EXTRACTIVE SUMMARIZATION

Goal: Given a corpus of documents,generate summary of each by extracting key sentences.

Understanding

- Text summarization is defined as the process of refining the most useful information from the source document to provide an abridged version for the specific task.In Extractive summarization we deal with finding key sentences or phrases from the original text and with help of this key sentences we need to write summary for document.
- There are various Techniques to solve this problem as mentioned in paper, Text Summarization Techniques: A Brief Survey.
  We will use Naive Bayes method to extract meaningful summary for given document.

DataSet

We will use **Dataset: CNN-Daily Mail dataset of news articles** this dataset consist of CNN, daily mail news articles containing summarized bullet points, which we can treat as summaries. In other words, the original articles are text and the bullet points are summaries.

Naive Bayes Implementation

The Extractive Summarization problem can be modeled as a classification problem, where each sentence can be classified as either a summary sentence or non-summary sentence.

Steps:

1) Document is provided for processing.
2) Sentences are extracted from document.
3) Sentences are broke into segments.
4) Features are estimated to classify sentences. Features can be location of segments, avg term freq of words occurring in segments, title words in segment, Keyword in segment etc
5) Now we can use Naive classifier to train summarizer using dataset:**CNN-Daily Mail dataset of news articles** to extract important sentences based on the feature vector.
6) After extracting sentences we will apply below processes to get proper results:

Basic Extractive Processes

- **Coverage:** It recognizes the necessary information that covers the diverse topics in input documents.
- **Coherency:** Ordering of retrieved sentences in order to get a meaningful and context based summary.We can use a ranking system to order the retrieved sentences.
- **Redundancy elimination:** Filter out the sentences having same meaning. We can use similarity measure to identify duplicates and eliminate them.

7) We will now apply our classifier on test dataset for evaluation of performance.

Above is the current Idea for implementation this can change as we further proceed in our work.