**Text Summarization Process Map**

**Approach**

Text summarization using clustering: Representing every sentence as a vector by using different features and clustering them in vector space by cosine similarity. From every cluster get the top 1 or 2 sentences depending on the length of summary needed to be generated.

**Current method**

Working on Degree centrality, Eigenvector Centrality and LexRank.

1.Degree centrality of a sentence is defined as the degree of the corresponding node in the similarity graph.  
2. Eigenvector Centrality and LexRank algorithm measure the importance of sentences in the graph by considering its relative importance to its neighboring sentences.

Representing the document as a graph ( [(v,e)=(sentence, similarity weight)] ). Similarity is found by using **idf-modified-cosine** in Degree centrality and prestige concept is adopted to find the lex rank to remove the negative effect (non-important sentences show more similarity among each other but they should not get included in summary) of Degree centrality.

**Reference for this approach:**

1. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization by Gunes Erkan and Dragomir R. Radev

Praposed ideas: To find the similarity based on the sense of the sentences instead of just tf and idf values.

(i) Can include Thematic features (eg: count of nouns, upper case words, etc).

(ii) Same POS tags with negation (eg: Yuvi played well in last WC but he will **not** be playing in this WC)

(iii) Synonyms can be used for getting the sense similarity of the sentences.

(iv) Position of sentences in a paragraph

(v) n-grams as a unit for calculating tf and idf.

Work till now:

Finding the idf-modified-cosine and similarity between every sentence and degree centrality.

Working out the method to find the threshold in terms of edge weigth for Degree measure.

**Future work:**

Applying the **power method** to include prestige concept and find the lex rank.

Complete the model and then include new ways to improve it.

**Challenges:**

1. In Degree centrality method finding the threshold for degree of node to include it into the summary.

**Previous works with reference**:

1. Text Summarization using a Machine Learning Approach (Naive Bayes and C4.5 as classifier)

Reference: Automatic Text Summarization using a Machine Learning Approach by Joel Larocca Neto Alex A. Freitas Celso A. A. Kaestner

2. Text Summarization via Hidden Markov Models. Used five features in the development of a Hidden Markov model for text summarization.

Reference: Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition by John Conroy and Dianne P. O'Leary

3. QR Method of Text Summarization. Idea behind the algorithm is given the sentences in the document and a measure of the importance of each, choose the most important sentence to add to the summary.

Reference: Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition by John Conroy and Dianne P. O'Leary

4. Text Summarization using Term Weights

Reference: Text Summarization using Term Weights by R.C. Balabantaray, D.K. Sahoo

5. Using Lexical Chains for Text Summarization

Reference: Using Lexical Chains for Text Summarization by Regina Barzilay and Michael Elhadad

6. Sentence Centrality and Centroid-based Summarization. The centroid of a cluster  
is a pseudo-document which consists of words that have tf×idf scores above a predefined  
threshold

Reference: centroid-based summarization by Radev, Jing, & Budzikowska

7. Centrality-based Sentence Salience, Degree Centrality, Eigenvector Centrality and LexRank. By using the idf-modified-cosine similarity between sentences and threshold for degree centrality. Then for removing the negative effect of degree centrality using LexRank for finding the sentences to be included in summary using lex rank and continuous lex rank.

Reference : LexRank: Graph-based Lexical Centrality as Salience in  
Text Summarization by Gunes Erkan and Dragomir R. Radev

8. Summly an application for generating summaries for news article sold to yahoo.