

**News Article Text Summarisation**

**Capstone Project Report**

**2018-2019**

**Course: PGP in Data Science**

**March-July 2019**

**DF-1809CM**

**Falgunee Warutkar**

**Prachi Bari**

**Prachi Ghadge**

# Abstract

**Text Summarization has always been an area of active interest in the academia. In recent times, even though several techniques have being developed for automatic text summarization, efficiency is still a concern. Given the increase in size and number of documents available online, an efficient automatic news summarizer is the need of the hour. In this paper, we propose a technique of text summarization which focuses on the problem of identifying the most important portions of the text and producing coherent summaries. In our methodology, we do not require full semantic interpretation of the text, instead we create a summary using a model of topic progression in the text. In this report we document our observation as we explore and build various nlp and deep learning models to generate summary of news articles.**

**Acknowledgement**

**We take this opportunity to thank our Project guide Prof.Vinay Kulkarni for their valuable guidance and Aegis School of Data Science for providing the required facilities, Internet access and important books, which were indispensable in the completion of this Project . I am also thankful to all the staff members of the Department for their valuable time, support, comments, suggestions and persuasion.**

**Falgunee Warutkar**

**Prachi Bari**

**Prachi Ghadge**

**Introduction**

[Automatic summarization](https://en.wikipedia.org/wiki/Automatic_summarization) aims to produce a shorter version of an input text, preserving only the essential information.Text summarization is an area with much promise in today’s age of information overflow. In the domain of scientific literature, the rate of publications grows exponentially (Hunter and Cohen, 2006), which calls for efficient automatic summarization tools.

**Need**

Text summarization is growing as sub – branch of NLP as the demand for compressive, meaningful, abstract of topic due to large amount of information available on net. Precise information helps to search more effectively and efficiently. Thus text summarization is need and used by business analyst, marketing executive, development, researchers, government organizations, students and teachers also. It is seen that executive requires summarization so that in a limited time required information can be processed.

**Basic Concept**

This paper takes into all about the details of both the extractive and abstractive approaches along with the techniques used, its performance achieved, along with advantages and disadvantages of each approach. Extractive summarizations extract important sentences or phrases from the original documents and group them to produce a summary without changing the original text. An extractive text summarization system is proposed based on POS tagging by considering Hidden Markov Model using corpus to extract important phrases to build as a summary . Abstractive summarization consists of understanding the source text by using linguistic method to interpret and examine the text. Abstractive methods need a deeper analysis of the text. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate

**Application**

One challenge for data-driven summarization is the lack of data: currently, there are only a few large, high-quality collections of articles and summaries. In particular, datasets consisting of multi-sentence summaries are scarce. Due to this, the focus in the research has mostly been on short summarization of news articles, such as to generate a headline.

In the paper, we investigate the performance of several different automatic summarization methods on the datasets. We test extractive methods based on word embeddings, as well as abstractive methods based on neural machine translation. One of our findings is that while the abstractive methods perform decently on headline generation, they struggle with abstract generation of articles. Generation of long, coherent, multi-sentence text is a big current challenge in NLP.

**Problem statement**

In this new era,where tremendous information is available on the Internet,it is most important to provide the improved mechanism to extract the information quickly and most efficiently . It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the Internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it.In order to solve the above two problems, the automatic text summarization is very much necessary.

Automatic text summarization is an information access technique used to present only the most important information from a set of documents, thereby reducing the need to look into the actual documents.

In this project we aim to reduce the problem of information overload and redundancy by automatically summarizing text conversations and providing an efficient summary to the end user. The goal is to take an article and generate a short and concise summary that can be read in lieu of the original conversation. we have mainly focused on development of extractive methods of summarization with sentences as the basic units of summary. Therefore, the problem can be transformed to “the selection of the most informative sentences from the piece of text which best represents the document”.

**Scope**

The rate at which the information is growing is tremendous. Hence it is very important to build a multilingual summarization system.Through the study it is also observed that very less work is done using abstractive methods on Indian languages, there is a lot of scope for exploring.

