

ML Phase 1 Assignment

Topic: Natural Language Processing

URL to download the dataset: https://drive.google.com/file/d/14yA_hivQOTtw-wYVMOqX7edLnUxD0xEn/view?usp=share_link

Q1) Write a python script to convert the data into a dataframe (check the section below for the head and tail of the dataframe).

```
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...

In [7]: *# Answer here*

Q2) What is the format of given raw data?

In [1]: *# Answer here*

Q3) How many total files are given in the .txt format?

In [2]: *# Answer here*

Q4) How many files are having 'Premise' label?

In [3]: *# Answer here*

Instructions for the following questions:

- Don't split the data into train and test.
- If there is a requirement to pre-process the data, perform the operations on the entire data.

Q5) What is the maximum number of character level tokens in the raw 'text' column?

In [4]: *# Answer here*

Q6) What is the maximum number of word level tokens in raw 'text' columns?

In [5]: *# Answer here*

Q7) After applying text cleaning (i.e. text pre-processing), what is the maximum number of word level tokens in the clean 'text' column?

In [6]: *# Answer here*

Q8) What is the total number of unique vocab words in the entire corpus?

In []: *# Answer here*