**NLP Text Summarization**

**Problem Statement:**

Unstructured data is difficult to process, but the major portion of data today is unstructured in nature. Given a paragraph on some topic, use natural language processing techniques to

summarize the text.

**Pre-Processing:**

The text data is in raw form and may contain errors and unwanted text. Hence to achieve better, accurate and efficient outcomes, pre-processing the data is essential.

1. **Tokenization:** This is done to split the data, i.e. the paragraph, into sentences and these sentences are split into words.
2. **Lower casing:** All the capital letters in the words are converted to lower case. This is done because if the same word is used twice, but one has a capital letter, they would be read as different words without this step.
3. **Removing stop words:** Stop words are commonly used words used to combine sentences, e.g. a, an, the, is, etc. These words do not really have any importance as they do not help in distinguishing two documents, and so they are removed.
4. **Lemmatization:** it is used to remove any kind of suffix from the word and return the word in its original form (root word) and this word is always meaningful and belongs to the dictionary. The word generated after lemmatization is also called a lemma.
5. **Removing of non-ASCII characters:** These characters add no value to text-understanding and induce noise into algorithms. So they must be removed.
6. **Removing HTML tags:** HTML tags are like keywords which define how a web browser will format and display the content. Often, unstructured text contains a lot of noise, especially if we use techniques like web or screen scraping. HTML tags are typically one of these components which don’t add much value towards understanding and analysing text so they should be removed.
7. **Removing URLs:** A URL is a link for any given resource on the internet. A URL is unique for every resource but they all follow the same structure. A URL will be different in every text and a given text may contain so we need to first identify the URL from its format and remove it.

**Text Summarization:**

**Approach: TextRank**

It is a graph-based ranking model for text processing which can be used in order to find the most relevant sentences in text and also to find keywords. It is an extractive and unsupervised 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, scores are assigned to each sentence, and the top ranked sentences are used in the summary.

This algorithm is based on the PageRank algorithm, which used by Google Search to rank web pages in their search results. It gives the probability distribution of a user arriving at a particular page by transitioning from one page to another.

In a similar fashion, TextRank applies this algorithm on sentences instead of web pages, and the similarities are stored in a square matrix.

All the text of the paragraph is first concatenated, and then it is split into different sentences.

Then the vector representation (word embedding) is found for each of these sentences, and their similarities are calculated and stored in a square matrix.

The square matrix is then converted into a graph for calculating sentence ranks, where the sentences are the vertices and the edges are based on content overlap, i.e. the number of words two sentences have in common.

This network of sentences is fed into the PageRank algorithm to identify the importance rank of every sentence.

The top ranked sentences are extracted and used in the summary.

**Conclusion:**

TextRank is a one of the simpler algorithms for text summarization and has many advantages, so it is a natural choice for extractive summarization.

It is unsupervised: it does not require training or input to process the given data.

It is language independent: it works solely depends on word concurrence and does not require knowledge of grammar. So for the most part, it works regardless of the language used in the data.

TextRank is a widely used, well developed and easy to understand algorithm, making it a good choice for summarizing a given paragraph.