top words has declined, possibly indicating a wider variety of topics in the modern era. The 2010s shouldn't be compared to other decades for frequency, since the data only represents 60% of the decade.

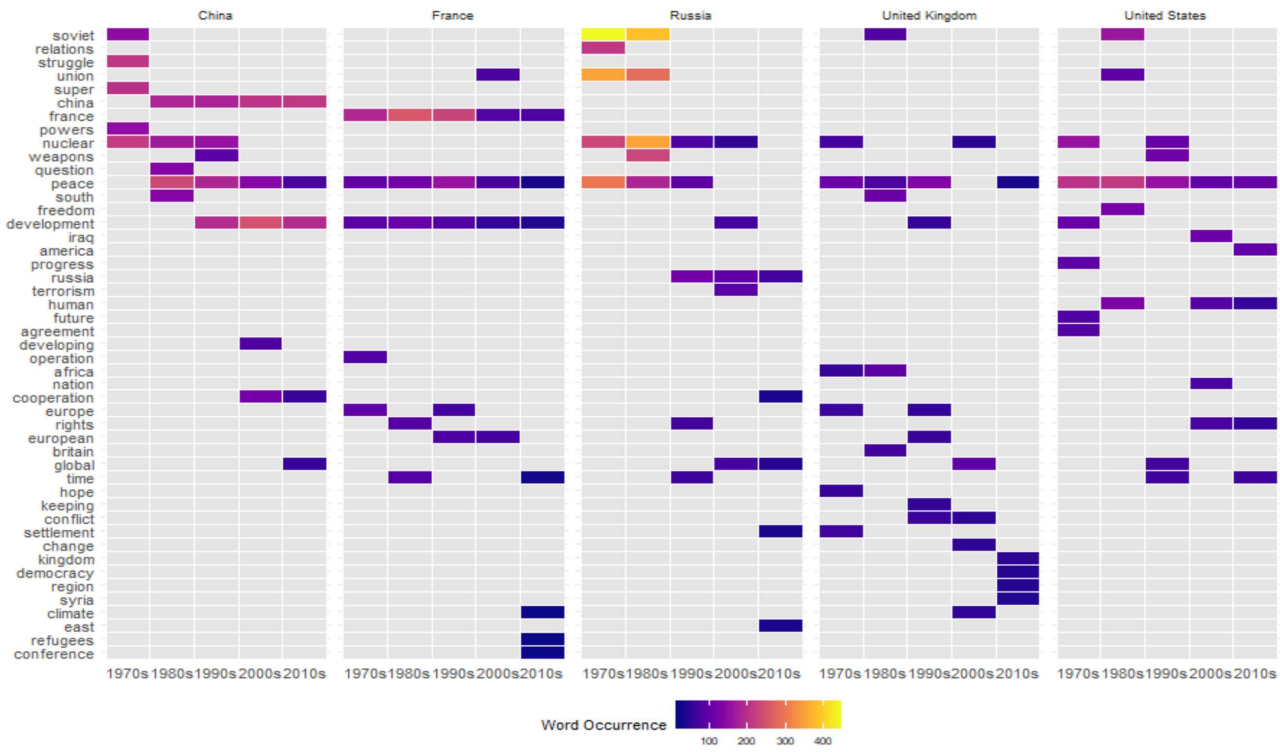

Figure 3: Top five words by decade for PMUNSC

5. How has lexical diversity changed over time?

The average number of unique words used by the world has declined since the 1970s, as shown by the pirate plot in figure 4.

