Quiz 4 - Text Pre-processing

Total points 10/10

Email *

Mohamadhirfan1@gmail.com

0 of 0 points

Full Name *

Mohamadh Irfan

Quiz 3 - Text Pre-processing

3 of 3 points

Instructions:

1. **Download the dataset from here: https://drive.google.com/file/d/14yA_hivQOTtwwYVMOqX7edLnUxD0xEn/view?usp=share_link**
2. **Write a python script to convert the data into a dataframe (check the section below for the head and tail of dataframe).**

Write a python script to convert the data into a dataframe as shown below:

```
df.head()
```

	label	text
0	MajorClaim	we should attach more importance to cooperatio...
1	MajorClaim	a more cooperative attitudes towards life is m...
2	Claim	through cooperation, children can learn about ...
3	Premise	What we acquired from team work is not only ho...
4	Premise	During the process of cooperation, children ca...

```
df.tail()
```

	label	text
6084	Premise	indirectly they will learn how to socialize ea...
6085	Premise	That will make children getting lots of friends\n
6086	Premise	they can contribute positively to community\n
6087	Premise	playing sport makes children getting healthy a...
6088	Claim	playing sports will give good effects on child...

✓ What is the format of given raw data? * 1

- ○ CSV Files
- ◉ Text Files
- ○ JSON files
- ○ Database files

✓ How many total files are given in the .txt format? * 1

- ○ 6088
- ◉ 6089
- ○ 6098
- ○ 6090

✓ How many files are having 'Premise' label? * 1

- ○ 751
- ◉ 3832
- ○ 1506
- ○ None of the above is correct

Instructions for the following questions: 7 of 7 p[illegible]ts

1. Don't split the data into train and test.
2. If there is a requirement to pre-process the data, perform the operations on the en[illegible] data.

✓ What is the maximum number of character level tokens in raw 'text' column? 1

- ○ 735
- ○ 78
- ○ 543
- ◉ 345

✓ What is the maximum number of word level tokens in raw 'text' columns? * 1

- ○ 63
- ○ 75
- ◉ 67
- ○ 76

✓ After applying text cleaning (i.e. text pre-processing), what is the maximum number of word level tokens in clean 'text' column? 1

- ◉ 31
- ○ 21
- ○ 30
- ○ 20

✓ What is the output if you apply all the text pre-processing on file_1? (Select all the correct options) 2

- [x] attach import cooper primari educ
- [x] attach importance cooperation primary education
- [x] Number of word level token is same irrespective of stemming or lemmatisation in the text cleaning step.
- [x] Number of word level tokens = 5

✓ What is the total number of unique vocab words in the entire corpus? * 2

- [ ] There are total 5970 1-gram vocabulary words
- [x] There are total 41969 2-gram vocabulary words
- [x] There are total 79726 3-gram vocabulary words
- [x] There are total 5974 1-gram vocabulary words