Aim:

Prepare detail statement of problem for the selected mini project and identify suitable process model for the same with justification.

Abstract:

Text Summarization solves climacteric problems in furnishing information to the necessities of user. Due to explosive growth of digital data on internet, information floods are the results to the user queries. This makes user impractical to read entire documents and select the desirables. To this problem, summarization is a novel approach which surrogates the original document by not deviating from the theme helps the user to find documents easily. Summarization area was broadly spread over different research fields, Natural Language Processing (NLP), Machine Learning and Semantics, Deep Learning.

Technical Details:

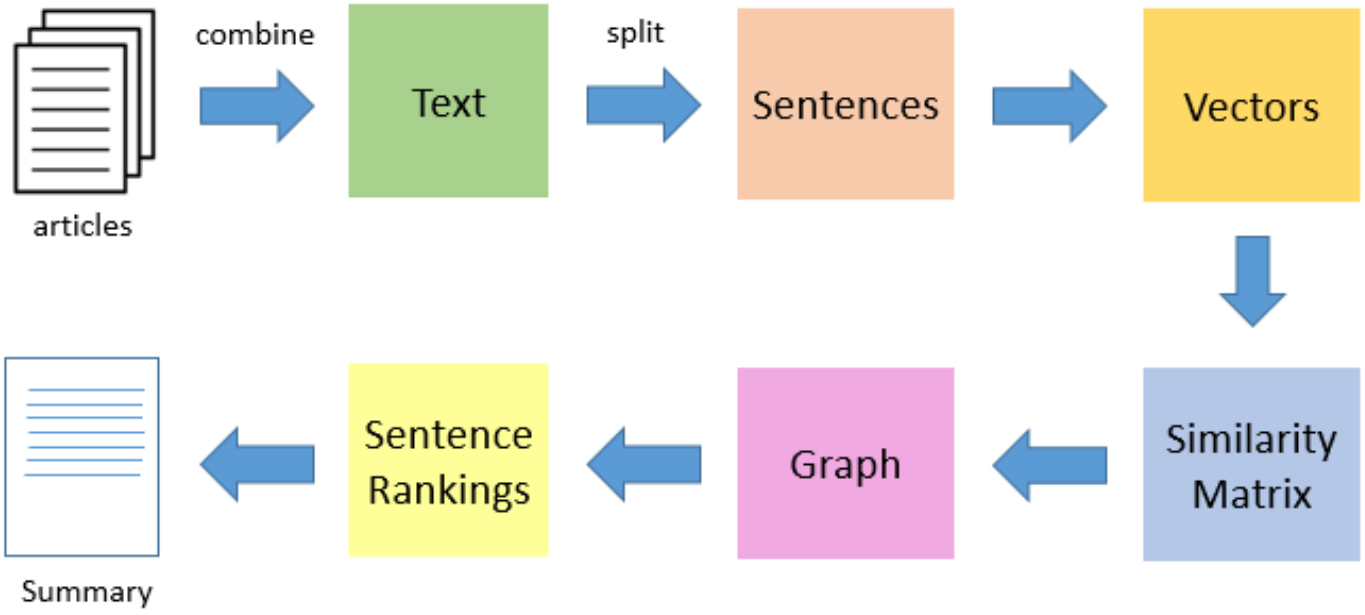
Extraction-based summarization:
The extractive text summarization technique involves pulling keyphrases from the source document and combining them to make a summary. The extraction is made according to the defined metric without making any changes to the texts.

Abstraction-based summarization:
The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method. The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text—just like humans do. Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that's why the use of extraction is still popular.

Technical Flow:

Extractive-based Text Summarization:
The algorithm used here is text-rank algorithm.
The text rank algorithm works as follow:-

- The first step would be to concatenate all the text contained in the articles
- Then split the text into individual sentences
- In the next step, we will find vector representation (word embeddings) for each and every sentence
- Similarities between sentence vectors are then calculated and stored in a matrix
- The similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation
- Finally, a certain number of top-ranked sentences form the final summary

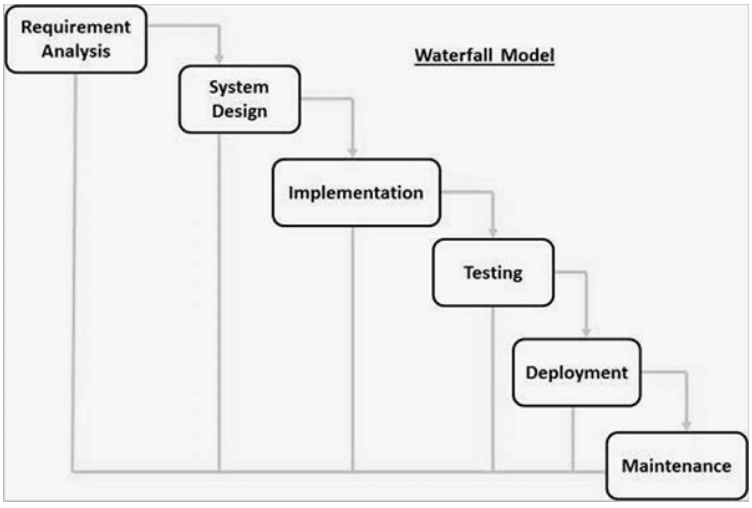


Societal Impact:

1) Reading the entire article, dissecting it and separating the important ideas from the raw text takes time and effort. Reading an article of 500 words can take at least 15 minutes. Automatic summary software summarize texts of 500-5000 words in a split second. This allows the user to read less data but still receive the most important information and make solid conclusions.
2) Some softwares summarizes not only documents but also web pages. This highly improves productivity as it speeds up the surfing process. Instead of reading full news articles that are full of useless information - the summaries of such web pages can be precise and accurate - but still 20% the size of the original article.
3) A unique features that some software have, is the ability to declare a word whose sentences that include it will automatically appear at the summary. These critical words are usually words with tactical importance, such as 'bomb', 'explode', etc. While humans can oversee an important sentence, computers will not miss it so important ideas

Software Process Model:

The Waterfall Model:



Waterfall approach was first SDLC Model to be used widely in Software Engineering to ensure success of the project. In "The Waterfall" approach, the whole process of software development is divided into separate phases. In this Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially.

Reasons to use waterfall model:

1. The requirements of our model are well defined.
2. The design of our system is simple and understandable, and design changes is allowed.
3. The implementation is divided into checkpoints, and some parts of the implementation are feasible to change to improve accuracy.
4. There is no integration required from outer scope.
5. Deployment of this system is very easy since it is generic.
6. No such maintenance is required for this system.