**CSCE 771 Final Project**

**Background -**

We live in a society in which the use of Artificial Intelligence has greatly improved efficiency of multiple processes, reading and interpreting text being one of them. Automatic text summarization is a field within Natural Language Processing, which can enable the user to generate machine based summaries of various text pieces like news feeds, articles, food blogs, and even research papers. There are 2 types of automatic text summarization, i.e., extractive text summarization and abstractive text summarization. Extractive text summarization is a means to summarize text by selecting the most important phrases or sentences from the text piece itself. The summary is created by combining all the selected sentences. Every line and word in the machine generated summary in this case comes directly from the text piece without paraphrasing it. Abstractive based summarization on the other hand generates the summary based on new phrases and sentences, essentially generating a paraphrased summary of the original text piece.

**Problem -**

The need to summarize text automatically is to save time and provide people who are not proficient in a particular field with enough insight and the key points to make them understand the crux of the text piece without reading it completely. My motivation for this project is due to the pace at which the user is able to consume information from a summary rather than going through the whole text piece. There have been times where I wished to read the summary of a text piece to save time and still be able to get all the key points from the piece itself, and I was forced to read through the whole piece due to the lack of a summary. I plan to use this project for certain types of text pieces to make mine and other peoples’ lives easier and save time. The expected output from this program would be a relatively short text piece derived from the main text piece, which will provide the user with key points and important information through Natural Language Processing.

**Approach / Solution -**

For the purpose of this project, I have focused on extractive text summarization, due to it being a shorter version of the original text with exact words, sentences and phrases. I have used various libraries like os, spacy, string and csv for this project.

The os library is used to check if the file exists in the current directory, as the input would be generated by a text file. The spacy library is used for it’s in-built functionality of stopwords, which are basically the most commonly used in a given language. We use stopwords for this project in order to focus on the important words instead. For instance, when using a search engine, we enter a phrase like, “where can I find a chocolate cake?” the search engine shows us with a bunch of important links to our proposed question in the phrase. It uses stopwords to filter out common words like “where”, “can”, “I” and “a” and focuses on the keywords which are “find”, “chocolate”, “cake” which helps the search engine to find links related to these keywords. I used stopwords in this project to calculate which sentences are the most important ones.

The string library is used for its in-built functionality of punctuations as the stopwords do not contain punctuations, and if the punctuations are counted, then the most important sentences can change based on what punctuations have the maximum occurrence. Finally, the csv library was used to write the machine generated summary and the actual summary from the dataset to a csv file, which is used later on to evaluate the summaries. The project is user friendly and asks the user for a text file input and the percentage of the original text piece the user wants to see in his summary. For instance, if a user specifies 20, than the summary will be 20% of the original text piece’s length.

The approach for this project is to get a text file and percentage input from the user, clean the text using stopwords and punctuations, tokenize the sentences and then tokenize the words. Create a word and it’s weighted frequency’s dictionary (weighted frequency is found by dividing each words total frequency by the frequency of the maximum occurring word). Give the sentences a score based on the above processes and convert the percentage into number of sentences to then select the calculated number of sentences based on the score of each one.

**Demonstration -**

**Following is the text document:**

Ink helps drive democracy in Asia

The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisible ink and ultraviolet readers in the country's elections as part of a drive to prevent multiple voting.

This new technology is causing both worries and guarded optimism among different sectors of the population. In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President, Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections. The US government agreed to fund all expenses associated with this decision.

The Kyrgyz Republic is seen by many experts as backsliding from the high point it reached in the mid-1990s with a hastily pushed through referendum in 2003, reducing the legislative branch to one chamber with 75 deputies. The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy, the Soros Foundation and the Kyrgyz government have all contributed to purchase transparent ballot boxes.

The actual technology behind the ink is not that complicated. The ink is sprayed on a person's left thumb. It dries and is not visible under normal light.

However, the presence of ultraviolet light (of the kind used to verify money) causes the ink to glow with a neon yellow light. At the entrance to each polling station, one election official will scan voter's fingers with UV lamp before allowing them to enter, and every voter will have his/her left thumb sprayed with ink before receiving the ballot. If the ink shows under the UV light the voter will not be allowed to enter the polling station. Likewise, any voter who refuses to be inked will not receive the ballot. These elections are assuming even greater significance because of two large factors - the upcoming parliamentary elections are a prelude to a potentially regime changing presidential election in the Autumn as well as the echo of recent elections in other former Soviet Republics, notably Ukraine and Georgia. The use of ink has been controversial - especially among groups perceived to be pro-government.

Widely circulated articles compared the use of ink to the rural practice of marking sheep - a still common metaphor in this primarily agricultural society.

The author of one such article began a petition drive against the use of the ink. The greatest part of the opposition to ink has often been sheer ignorance. Local newspapers have carried stories that the ink is harmful, radioactive or even that the ultraviolet readers may cause health problems. Others, such as the aggressively middle of the road, Coalition of Non-governmental Organizations, have lauded the move as an important step forward. This type of ink has been used in many elections in the world, in countries as varied as Serbia, South Africa, Indonesia and Turkey. The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed, improper use of this type of ink can cause additional problems. The use of "invisible" ink is not without its own problems. In most elections, numerous rumors have spread about it.

In Serbia, for example, both Christian and Islamic leaders assured their populations that its use was not contrary to religion. Other rumours are associated with how to remove the ink - various soft drinks, solvents and cleaning products are put forward. However, in reality, the ink is very effective at getting under the cuticle of the thumb and difficult to wash off. The ink stays on the finger for at least 72 hours and for up to a week. The use of ink and readers by itself is not a panacea for election ills. The passage of the inking law is, nevertheless, a clear step forward towards free and fair elections." The country's widely watched parliamentary elections are scheduled for 27 February.

