NLP Assignment 1

Aryan GD Singh (2019459)

Methodology

1. Preprocessing of data
2. Performing tasks on the new data

Task specific information is written in the main.ipynb file.

Preprocessing

1. Remove duplicate entries from the dataset
2. Remove null entries from the dataset

Assumptions

While checking phone numbers, only 10 digit numbers(along with 1 or 2 prefix digits) are checked.

Capitalised word is one with all its letters in uppercase, and with a length greater than 1.

For monetary quantities, only (£, $, €, ¥) symbols and a few words for currencies are checked (pounds, dollars, rupees).

Clitics are assumed to be of maximum length 2. So for example 'est' is not a valid clitic.

A message starting or ending with a word means it has to match the complete word itself, not the word as a part of another larger word.

For example-
If the input is ‘foot’, and we have to check for start -

‘foot long burger’ will match
‘football is great!’ will not match