**ASSIGNMENT 3**

**EXPLORATION OF AN INTERESTING TEXT CORPUS**

**Introduction:**

There are a vast variety sources where data is generated in the form of text. Text data can be customer reviews, complaints, sentiments, trends etc. All this data is unstructured and to get meaningful insights we need to analyze this data. Exploration of such data makes it required to find out trends and apply this knowledge to make data driven decisions. Making comparisons is a useful way to find trends.

**Data Description:**

I have used corpus from the one available on blackboard which contains text data of politicians speaking on the topic of Iraq War Resolution. I made a choice to create a new corpus from the existing ones about this topic. This decision was made to explore the speech in writing about the same topic of two different people in order to find out their style of talking, usage of words, their opinion. It can also be used to find the trend about the topic at hand by making comparisons between their speeches. I also tried to combine speech on Risen Oil Prices.

I collected data in a .txt file to be used for this exploratory analysis. I randomly selected the topic from the text and then combined them from different documents. This made the final .txt ready to be worked on for exploration.

**Methods:**

Vectorization is a process of converting text into word vectors(numbers) to be able to count and analyze them. The text file is loaded using Python open and read functions to be read as a string of sentences. Before loading the final file, regular expressions filtering was done to remove certain words and combinations like <DOC> etc. present.

These sentences were split into vectors(words) using split function. Removal of stop words, punctuation was done on the data preparation side. Stemming was not used as the corpus was a short text meant to find interesting patterns in the text. Big corpse would need some vocab reduction technique to be applied so as to have a limited vocabulary of words.

The sentences are now vocabulary of separate words having removed stop words including a, an, the, that etc. along with default English stop words. These words are then tokenized. The various vectorization methods used here are Unigram, Bigram and Trigram.

Unigram method was used to find out which words are common in an individual’s speech. There were 10 words which came out from 29 documents(sentences). There was not much difference between the number of words based on these 3 approaches. There was a slight difference in count based on 1, 2 or 3 combinations of words. The performance remained almost same but gave few findings which were kind of interesting.

I used words taken out of the string earlier to count the frequency of words that come in the speech to understand their style of speaking. This can also be used to compare both the speeches.

The text had words such as IRAQ and Iraq, iraq etc. which were left as it is to find how they are labelled and treated. These were counted separately and not as one word.

These different vectorization methods are applied to find out what to count and how to count vectors(words) to make meaningful sense out of it. These methods tell us how the words appear in an individual’s speech or writing helpful in deciding the intricate differences between them even though they may seem same.

**Results:**

The total number of words for Unigram were 10, Bigram 9 and Trigram 9. Initially starting with unigram would enable to us to make out the differences if the combinations are changed and how many single words make sense at first before moving to others. The bag of words hardly changed. The corpus had a text which was short which lead to the shortcomings on the performance of these methods. One noticeable observation which came during the Oil text data was using Unigram oil and companies were 2 different words only but with bigram usage the word “oil companies” had a considerable count. This was one of the improvements observed by using different vectorization methods. Companies can mean other consumer ones and not necessarily oil companies. Oil and companies can be 2 different words or used together for the context. Price and prices were counted separately counted for all the 3 methods. These 2 words were counted 1 extra time for bigram and trigram methods which is an intuitive finding. The results for TFIDF were same as unigram for these words.

The most frequent words were counted to see which words based on the context such as Oil Prices, Iraq War Resolution do the speakers say the most. This can be different in writing than when they speak. For Oil, the most frequent word was relief being counted 17 times. The word frequency count is very useful as in this scenario. Relief, week, prices, oil, companies all these words are most frequent which makes anyone who has no knowledge about the speech or article’s topic to figure out what is being said.

For the Iraq War Resolution, I saw that for one person the count of word War, support, President’s was different from another. In one speech, the count of word support was higher which made me think I supports this decision and is more peace-loving person. Another person’s usage of War being high made me read the text again to confirm that he didn’t support the idea and the count frequency count of support was very low. Furthermore, I was able to track the behavioral changes or the different styles of speaking of 2 people. This frequency can also be used for a particular to distinguish his way of speaking in public. It can also be used to make distinctions such as very positive, negative, criticizing based on their reactions and styles.

All this was helpful to find the trends in a text which includes most frequent(trending), see the least trending things based on combination of words (NGrams). It also helps to make comparisons about people are based on their speech and writing. We can tell in advance based on this analysis if this is a same person we know. Patterns in word frequency, usage can be quantified to find out the trending things based on the topic.

**Conclusion:**

Comparisons reveal certain patterns helpful to predict based on how much a person can deviate from their individual style. Patterns can also be based on least frequency of the word. Vectorization based on different methods brought up some different aspects which were not possible if using only unigram. Big text corpus would be very helpful to make these deductions out of the text. Even more helpful would be to have based on a topic to be able to make comparisons. Corpus’s based on various persons are useful to segregate two different people based on their thinking which is reflected in writing or speech. It is better to use a good amount of text to be able to capture these nuances which make a great deal of difference. Here the interesting part was comparing 2 people on how they react on a topic, the words they would use to identify them and find the trending topics or trends in a speech.