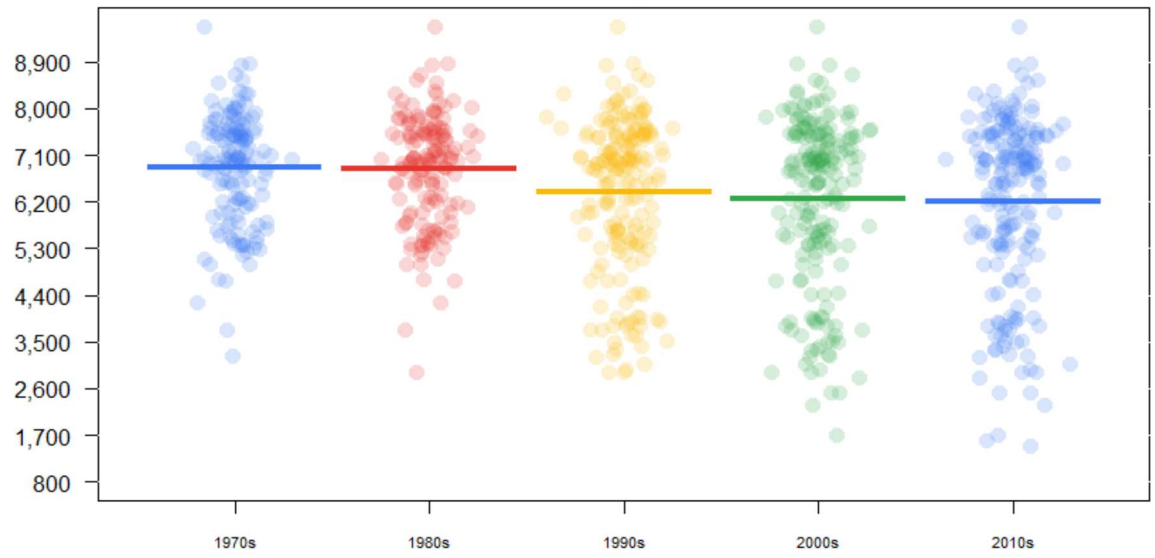

Figure 4: Lexical diversity distribution by decade

6. How has overall sentiment changed over time?

Polarity for this question is calculated using the number of positive words minus the number of negative words. A percent positive was calculated by dividing the number of positive words by the total number of words. Two distinct negative polarity periods can be clearly seen in figure 5: One begins in 1981 and ends in 1986, and the other is in 2001. The first approximately corresponds to the global timeline of the early 1980's recession. The second is aligned with 9/11.

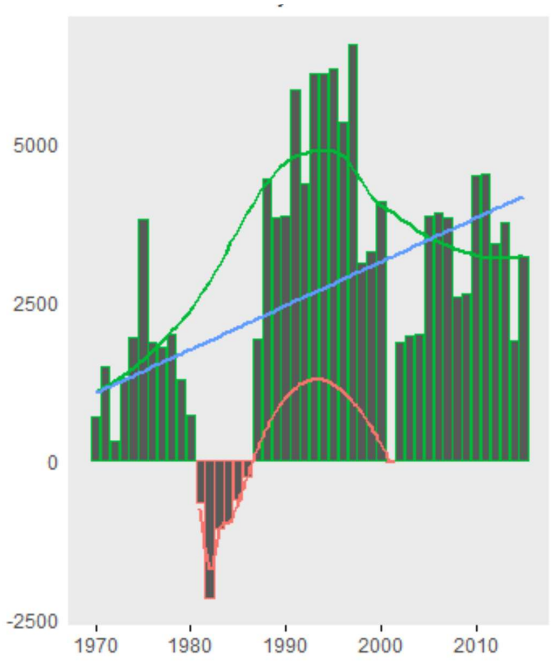

Figure 6: Polarity over time

Figure 7: Percent positive over time

7. Can sentiment gauge positive or negative mood?

Using the National Research Council Canada (NRC) emotion lexicon, a simple faceted bar chart shows mood most clearly. However, a chord chart is far more aesthetically pleasing and tells a more nuanced story.

