Text Summarization (Extractive Method)

Text summarization is a process of creating concise versions of the original text while retaining key information. Humans are naturally good summarizers for we have the ability to understand the overall meaning of a text just by reading it. But how can machines do the same ? Well this is what we are going to discuss in this article.

There are numerous applications of the automatic text summarization including following:

1. Reading lengthy customer reviews and converting them into smaller and meaningful versions to be used to take necessary actions.
2. Converting news articles into a short summary. Mobile app **inshorts** is an example of this.
3. Creating concise summary reports from business meeting notes.

There are two types of text summarization methods, namely:

1. Extractive Text Summarization
2. Abstractive Text Summarization

In the approach I've taken, we'll be using TF-IDF to calculate the frequency of words and tag their importance.

TFIDF, short for term frequency–inverse document frequency, is a numerical measure that is used to score the importance of a word in a document based on how often it appears in that document and a given collection of documents. The intuition behind this measure is : If a word appears frequently in a document, then it should be important and we should give that word a high score. But if a word appears in too many other documents, it's probably not a unique identifier, therefore we should assign a lower score to that word.

Formula for calculating tf and idf:

- **TF(w)** = (Number of times term w appears in a document) / (Total number of terms in the document)
- **IDF(w)** = log_e(Total number of documents / Number of documents with term w in it)

Hence tfidf for a word can be calculated as : **TFIDF(w) = TF(w) * IDF(w)**

Approach :

1. Import necessary packages like NLTK, networkx, numpy, regex and initializing WordNetLemmatizer
2. Text preprocessing to remove all the special characters.
3. Next we calculate the frequency of a word in a document.
4. Then we calculate the sentence score incorporated with POS Tagging and TF-IDF score.
5. Finally the score for each sentence is calculated and the most important one is taken out based on the retention rate provided by the user.

Note - Retention rate is the rate decided by the user that gives us the amount of (in percentage) information to be retained from the original input, or how compact you want the summary to be.