**Final Project**

**Report**

Shraddha Khadepatil

Varun Kadian

 Prof. Sajeeva Salgadoe

AIDI – 2004– 02

April 06, 2022

**Introduction**

Presently, the enormous volume of data is expanding quickly especially in the digital format. It very well may be organized information like data sets, organization heritage information, or unstructured information like text, pictures and so forth. Around 85 and 90% of information is held in unstructured manner. Hence, text mining is important for extricating and overseeing helpful data from unstructured arrangements of information, for example, news reports, messages, and pages, utilizing a different message mining method. Thus, text mining has turned into a significant and dynamic exploration field. It is notable that text mining strategies have generally been produced for the English language because most electronic information is in English (though our method works in more than 30 languages). Utilizing this for our potential benefit, it is an undeniable following stage to utilize these methods for filtering through the large number of accessible on-line information to mine statistical data points from different sources and afterward sum up them productively to use in different occasions. In this research, the data extraction of news stories in view of NLP procedures to sum up the text. The outline interaction includes separating, featuring, and arranging data, which is compact, cognizant and dedicated to the news article. The critical assignment in our research is as per the following,

1. Scrape on-line articles from news sites considering a given URL.
2. Partition whole articles collectively of sentences, which goes about as the dataset for additional handling.
3. Addressing sentences in a machine coherent and justifiable configuration.
4. Recognizing semantic complications between sentences to rule out repetitiveness in end-result.
5. Bunching comparative sentences to recognize semantically various sentences.
6. Picking sentences among bunches which are much more relevant.
7. Organizing the sentences sequentially to show the summarized article along.
8. Performing a sentimental analysis on the summed-up article to show its nature for example extremity and subjectivity.
9. Developing a website page to show the outline of the article utilizing NLP fueled libraries and the opinions related with the article.
10. Deploying the model and webpage on the server so it can be accessed from anywhere.

**Scope**

The general extent of the venture is to have a more profound information on the strategies in Machine Learning, Deep Learning and Data Analysis to produce compact synopsis of news stories, which allows a client to see a rundown given a news story or of the most recent news (provided the article link). The degree additionally includes comprehension of why RNNs are fruitful at expressing sentences and how they treat some words more significant than the others by relegating the fitting weights to the words. The extent of this venture can be reached out by creating sentimental analysis of those short rundowns/summaries. Some further comparison with state-of-the models and the best performance benchmark ROUGE results are left over the future scope of this research.

**Methodology**

A pretrained library is used to scrape and summarize the news article and then sentimental analysis is performed on the summarized text in order to better understand the nature of the article (whether the article is positive or negative in nature or the article is subjective or objective).

Newspaper3k is a library utilized for web scraping of news articles. It uses the solicitations library and has BeautifulSoup as a reliance while it goes through the lxml. This library scratches the whole article text but can likewise scratch for different sorts of information, for example, the publishing date, authors, URL, and pictures. If we wish to just know what's going on with the article without perusing the entire article, Newspaper3k can likewise deliver a rundown of the article.

After gathering the news, it can then be incorporated and saved into various arrangements like CSV, JSON, and even pandas. Newspaper3k additionally works in more than 30 dialects like Arabic, Dutch, English, German, French, Spanish, Russian and many more. The Newspaper3k Python library can likewise accomplish further developed capacities, for example, finding RSS channels, scratching for article URLs from a principle news source, and even multi-string extraction in the event that there is a need to scratch for more than one article.

To perform the sentimental analysis on the text, TextBlob library is used. It is a library worked for Natural Language Processing (NLP) and it is based upon Natural Language Tool Kit (NLTK) to accomplish most of its assignments. Numerous different undertakings can be achieved utilizing TextBlob for example Message Classification and Language Translation and so on, however we will be just zeroing in on sentimental analysis in our project.

Whenever we compute the opinion of a text through TextBlob, it gives us numeric qualities to extremity (polarity) and subjectivity. The numeric incentive for extremity depicts how much a text is negative or positive. Additionally, subjectivity portrays how much a text is unbiased or abstract.