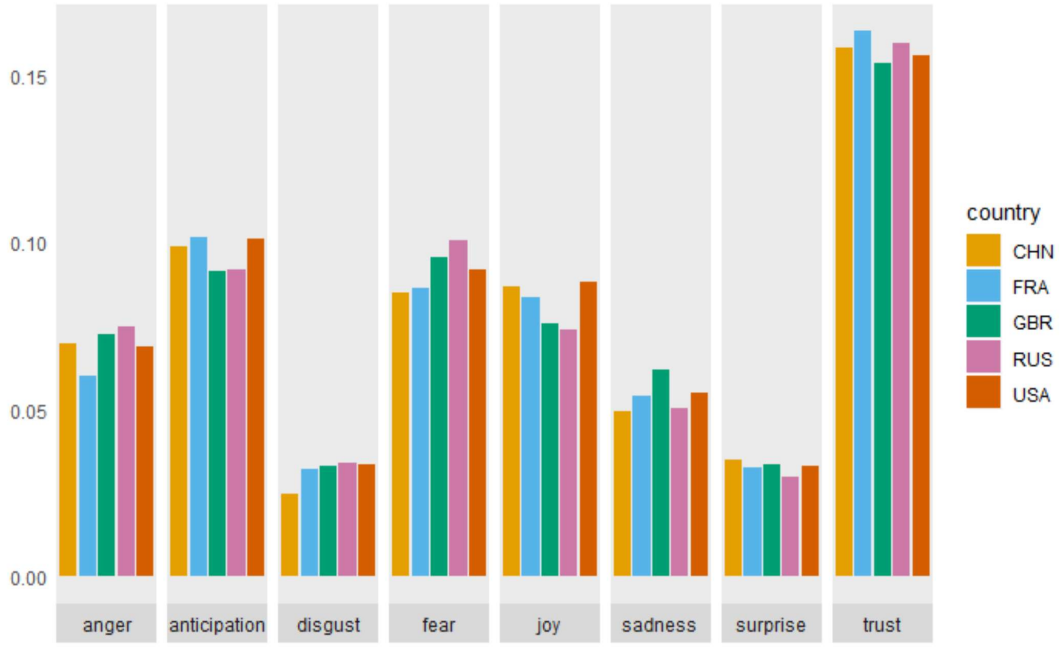

Figure 8: Sentiment of speeches by country for PMUNSC

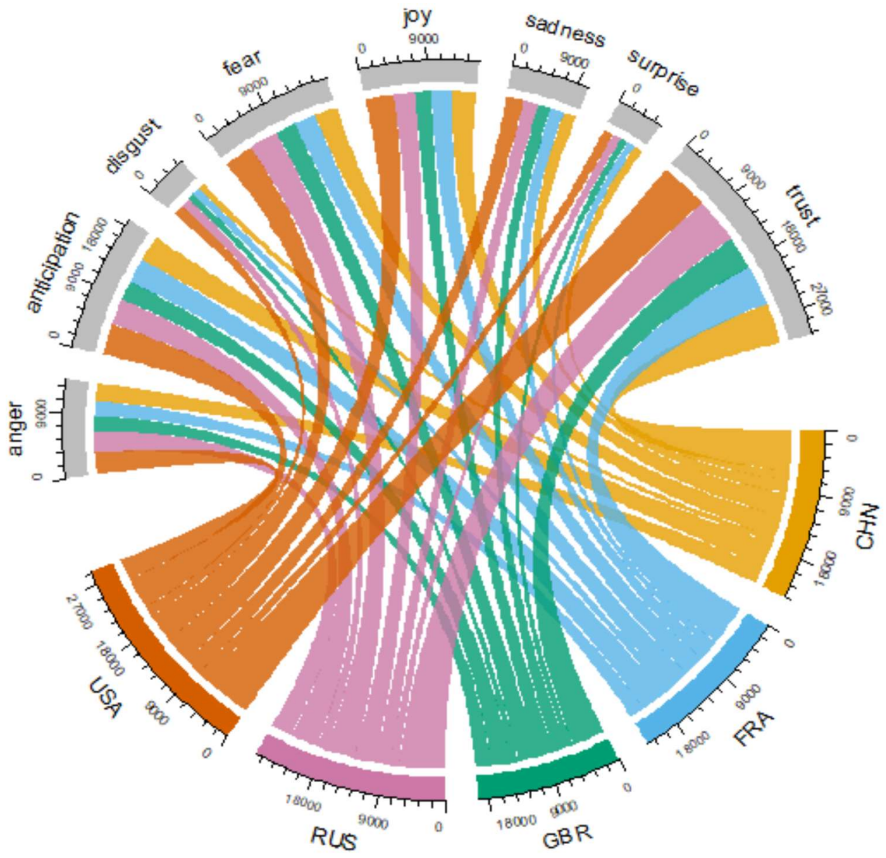

Figure 9: "Mood Ring", sentiment of speeches by country for PMUNSC

8. How does including multiple-word phrases clarify the themes?

"Human rights" is by far the #1 topic, followed by "Middle East". After that, topics begin to level out, led by "nuclear weapons". The themes are certainly more defined with multiple word phrases over single words.

