**NLP text prepossessing task**

This task aims to clean the text data to the required format by fulfilling the following requirements

1. Remove multiple spaces
2. Delete usernames and hashtags
3. Delete anything in quotation marks
4. Delete URLs
5. Delete email address
6. Replace words with the need to (refer to the replace word table)
7. Replace multiple punctuation with the last punctuation appeared (eg: ?!?! -> !)
8. Remove multiple non common punctuation (eg: ((((&&&hi -> hi)
9. Add a space before all remaining punctuation marks except for the decimal point between numbers (eg: $1.43)

|  |  |  |
| --- | --- | --- |
| Steps No | Steps | Code example |
| 1 | Removes any leading (spaces at the beginning) and trailing (spaces at the end) characters |  |
| 2. | Remove multiple spaces in between the sentence |  |
| 3 | Remove username and hashtags |  |
| 4 | Delete anything in quotation marks |  |
| 5 | Delete email address |  |
| 6 | Replace words with the need to |  |
| 7 | Replace multiple punctuation with the last punctuation appeared (eg: ?!?! -> !) |  |
| 8 | Remove multiple non common punctuation (eg: ((((&&&hi -> hi) |  |
| 9 | Add a space before all punctuation marks except for the decimal point between numbers and $ before number |  |

Most of the functions can be fulfilled by the regular expression operations. Click [here](https://docs.python.org/3/library/re.html) for the documents. A sample of NLP processing code can be found [here](https://colab.research.google.com/drive/1GWhAO4fwGQ6enoiwLg4AOBBBJrJ4L0-A?usp=sharing).