Logo, company name

Description automatically generated

**CSE523 Machine Learning**

**Prof. Mehul Raval**

**Weekly report - 2**

**Group number: 17**

**Group name: The Mandelbrot set**

|  |  |
| --- | --- |
| **Name** | **Enrolment Number** |
| Aastha Gaudani | AU2040032 |
| Khushi Patel | AU2040068 |
| Devyash Shah | AU2040152 |
| Simran Khoja | AU1910606 |

**Report**

This week we studied a research paper related to our topic to get more information on different way of approaches for our project. Below is the summary of what we understand.

**Link of the paper** - <https://arxiv.org/pdf/1707.02268.pdf>

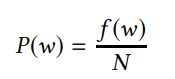
**Summary of our paper:**

**Introduction:**

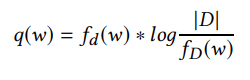
The increasing availability of online information has made automatic text summarization an important task. Extractive and abstractive summarization are two main approaches to summarization. Extractive summarization methods aim to identify important sections of the text and generate them verbatim, while abstractive summarization methods interpret and examine the text to generate a new, shorter text. Extractive summarization has been the focus of most summarization research, as it is generally more effective than abstractive summarization. This paper provides an overview of extractive summarization methods, which extract important sentences from a text to create a concise and fluent summary.

**Frequency driven approaches:**

Word probability is the simplest method, which calculates the probability of a word in a document as the number of occurrences of the word divided by the total number of words in the document. The SumBasic system uses this approach to determine the importance of sentences, assigning a weight to each sentence that is equal to the average probability of the words in the sentence.

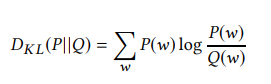


TFIDF is a more advanced technique that assesses the importance of words by identifying very common words that should be omitted from consideration and giving low weights to words appearing in most documents. The weight of each word in a document is computed by multiplying its term frequency by the logarithm of the inverse document frequency. TFIDF weights are good measures for determining the importance of sentences and are easy and fast to compute. Many existing summarizers utilize this technique or some form of it.



**Bayesian Topic Model:**

Existing multi-document summarization methods have limitations in disregarding topics embedded in documents and utilizing heuristic-based sentence scores. Bayesian topic models, such as Latent Dirichlet Allocation (LDA), offer a powerful approach to representing lost information and scoring sentences using the Kullbak-Liebler (KL) divergence measure. LDA models represent documents as a random mixture of latent topics, with each topic being a probability distribution over words. Recent studies have shown the effectiveness of LDA-based models in multi-document summarization tasks, such as BayeSum and a Bayesian sentence-based topic model. Hybrid models that incorporate hierarchical topic modeling and regression-based scoring have also been proposed. These approaches enable better representation of document topics and sentence scoring, leading to improved summarization performance.



So this much we interpreted from our report, due to the time constraints we could not manage to implement the same on our dataset. Right now we are doing tokenisation on our dataset.