These vocabularies are alluded to in "en-sentiment.xml". Looking at the vocabularies record, we can see that it doesn't contain any stop words i.e., the, he, have, and so on the grounds that they have no opinion. Other than that, each word is characterized in the vocabulary document with their grammatical form, extremity, subjectivity, intensity, and confidence. While ascertaining an opinion for a solitary word, TextBlob utilizes the "averaging" procedure that is applied on upsides of extremity to figure an extremity score for a solitary word and thus comparable activity applies to each word and we get a single polarity for longer texts. There is an exceptionally intriguing thing about TextBlob that it handles the modifiers which intensifies the significance of message as per its example. At the point when a modifier word is utilized, TextBlob will overlook extremity and subjectivity and simply use power to process the sentiment of the text. The NLP model is developed using Python Programming Language where we used Newspaper3k library. We have used Flask for developing web application. We used Heroku for model deployment

Text

Description automatically generated

**Results**

This work presents a framework to naturally gather, examine and sum up internet based news articles in light of a client submitted query. The dataset used for summarization is ad-hoc and is generated on-the-fly. Using pre-trained libraries like newspaper3k and textblob, the pre-processing and scraping becomes much easier and quick. Results of the model were shown on the web application as summarized text along with it’s sentiments. The sentimental analysis was done on the polarity and subjectivity of the article, making users more aware towards the nature of the news.

The model was performed on more than 50 news stories to survey its credibility. The articles were looked over CBC, CTV News and BBC and were distributed during the period January, 2022 - April, 2022. The article length changes from 1000 words to 7000 words, with a normal length of 150 sentences and 3047 words. The length of the rundowns were 5% of the length of the published articles.

The summary of a test article is as follows:

'A win for the NDP’s base but in the eyes of most Canadians, it’s Trudeau who’ll get the credit. One thing is certain, after the presentation of Steven Guilbeault’s detailed plan, there will be lots of money for the promised fight against climate change. I attended his nomination meeting and was surprised by the stark partisan tone he struck in defending the Liberals’ record on climate. Trudeau broke a key election promise with the new Liberal climate plan. Tom Mulcair was the leader of the federal New Democratic Party of Canada between 2012 and 2017.'

**User Interface**

A picture containing text

Description automatically generated

**Text

Description automatically generated**

**Future Scope**

The future work extended would be to include different media types like image, video audio etc., to try different and more advanced density based clustering techniques or deep learning neural networks to get better results, to employ various parallel programming techniques, multi-threaded approaches to increase the speed of summary generation. The proposed model can summarize articles in over 30 language but the tests were done only on English news articles, further testing in other languages will help in making the model robust and versatile. The current website only boasts a single algorithm while further development of the framework is planned to have at least 3 algorithms to choose from, where user will have the option to select the type of algorithm to be used for summarizing the news article.

Apart from an interactive website, comparison of our model with other state-of-the-art model will be carried out at a later stage. Evaluation ROUGE metrics will also be introduced along with summaries and sentiments to better understand the model’s performance. ROUGE tests are considered as the benchmarks for text summarization and will act as a self-assess metric for our model.

For future development, we can also add sentimental analysis to the text getting summarised. We can add polarity to the sentiment of the text as 0 and 1 for negative and positive sentiment respectively.

**References**

Fahad, A. H. (2021, December 30). *Sentiment Analysis — Let TextBlob do all the Work! - Red Buffer*. Medium. https://medium.com/red-buffer/sentiment-analysis-let-textblob-do-all-the-work-9927d803d137

freeCodeCamp.org. (2017, October 24). *An intro to ROUGE, and how to use it to evaluate summaries*. https://www.freecodecamp.org/news/what-is-rouge-and-how-it-works-for-evaluation-of-summaries-e059fb8ac840/

GeeksforGeeks. (2021, November 30). *Newspaper: Article scraping & curation (Python)*. https://www.geeksforgeeks.org/newspaper-article-scraping-curation-python/

Helstowski, C. (2021). *Newspaper3k – A Python Library For Fast Web Scraping*. Finxter. https://blog.finxter.com/newspaper3k-a-python-library-for-fast-web-scraping/

Hristozova, N. (2022, March 30). *To ROUGE or not to ROUGE? - Towards Data Science*. Medium. https://towardsdatascience.com/to-rouge-or-not-to-rouge-6a5f3552ea45

*Newspaper3k: Article scraping & curation — newspaper 0.0.2 documentation*. (2021). . https://newspaper.readthedocs.io/en/latest/

P. (2022, March 12). *Web Scraping for News Articles using Python*. ProxyScrape. https://proxyscrape.com/blog/web-scraping-for-news-articles-using-python/

Radityohutomo, Y. (2018, April 13). *Sentiment Analysis Using TextBlob - Yora Radityohutomo*. Medium. https://medium.com/@yoratyo/sentiment-analysis-using-textblob-c57728ded1ec

Shah, P. (2021, December 15). *Sentiment Analysis using TextBlob - Towards Data Science*. Medium. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524