A Minor Project Synopsis on

Text Summarization and Correlation

Submitted to Manipal University, Jaipur

Towards the partial fulfillment for the Award of the Degree of

BACHELORS OF TECHNOLOGY

In Information Technology

2019-2020

By

Bhavya Joshi

179302043

Pawan Kartik

179302044

MANIPAL UNIVERSITY
JAIPUR

Under the guidance of

Mr. Virender

Department of Information Technology

School of Computing and Information Technology

Manipal University Jaipur

Jaipur, Rajasthan

Introduction

Text Summarization is an important problem that is widely popular in the field of Natural Language Processing. Because of the ever increasing nature of data, including textual data, there is a wide need to condense it into a form more convenient for usage in many applications.

Through our project we intend to research and explore the numerous methods of text summarization present today. Methods are generally broken down into 2 main categories: -

- Extractive Summarization: In these methods the text is generally broken down into sentences or tokens and are ordered according to some measure of importance. Those parts are then combined together to form the summary of the text. Thus the summary will contain some of the exact sentences as in the original text.
- Abstractive Summarization: These methods contain new sentences derived from the meaning of the original text. These methods are generally much harder to apply as for an error free summary the computer must fully understand the natural language, derive the meaning of the full text and then use words to express that meaning.

Abstractive methods are generally based on neural networks. There are currently not many methods that are able to generate a perfect summary which contains sentences of perfect grammar and semantic meaning.

Through our project we wish to objectively use the various methods to generate summaries of various datasets and determine which of them gives the best summary. To evaluate summaries there are various methods –

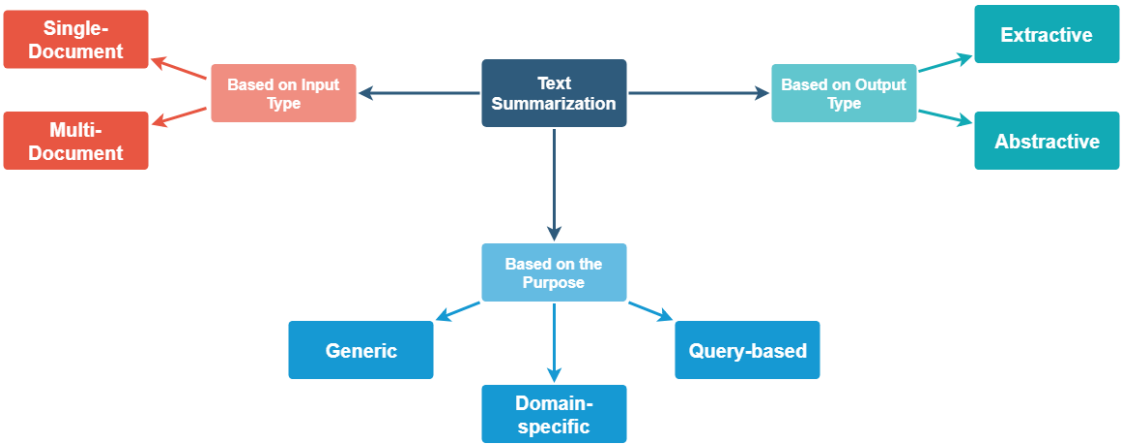
1) The simplest method is for someone to evaluate the summary and the extent of important information that it gives.
2) The second method is to use ROGUE – n where a series of n-grams is taken from the reference summaries and the generated summary. If p is "the number of common n-grams between generated and reference summary", and q is "the number of n-grams extracted from the reference summary only". The score is calculated as: ROUGE-n = p/q

Motivation

The motivation of this project was mainly an interest in undertaking a challenging project in a stimulating area of research and to try to improve the existing methods of text summarization after comparison. The ever increasing availability of documents and textual data demands thorough research into the area of automatic text summarization.

The number of methods of both exhaustive and abstractive summarization is immense and an overview of them is required in order to understand the complex topics involved and there is a lot of room for improvement as there is immense potential to introduce new factors in existing methods or to try out evaluating them using new techniques.

Automatic text summarization is a very challenging task, because when we as humans attempt to create a summary of a text we usually read it entirely to develop our understanding and derive meaning from it, and then write down a summary according to what we judged to be the important parts of the text. For a computer to do the same they would require human knowledge and language capability to the same degree as we do which makes automatic text summarization a very difficult and non-trivial task. Earlier attempts included attempting to extract important sentences from the text using features like sentence or phrase frequency. Since then there have been numerous attempts at solving the complex problem of text summarization.



Project Objective

This project aims to implement a technique for summarization and correlation and at the same time address some of the existing drawbacks.

Text Summarization can be broadly divided into 2 categories based on the approaches:

1. Exhaustive (statistical)
2. Abstractive

Exhaustive method was the foremost technique introduced for text summarization dating back to the 1950s during the absence of powerful computation hardware and the advancements of neural networks and deep learning. Hans Peter Luhn, an IBM Research employee published one of the first papers in this field following this method. Exhaustive can further be divided into subtypes depending on the approaches.

