NLP Experiment 3

Aim: To implement NLP Pre-processing Tasks

Theory:

Preprocessing steps:
Removing inconsistent data, in the process of web scraping a few longer synopses had hyperlinks to expand their content, the dataset has text content of those hyperlinks as well. We had to remove the inconsistent part of the data.

Example:

Removing rows that were outside of the default English character set, there were multiple instances where a foreign script was scraped but since the system did not support the same, there was a loss of data and random characters added noise in those rows.
This was done using a regular expression: re.compile(r'^[a-zA-Z\s]*$')

The above regex matches with all English characters, those unmatched with the regex are replaced with a white space (‘ ’) and all blank/white space data (that is all cells of the data frame that are empty) are removed from the data frame.

Example:
We remove all the punctuations in each synopsis using a regex:
re.sub(r'[^\w\s]', '', sentence)
The above regex substitutes each character that is not a word character /w or a /s a white space character which effectively removes punctuations from a sentence.
We convert all of the text data into lowercase.
We split all sentences into tokens and each unique token is assigned a unique number representing the token.
We represent each sentence into a sequence of those numbers in this method.

Libraries and Tools Used:

- Pandas (used for manipulating data)
- re (for matching and removing noisy data from the dataset)