New metrics can be investigated which can be used in automatic evaluation environment to measure the overall quality such as grammar, readability, prominence and relativeness.

The community is gradually progressing towards abstractive summarization. Although a complete abstractive summarization would require deeper natural language understanding and processing, a hybrid or shallow abstractive summarization can be achieved through sentence compression and textual entailment techniques. Textual entailment helps in detecting shorter versions of text that entail with same meaning as original text. With textual entailment we can produce more concise and shorter summaries.

Text Conversation could be explored more.

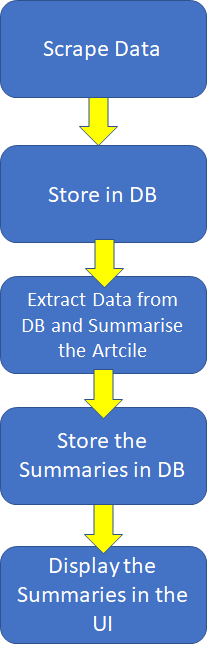
Research in summarization continues to enhance the diversity and information richness, and strive to produce coherent and focused answers to users information need.

**Implementation**

We have implemented the product using Python , Python NLTK , FLASK and MySQL.We have also tried using LSTM to generate Headline for news article.

Our current cycle is :

This Cycle is executed Daily

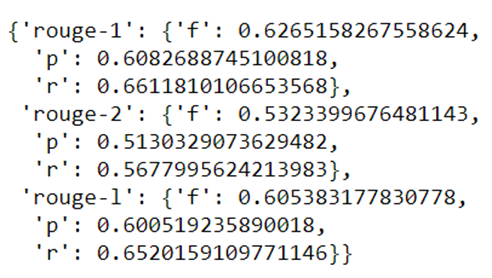


**Results :**

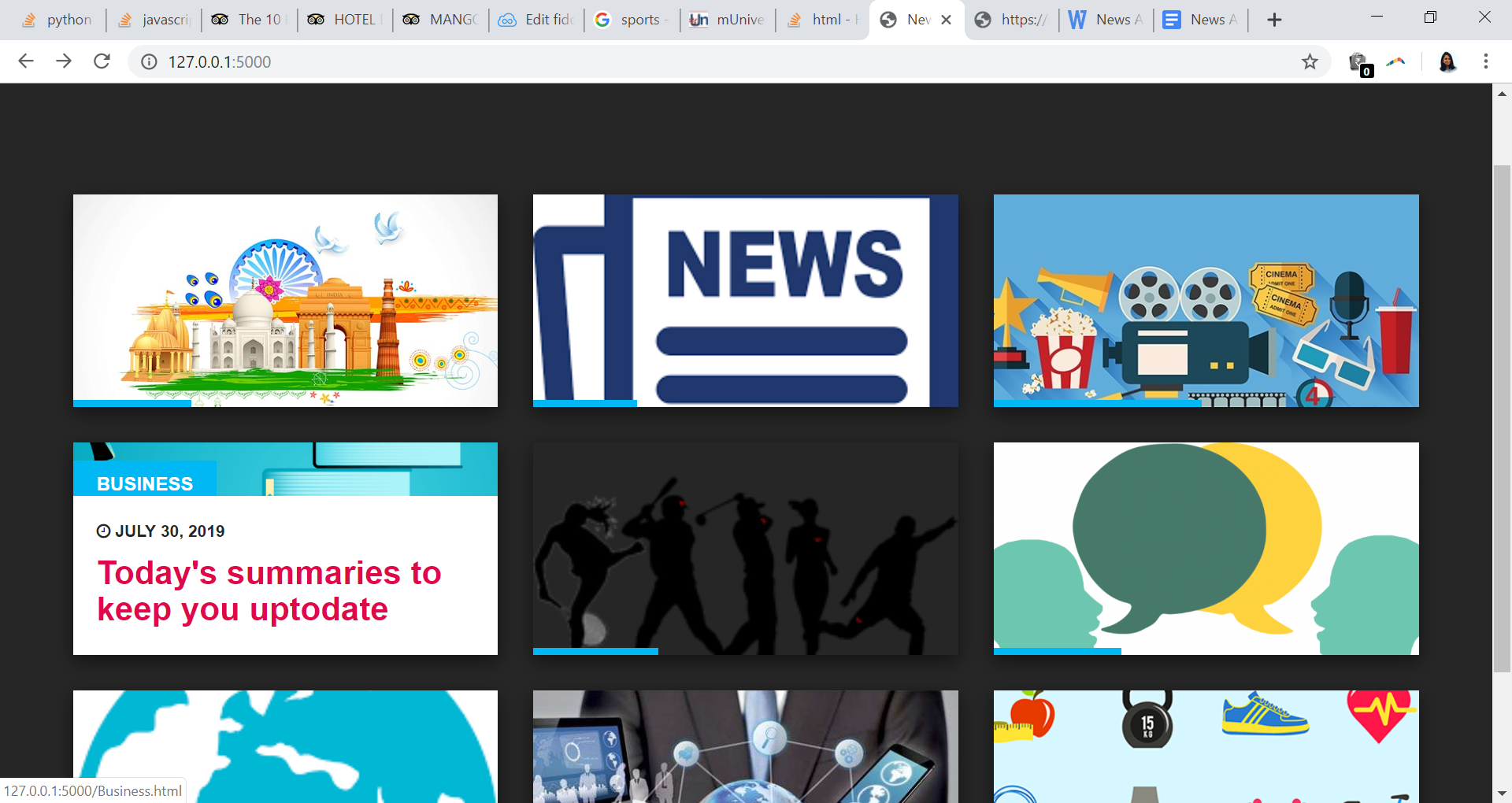
****

We have summarised the news article with 3000 words to 200 words , which will help the reader get a gist of the entire article.

For the technical part we have measured the accuracy of the generated summary using ROUGE score:



**Deployment**

****

Landing Page of the product.

Here we have the major categories users can browse through categories like Nation , Lifestyle , Sports , Business etc. and check the top news in each category by the top News Houses in India.

This product is a News Aggregator as well as News Summariser.

**Conclusion**

As natural language understanding improves, computers will be able to learn from the information on-line and apply what they learned in the real world. Combined with natural language generation, computers will become more and more capable of receiving and giving instructions. Due to rapid growth of technology