Abstract:

In the Modern Era online newspaper has increased tremendously, making as much sense of news data is an interesting resource for many research disciplines. From the source like Times of India; a news related article is collected and analyzed. So, I used a news-please [1] an open-source crawler and extractor of news. The machine learning model using python and NLTK are used to extract the features from the corpus. The work documented in this paper is related to web scraping and Feature Extraction from text [2].

Introduction and Background:

For researchers domain news article is an interesting data source. The process gathering news data typically consist of two phases: (1) scraping news URLs websites and (2) extracting information from news articles.

Scraping news URLs websites can be achieved by using web scraping frameworks such as Beautiful-Soup for python which create the soup object.

Extracting information from news articles a news-please [1] is used to crawler the news data from news URLs.

After collection of the necessary amount of data from the news website, records containing more null values are deleted. From this dataset news articles (collected from TimesofIndia) creates a document-term matrix from the corpus. Usually, these texts are composed of many uninformative words such as ‘a’, ‘is’, ‘the’ etc. To eliminate such kind, they were pre-processed before supplying them to the Bag-of-Words model.

Text in every news article is a composition of both informative and uninformative chunks of data. Uninformative chunks of data include stop-words (‘the’, ‘is’, ‘are’, etc.), digits, punctuations, named-entities etc. They are used to add order, meaning and to serve grammatical purposes while reading and understanding the text. But they don’t necessarily present the main idea in the text. Informative chunks are more descriptive and presents the main idea in the text. To extract informative words, these uninformative chunks need to be eliminated. Punctuations were replaced with empty spaces and stop-words were removed using stop-word removal in NLP. The acquired text was divided into tokens and processed individually. Using pos\_tagging in NLP, named-entities were identified from the tokens and removed. On the rest of the tokens, NLP techniques like lemmatization and stemming were applied to convert each token (word) into its respective base form. In the end, the processed tokens were joined to form a filtered text filled with informative words. This method was applied to every article to convert them into their respective filtered texts. Then each filtered text was converted into a vector form and supplied as an input to the model.

References:

[1] F. Hamborg, N. Meuschke, C. Breitinger, and B. Gipp, “news-please: A Generic News Crawler and Extractor,” in *Proceedings of the 15th International Symposium of Information Science*, 2017.

[2] Andrew Hintermeier, “Feature Extraction from Text” in https://andhint.github.io/machine-learning/nlp/Feature-Extraction-From-Text/