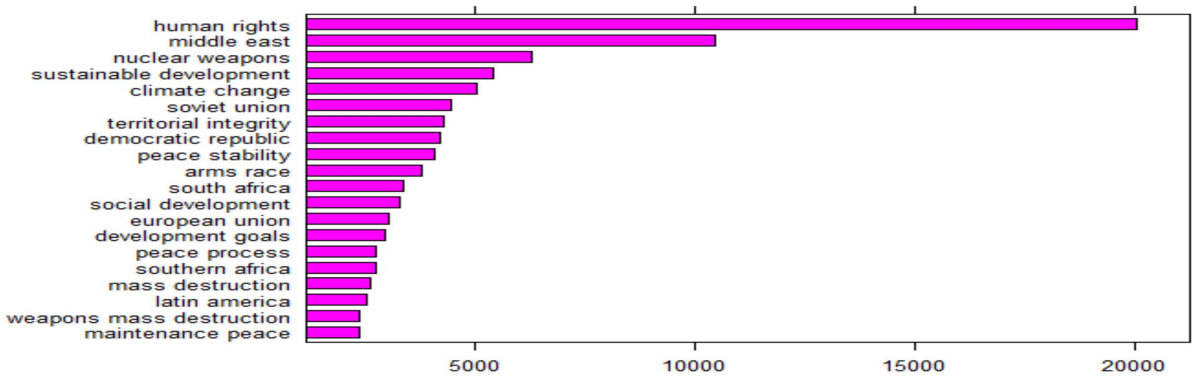

Figure 10: Multiple word noun-based phrases

9. How effective is the LDA model and RAKE for deriving themes?

LDA seems to be effective in identifying the topics of interest for each country as figure 11 illustrates. One can accurately guess which topic corresponds to which country.

Figure 11: Top words categorized into topics by LDA algorithm for PMUNSC

The chord chart shows each topic primarily dominated by a single country, and other countries have varying levels of "interest" in the topic. This chart indicates the French and China topics draw the least attention from other countries. The topic corresponding to the United Kingdom draws the most attention from other countries.

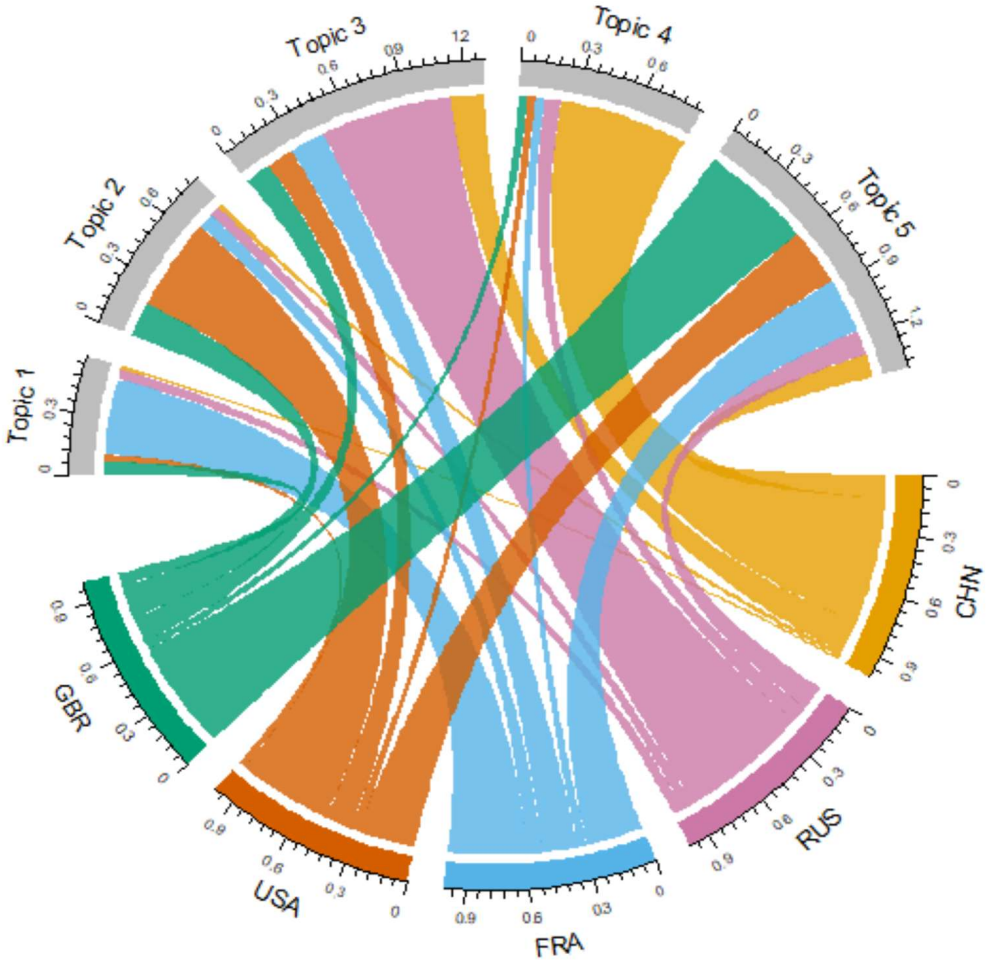

Figure 12: Topic participation by country for PMUNSC

The RAKE co-occurrence analysis interestingly exposed many proper names that are not brought to the surface by other techniques. This stands to reason since first and last names are used together frequently. A data set of historical world leaders and diplomats with locations and

dates would be useful to identify the individuals for context. Perhaps it would be best to isolate all proper names for separate analysis. That activity is beyond the scope of this project.