David Mikosz works for the IFES, an international, non-profit organisation that supports the building of democratic societies.

**Following is the output (summary) generated by the project:**

The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed, improper use of this type of ink can cause additional problems.Ink helps drive democracy in Asia

The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisible ink and ultraviolet readers in the country's elections as part of a drive to prevent multiple voting.

The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy, the Soros Foundation and the Kyrgyz government have all contributed to purchase transparent ballot boxes.

In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President, Askar Akaev, pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections.At the entrance to each polling station, one election official will scan voter's fingers with UV lamp before allowing them to enter, and every voter will have his/her left thumb sprayed with ink before receiving the ballot.These elections are assuming even greater significance because of two large factors - the upcoming parliamentary elections are a prelude to a potentially regime changing presidential election in the Autumn as well as the echo of recent elections in other former Soviet Republics, notably Ukraine and Georgia.Widely circulated articles compared the use of ink to the rural practice of marking sheep - a still common metaphor in this primarily agricultural society.

The use of ink has been controversial - especially among groups perceived to be pro-government.

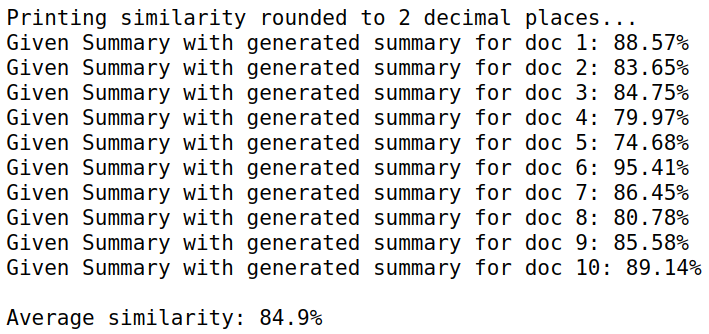
However, the presence of ultraviolet light (of the kind used to verify money) causes the ink to glow with a neon yellow light.

**Evaluation -**

I have used 2 methods for evaluating this project, i.e., evaluate using cosine similarity and evaluate using human beings.

**Evaluating based on cosine similarity:**

Cosine similarity calculates the angle between 2 vectors in a cartesian plane and we then take the cosine of that angle to find the similarity between 2 documents. Following is a snapshot taken from the “similarity\_checker.ipynb” file:



**Evaluating based on human beings:**

I provided 4 people with the original text file (not summarized), my project’s machine generated extractive summary and the text summary of the associated document from the dataset. They were then asked to read through the original document and the 2 summaries to answer some of the questions I prepared for their interview.

**Following are the questions prepared for the interview:**

1. Can you tell us what your name is?

2. Can you tell us a little bit about yourself?

3. Do you testify that you will answer all the following questions based on your own judgment?

4. Do you think that the machine generated summary accurately represents whatever is in the original document?

5. Do you like one of the 2 summaries more, if so, please give the reason?

6. On a scale of 0-100 please rate the summary, with 0 being not at all living up-to your expectations and 100 being succeeding or meeting your expectations?

**Following are the details about each candidate’s choice of the articles from the dataset:**

**Candidate #1**

Name of the candidate: Jueling Chen

Category of Document: Business

Title of the document: Nigeria to boost cocoa production

Percentage of summary: 20

Rating: 100

**Candidate #2**

Name of the candidate: Hriday Patel

Category of Document: Tech

Title of the document: PlayStation 3 chip to be unveiled

Percentage of summary: 30

Rating: 90

**Candidate #3**

Name of the candidate: Mikka Maja

Category of Document: Entertainment

Title of the document: Hillbillies singer Scoggins dies

Percentage of summary: 40

Rating: 70

**Candidate #4**

Name of the candidate: Ian Villavicencio

Category of Document: Sport

Title of the document: Barbarians 19-47 New Zealand

Percentage of summary: 30

Rating: 85

**Average rating of the candidates: 86.25%**

**Following is the link to a google drive containing the interviews:**

<https://drive.google.com/drive/folders/1DAQxYwrw4IhkqqNgnWqOx0dao38EBXrN?usp=sharing>

**Discussion -**

The project was evaluated using 30% of the original text as the summary, using the 2 methods discussed above in the evaluation section. The results were found to be approximately 85% accurate, which is in-fact very close to the human evaluation rating, i.e., 86.25%.

Candidate #1, Candidate #2 and Candidate #4 were very satisfied with the summaries that the project generated, but Candidate #3 (Mikka Maja) was not very satisfied with the results. The reason was that the selected percentage on which the text piece would be summarized proved to be inefficient. We later on increased the percentage of the same document, and she was very satisfied with the results giving a rating of 85, which boosts the average ratings of the candidates to 91.25%.

This shows us that choosing the percentage for the summary is co-related to the candidates’ satisfaction and it is the most crucial step when generating automatic text summaries.

In conclusion, I created an automatic text summarizer using Natural Language Processing techniques to solve the problem of spending vasts amount of time on reading articles, by summarizing the text piece and providing the users with key points in the article. The project has proven to be successful and was well liked by all the candidates, with mostly high levels of satisfaction. I also took part in evaluating a few articles, and I found the results to be mostly highly satisfactory as well.

I hope that I can work on this in the future to better analyze the needs of each and every category to predetermine the percentage of summary and produce slightly more satisfactory results.