Text Summarization Report

1. TF-IDF Based Summarizer:

Term Frequency: For every word in given sentence we have calculated term frequency

$$TF = \frac{Number\ of\ Repetition\ of\ W_i\ in\ sentence}{Number\ of\ words\ in\ sentence}$$

Inverse Document Frequency: For every word in complete text given IDF Score.

$$IDF = log(\frac{Total\ Number\ of\ Sentences}{Number\ of\ sentences\ containg\ word\ W_i})$$

TF-IDF: Assigning weight for each word based on TF and IDF score

$$Weight(W_i) = TF * IDF$$

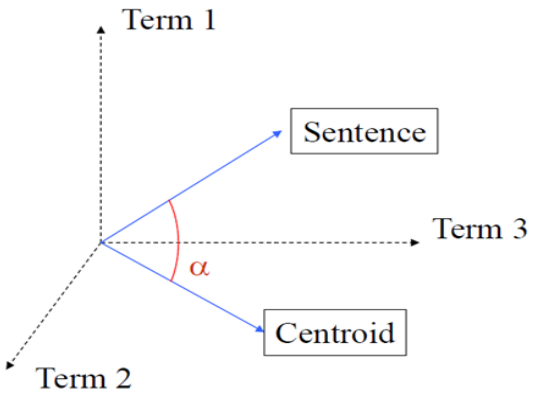
Sentence Weightage:

$$Sentence\ Weight = \frac{Sum\ of\ weight\ of\ all\ words\ in\ sentence}{Number\ of\ words\ in\ Sentence}$$

Sentence Extraction:

sentence weight> average sentence weight * tuning factor

2. Centroid Based Summarizer:



Centroid Vector: We have taken those words in a text which have higher tf-idf score than the average tf-idf score.

Sentence Vector: All the words along with tf-idf score in a sentence act as vector

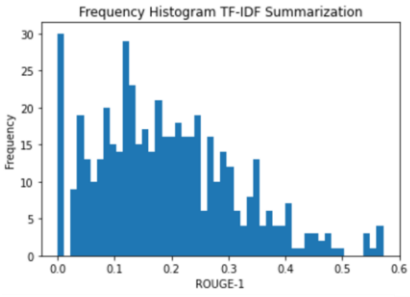
Cosine Sentence Similarity: Weight of sentence given based on the cosine similarity between the centroid vector and sentence vector

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

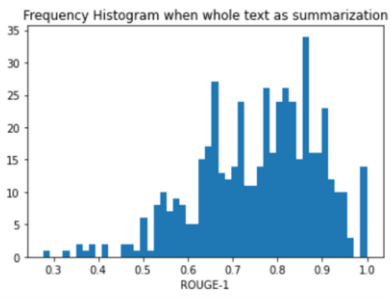
Sentence Extraction:

Cosine sentence similarity>avg cosine sentence similarity*tuning factor

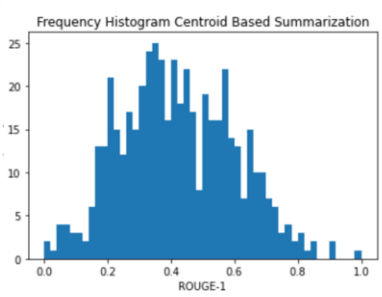
3. TF-IDF vs Centroid based summarizer



TF-IDF Summarizer
Avg. ROUGE: 0.18



Whole text as a Summary
Avg. ROUGE: 0.76



Centroid Summarizer
Avg. ROUGE: 0.42

All the above results are based on 500 instances of data.

Given that the gold reference summary is abstractive in nature, in this scenario, by using extractive summarization, the maximum ROUGE-1 score attainable will always be less than the ROUGE-1 score of the whole text as summary.

We are getting an average ROUGE-1 score as 0.18 for TF-IDF summarizer and 0.42 as Centroid based summarizer. It is evident that centroid based summarizers are performing better for giving the central idea of text.