Text Summarization

**Author: Ioana-Maria Popescu**

**Abstract**

A growing number of academics have become interested in the field of artificial intelligence known as natural language processing during the past few years. In order to conduct research and build natural language processing (NLP) and machine learning systems, Reuters offers a sizable corpus of articles. From employing word embedding regularization along with a vector space model representation to more complex message passing attention networks that are able to comprehend and categorize articles of this Corpus, previous methods that have demonstrated to yield the best results range. In order to categorize the texts in the Reuters-7083 dataset, the current research aims to construct a system using a few well-known preprocessing techniques and a Multinomial Naive Bayes algorithm. This is a document whose intent is to present all the steps implemented to create a text summarization project.

# Introduction

Data is to this century what oil was to the previous one, driven by modern technology advancements. The collection and distribution of vast amounts of data today parachutes our world. In fact, according to a projection by the International Data Corporation (IDC), the total volume of digital data that is transmitted annually around the world would increase from 4.4 zettabytes in 2013 to 180 zettabytes in 2025. With so much material being shared online, machine learning algorithms must be created that can automatically condense lengthy texts into concise summaries and provide accurate summaries that elegantly convey the intended messages.

Text summarization refers to the technique of shortening long pieces of text with the intention to create a coherent and fluent summary having only the main points outlined from the documents.

The remainder of the essay is structured as follows: In Section 2, we outline the entire pre-processing stage as well as the feature selection process. In Section 3, we go over how to represent our data, how to adapt it to the learning algorithm, and how to use several Python modules to implement the NB (Naive Bayes). Finally, in Section 4, we will wrap up the paper and compare our findings to those that have already been established by other researchers.

# Pre-processing and Feature Selection

The first step of the implementation is to specify a set of rules which all the words need to follow to have a unified and coherent set of words. These rules include eliminating punctuation, numbers, and white spaces. Once this step is over, all the words and topics are going to be extracted from documents. To do that, the implemented project search key phrases presented in all given documents like in a pattern and extract them accordingly. As an example, a list of words is created by fetching all the words contained in HTML tags such as “<text>” and “</text>”, respectively “<title>” and “</title>”. The implementation for getting the document’s topics is like the one explained above, but the searched tag is “metadata/codes”. Simulatively, the words and topics are counted to have an idea of the most and least important words and topics in the documents.

## Advantages

* **Instantly effective**

It takes time and effort to read the entire article, deconstruct it, and separate the significant concepts from the raw text. It takes at least 15 minutes to read a 500-word article.

In a fraction of a second, automatic summary software summarizes texts of 500-5000 words. This enables the user to read less data while still getting the most critical information and drawing sound judgments.

* **It Functions in Any Language**

Many summarizations software can work in any language, which is a capability that most humans lack. Because summarizers are based on linguistic models, they can automatically summaries texts in a wide range of languages, from English to Russian. As a result, they're great for persons who read and work with multilingual information.

* **Productivity is increased**

Not only do some software summaries documents, but they also summaries web pages. This boosts productivity by accelerating the surfing process.

Instead of reading entire news stories that are full of irrelevant information, summaries of such websites can be detailed and accurate while yet being only 20% of the original article's size. [1]

## Acknowledgment

The Acknowledgement will be written before References, as normal text:

**Acknowledgement:** **The name of the professor who supervised the work or project (if any) should be written here. Use a sentence like** “This work was supervised by Professor *first\_ name last\_name*, from *affiliation*”.

**In the Acknowledgement section should also be noted the founder (if any) in the form like “**This paper is founded from the research Grant M12” or “This work benefits from founds given by xx”.

# References

1. Utsav Mishra, *What Is Text Summarization in NLP?*, https://www.analyticssteps.com/blogs/what-text-summarization-nlp, 2022.

Ioana-Maria Popescu

“Lucian Blaga” University of Sibiu

Advanced Computing Systems

Romania

E-mail: maria.popescu@ulbsibiu.ro