and use of Internet, there is information overload. This problem can be solved if there are strong text summarizers which produces a summary of document to help user. Hence there is a need to develop system where a user can efficiently retrieve and get a summarized document. One possible solution is to summarize a document using either extractive or abstractive methods. Text summarization by extractive is easier to build. But text summarization by abstractive technique is stronger because they produce summary which is semantically related but difficult to produce. This report discussed different types of summarization methods used for summarizing a document

**Limitation**

Inaccurate extraction to essential sentences, low coverage and redundancy occurs sometimes.

Length limitation needed for an effective summary.

Redundancy elimination.

Coherency: optimal ordering of retrieved sentences to formulate the coherent context flow is the complex issue. In single document text summarization, one probable ordering sentence is given by the input text document itself. Still, this process is a nontrivial task.

**Reference**

**<https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/>**

**<https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f>**

[**https://towardsdatascience.com/text-summarization-in-python-76c0a41f0dc4**](https://towardsdatascience.com/text-summarization-in-python-76c0a41f0dc4)

[**https://arxiv.org/pdf/1804.08875.pdf**](https://arxiv.org/pdf/1804.08875.pdf)

[**https://ieeexplore.ieee.org/document/8336568**](https://ieeexplore.ieee.org/document/8336568)

[**https://waset.org/publications/4926/text-summarization-for-oil-and-gas-news-article**](https://waset.org/publications/4926/text-summarization-for-oil-and-gas-news-article)

[**https://waset.org/publications/4926/text-summarization-for-oil-and-gas-news-article**](https://waset.org/publications/4926/text-summarization-for-oil-and-gas-news-article)

[**https://machinelearningmastery.com/prepare-news-articles-text-summarization/**](https://machinelearningmastery.com/prepare-news-articles-text-summarization/)