Abstract Methods are still under development. They've gained prominence after the rise of Neural Networks and Deep Learning. SalesForce has published some groundbreaking papers in text summarization (2017, 2018). This type is still being researched on and its rise is catalyzed by new advancements in neural networks.

General comparisons between the existing text summarization methods:

Exhaustive (Statistical)	**Abstractive**
Does not have the ability to comprehend language's grammar.	Has the ability to comprehend the language's grammar.
Prone to committing grammatical errors.	Can avoid grammatical errors to some extent.
Cannot figure out relation between sentences (addressed wrt tfidf).	Can figure out relation between sentences to some extent.

Requires minimum amount of data to be trained.	Requires a vast amount of data to be trained (for language rules & summarization).
Easy to implement (wrt code).	Tough to implement (wrt code).
Easy on computing resources. Hence there are no significant constraints on hardware requirements.	Requires lots of computing resources. Significant constraints on hardware requirements.
Easy to use and produces fast results.	Comparatively tough to use and produces slower results.
Can be used for different languages without explicit training.	Cannot be used for different languages without explicit training.

Although exhaustive methods lack certain advantages that abstract methods do, methods like TextRank (based on PageRank algorithm) have shown significant improvements over other conventional exhaustive models.

Methodology & General Solution

The idea of this project is to explore both exhaustive and abstractive methods and look for certain improvements. Implementation of a text summarization technique requires certain prerequisite concepts such as:

1. Word Embeddings
2. Word2Vec
3. Natural Language Processing libraries
4. Cosine Similarity and other functions that define vector relations

5. Tfidf - Term frequency & inverse document frequency etc.

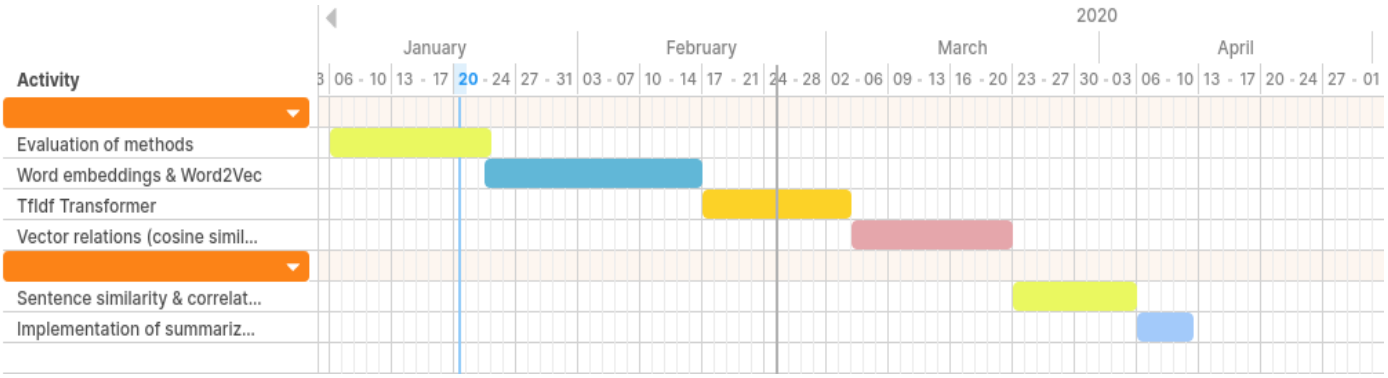
General steps involved in text summarization (with exhaustive method as a baseline):

1. Divide the paragraph into sentences. This can be achieved using a simple split or tokenizing functions (with full-stop as a delimiter).
2. A hash map of words is then built for tracking frequency of words (treated as case insensitive).
3. Words that do not carry significant weight are ignored (referred to as StopWords). Majority of NLP (Natural Language Processing) libraries or toolkits contain a list of stopwords for different languages.
4. Based on the technique being followed, i.e. Weighted Frequency model or Word Probability or some other technique, a sentence is assigned a weight. This important step is usually what differentiates different approaches and their accuracy. These techniques are called frequency driven approaches. Top k sentences (wrt to weights) are grouped (sorted under the natural occurring order in the given paragraph) form the summary.
5. Graph based approaches consider paragraph's sentences as nodes and sentences' similarities as edges between nodes. Top nodes (wrt to edge weight) form a summary for the paragraph. Certain graph based algorithms can be used to define –
    a. How 2 sentences are related?
    b. To what degree are the 2 sentences related? Etc.

The general idea is to explore ideas to answer the foremost question - how important is a sentence and how do we define this "importance"? Different approaches and different researchers view these questions differently due to their vague nature. Methods' accuracy varies based on the methods developed or used to tackle the above 2 questions. Ideally, a summary is considered better if -

1. It covers the general aspect of the paragraph.
2. Shortens or summarizes the paragraph considerably.
3. Minimizes the grammatical errors.

Gantt Chart: -



Requirements

Software:

a. Python3
b. Natural Language Processing Toolkit (NLTK)
c. Scikit-Learn.
d. Keras

Bibliography

1. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond - https://arxiv.org/pdf/1602.06023v5.pdf
2. Text Summarization Techniques: A Brief Survey - https://arxiv.org/pdf/1707.02268.pdf
3. https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-revisited/