NLP INTERESTSHIP

MILESTONE 1 : Introduction to NLP, NLP Applications, Corpus, Tokenization, Normalization, Stemming, Lemmatization, Stop Words Removal.

By Pallavi Pannu

Introduction to NLP

- According to industry estimates, only 21% of the available data is present in a structured form.

- The majority of the data exists in the textual form, which is highly unstructured in nature.

What is NLP?

- NLP is a field of linguistics and machine learning focused on understanding everything related to human language.

- The aim of NLP tasks is not only to understand single words individually, but to be able to understand the context of those words.

Some common NLP Tasks !!

1. **Classifying whole sentences:**

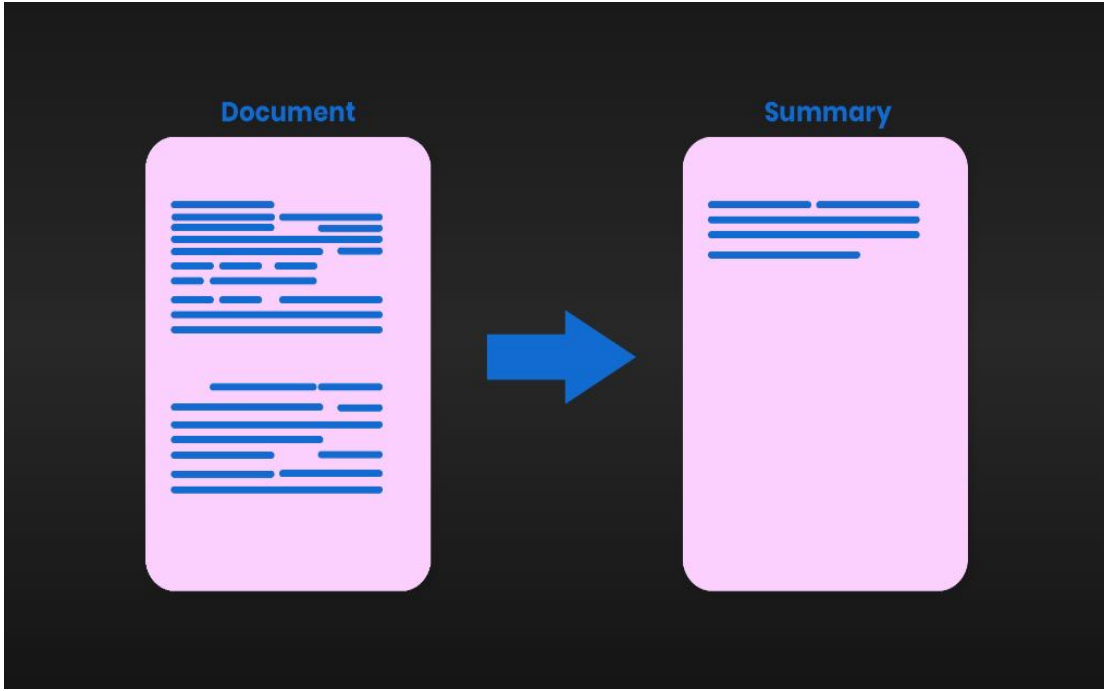
   Getting the sentiment of review, detecting if email is spam or not.

Some common NLP Tasks !!

2. Translating a text into another language.

Some common NLP Tasks !!

3. Text Summarization.



The first thing we need to do in any NLP Project is text preprocessing !!

Basics of NLP

1. **Corpus** - A Corpus is defined as a collection of text documents.

   For example a data set containing news is a corpus or the tweets containing Twitter data is a corpus.

Corpus > Documents > Paragraphs > Sentences > Tokens

Basics of NLP

2. Tokenization- It’s the process of breaking a stream of textual data into small units called tokens.

A token can be word, part of word, sentences, symbols, etc.

Why do we need tokenization?

1. A tokenizer breaks unstructured data and natural language text into chunks of information that can be considered as discrete elements.

2. The token occurrences in a document can be used directly as a vector representing that document.

3. This immediately turns an unstructured string (text document) into a numerical data structure suitable for machine learning.

Example of sentence tokenization

```
sent_tokenize('Life is a matter of choices, and every choice you make makes you.')
```

```
['Life is a matter of choices, and every choice you make makes you.']
```

Example of word tokenization

```
word_tokenize("The sole meaning of life is to serve humanity")
```

```
['The', 'sole', 'meaning', 'of', 'life', 'is', 'to', 'serve', 'humanity']
```

Normalization

Normalization is the process of converting a token into its base form. It is useful in reducing the number of unique tokens present in the text.

Structure of token : <prefix> <morpheme> <suffix>

Stemming

Reducing words to their basic form or stem.

“laughing”, “laughed”, “laughs”, “laugh” >>> “laugh”

Lemmatization

- It is a systematic step-by-step process for removing inflection forms of a word.

- It considers the context and converts the word to its meaningful base form, which is called Lemma.

- **Stemming** the word **‘studies‘** would return **‘studi‘.**

  Lemmatizing the word **‘studies‘** would return **‘study‘.**

Stop words removal

The words which are generally filtered out before processing a natural language are called **stop words**.

These are actually the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc) and does not add much information to the text.

Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”.

Thank You !!