Title and Group:

Group members: Chow Jun Wei (jun.chow.18@ucl.ac.uk, Brunei Darussalam, NIL, Data Science and NLP and Data Analyst), Esraa Sultan (Alzahrani.Esraa1@gmail.com, Saudi Arabia, Esri Saudi Arabia, Data Science).
Internship Batch: NLP 02

Problem Description

There are 20 different categories of articles, and our job is to classify, to which category does each article belongs to.

Business Understanding

Working for a newspaper office, our boss task us with classification of the news article being written. Previously, it was being classified by human (particularly the article writer). Our boss thought that article writer should focus on writing articles than classification job, and a human classifier is too expensive to hire, so he would like to develop an ML model to do the classification job.

Project Lifecycle

Day	14	15	16	17	18	19	20	21	22	23	24	25
Define Project Lifecycle	■											
Business Understanding	■											
Data Intake Report	■											
Data Understanding		■	■									
Data Cleansing		■	■									
Peer Review			■	■								
EDA				■	■	■						
Review Slide with Discussions						■	■					
Model Building								■	■	■		
Model Presentation										■	■	■

Data Intake Report

Name: NLP Group Project
Report date: November 14, 2021
Internship Batch: NLP 02
Version:<1.0>
Data intake by: Chow Jun Wei, Esraa Sultan
Data intake reviewer:<intern who reviewed the report>
Data storage location: https://github.com/Wabinab/NLP_GroupProject_DG

Note: Since we have too many files, we will list their folders instead.

Note: At the end these files aren't used, rather we changed to this:

```
from sklearn.datasets import fetch_20newsgroups

all_xs, all_y = fetch_20newsgroups(subset="all",
      remove=('headers', 'footers', 'quotes'),
     shuffle=True,return_X_y=True)
```

The files are arranged such that we retain the original data method: each article are their own .txt files. There are more than 1000 files so if we list them here it would take too long. Rather, we would group by each category instead and mention how many files there are. Hence, "Total number of observations" and "Total number of features" would be the NIL for all. We changed "Total number of features" to "Base Folder". And one file called "errors.txt" containing the files that cannot be processed due to reasons (mostly due to cannot decode with UTF-8 and we aren't sure about what encoding it uses so it's ignored).

Tabular data details:

Total number of observations	NIL
Total number of files	387
Base Folder	Alt.atheism
Base format of the file	.txt
Size of the data	Total: 2.1MB

Total number of observations	NIL
Total number of files	185
Base Folder	Comp.graphics
Base format of the file	.txt
Size of the data	Total: 1.5MB

Total number of observations	NIL
Total number of files	184
Base Folder	Comp.os.ms-windows.misc
Base format of the file	.txt
Size of the data	Total: 2.0 MB

Total number of observations	NIL
Total number of files	195
Base Folder	Comp.sys.ibm.pc.hardware
Base format of the file	.txt
Size of the data	Total: 924 kB

Total number of observations	NIL
Total number of files	132
Base Folder	Comp.sys.mac.hardware
Base format of the file	.txt
Size of the data	Total: 624 kB

Total number of observations	NIL
Total number of files	249
Base Folder	Comp.windows.x
Base format of the file	.txt
Size of the data	Total: 1.8 MB

Total number of observations	NIL
Total number of files	180
Base Folder	Misc.forsale
Base format of the file	.txt
Size of the data	Total: 820 KB

Total number of observations	NIL
Total number of files	234
Base Folder	Rec.autos
Base format of the file	.txt
Size of the data	Total: 1.1 MB

Total number of observations	NIL
Total number of files	168
Base Folder	Rec.motorcycles
Base format of the file	.txt
Size of the data	Total: 744 KB

Total number of observations	NIL
Total number of files	271
Base Folder	Rec.sport.baseball

Base format of the file	.txt
Size of the data	Total: 1.3 MB

Total number of observations	NIL
Total number of files	310
Base Folder	Rec.sport.hockey
Base format of the file	.txt
Size of the data	Total: 1.7M

Total number of observations	NIL
Total number of files	321
Base Folder	Sci.crypt
Base format of the file	.txt
Size of the data	Total: 2.0 MB

Total number of observations	NIL
Total number of files	193
Base Folder	Sci.electronics
Base format of the file	.txt
Size of the data	Total: 900 KB

Total number of observations	NIL
Total number of files	277
Base Folder	Sci.med
Base format of the file	.txt
Size of the data	Total: 1.7 MB

Total number of observations	NIL
Total number of files	272
Base Folder	Sci.space
Base format of the file	.txt
Size of the data	Total: 1.6 MB

Total number of observations	NIL
Total number of files	442
Base Folder	Soc.religion.christian
Base format of the file	.txt
Size of the data	Total: 2.5 MB

Total number of observations	NIL
Total number of files	400
Base Folder	Talk.politics.guns
Base format of the file	.txt
Size of the data	Total: 2.2 MB

Total number of observations	NIL
Total number of files	530
Base Folder	Talk.politics.mideast
Base format of the file	.txt
Size of the data	Total: 3.5 MB

Total number of observations	NIL
Total number of files	450
Base Folder	Talk.politics.misc
Base format of the file	.txt
Size of the data	Total: 2.8 MB

Total number of observations	NIL
Total number of files	384
Base Folder	Talk.religion.misc
Base format of the file	.txt
Size of the data	Total: 2.1 MB

Total number of observations	NIL
Total number of files	1
Total number of features	NIL
Base format of the file	errors.txt
Size of the data	4.1 kB

Note: Replicate same table with file name if you have more than one file.

Proposed Approach:

- Mention approach of dedup validation (identification)
- Mention your assumptions (if you assume any other thing for data quality analysis)

Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.
Please do not forget to remove this section while converting the file into pdf.