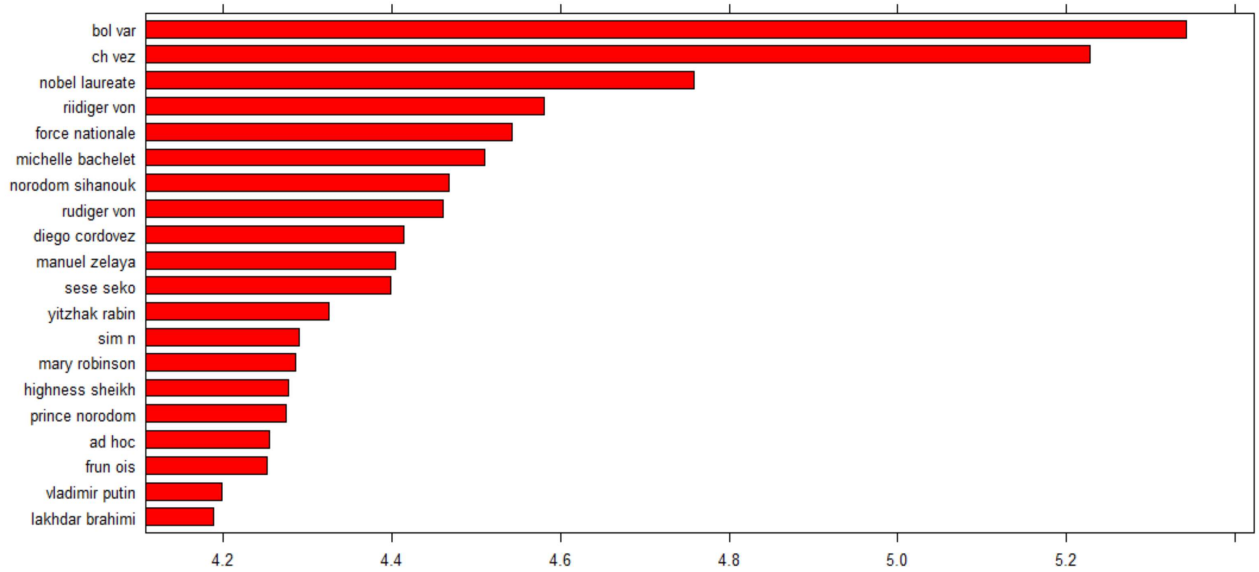

Figure 13: Two-word noun phrases using RAKE

Conclusion

The techniques of this analysis vary from simple to more complicated in nature. Each question provided value for the analysis and the means to exercise many data science techniques. Specifically, I exercised techniques for conception, planning, data collection, data processing, data cleansing, visualization, machine learning, documentation, and even source control.

References

AMR. (2018). Text analysis in R made easy with udpipe. Retrieved from https://towardsdatascience.com/easy-text-analysis-on-abc-news-headlines-b434e6e3b5b8

Liske, D. (2018a). Lyric Analysis with NLP & Machine Learning with R. Retrieved from https://www.datacamp.com/community/tutorials/R-nlp-machine-learning

Liske, D. (2018b). Machine learning and NLP using R: topic modeling and music classification. Retrieved from https://www.datacamp.com/community/tutorials/ML-NLP-lyric-analysis

United Nations. (2016). *UN general debates: Transcriptions of general debates at the UN from 1970 to 2016* [Data set]. Retrieved from https://www.kaggle.com/unitednations/un-general-debates

United Nations. (n.d.-a). General Assembly of the United Nations. Retrieved from https://gadebate.un.org/en

United Nations. (n.d.-b). Past Presidents. Retrieved from https://www.un.org/pga/73/about/past-presidents/

United Nations. (n.d.-c). Former secretaries-general. Retrieved from https://www.un.org/sg/en/content/former-secretaries-general