PROJECT - STATUS REPORT

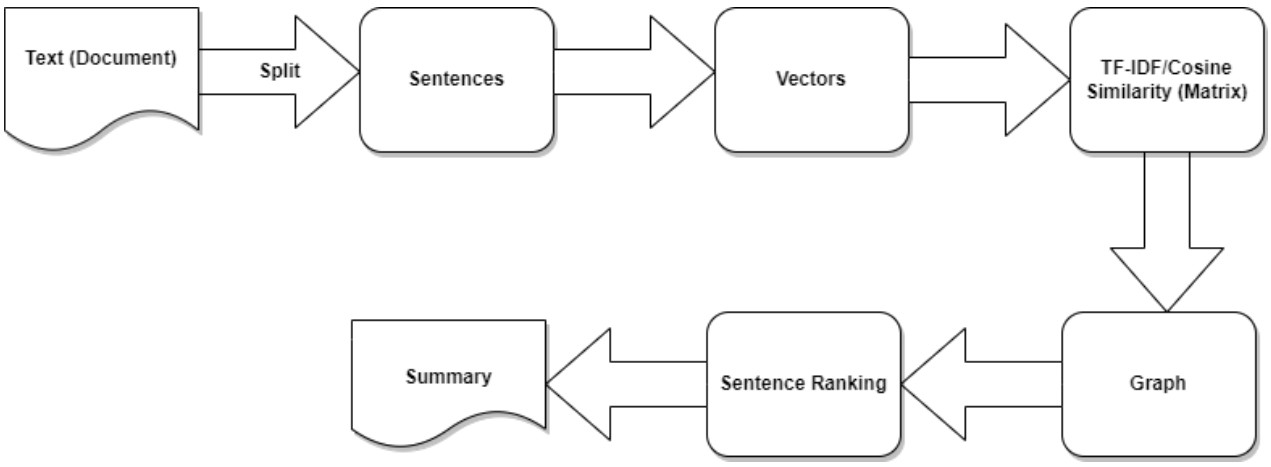
Project Topic: Document summarization using NLP techniques

Team Members:
1. Amit Kumar (Net-Id: axk210047)
2. Kirthi Menon (Net-Id: kxm190036)
3. Manpreet Sandhu (Net-Id: mxs200009)
4. Neha Ann John (Net-Id: naj210000)

Topic Gist:
Text summarization in NLP is the process of producing a concise and fluent summary while preserving key information and overall meaning. It can be performed in two ways:
1. Abstractive Text Summarisation
2. Extractive Text Summarisation

The abstractive method produces a summary with new and innovative words, phrases, and sentences whereas The extractive method attempts to summarize articles/documents by selecting a subset of words that retain the most important points. This technique weights the important part of sentences and uses the same to form the summary.

Algorithm to be Used:
Text Rank: It is an extractive and unsupervised summarization technique. It is based on the concept that words which occur more frequently are significant. Hence, the sentences containing highly frequent words are important. Based on this, the algorithm assigns scores to each sentence in the text. The top-ranked sentences make it to the summary.



Dataset Details: tennis_articles.csv In the dataset there are three columns: article_id, article_text, and source. We will use article_text column for generating the text-summary.

Coding Language / Technologies to be used: Python, PySpark