HW1 Part A Feature Engineering - 7 Points

- **You have to submit two files for this part of the HW (1) ipynb (colab notebook) and (2) pdf file (odf version of the colab file).**
- **Files should be named as follows: FirstName_LastName_HW_1A**

Install/Import Modules

```
if 'google.colab' in str(get_ipython()):
    !pip install  -U spacy  -q
    !python -m spacy download en_core_web_sm
```

```
2023-08-27 19:11:09.883106: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is optimized to use available CP
To enable the following instructions: AVX2 AVX512F FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2023-08-27 19:11:10.860168: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
2023-08-27 19:11:12.419490: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:996] successful NUMA node read from SysF
2023-08-27 19:11:12.419949: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:996] successful NUMA node read from SysF
2023-08-27 19:11:12.420103: I tensorflow/compiler/xla/stream_executor/cuda/cuda_gpu_executor.cc:996] successful NUMA node read from SysF
Collecting en-core-web-sm==3.6.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.6.0/en_core_web_sm-3.6.0-py3-none-any.whl (12
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 56.2 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.7.0,>=3.6.0 in /usr/local/lib/python3.10/dist-packages (from en-core-web-sm==3.6.0) (3.6.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-w
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm=
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-s
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm=
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm=
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-we
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm=
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (3.1.
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6.0) (
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web-sm==3.6
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-core-web
Requirement already satisfied: annotated-types>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.
Requirement already satisfied: pydantic-core==2.6.1 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.
Requirement already satisfied: typing-extensions>=4.6.1 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,>=3.6.
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->spacy<3.7.0,
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.0,>=3.
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8.1.8->spacy<3.7.
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.10.0,>=0.3.0->spacy<3.7.0,>=
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<3.7.0,>=3.6.0->en-core-web
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
```

```
# Import the spacy library for natural language processing
import spacy
# Import the pandas library for data manipulation and analysis
import pandas as pd
# Import the pathlib library for working with file paths
from pathlib import Path
# Import the re module for regular expressions
import re
# Import the random module for generating random numbers and samples
import random
# Import the BeautifulSoup module for parsing HTML and XML documents
from bs4 import BeautifulSoup

# Import the numpy library for numerical computing
import numpy as np
```

Task1: Feature Engineering and Preprocessing IMDB - 7 points

You can use regular expression or spacy for this task

- **PreProcessing**:
    1. Remove HTML tags and new line character (\n)
    2. Remove email, urls and punctuations

  For preprocessing, write your own simple functions and your final cleaned text should be saved in a new column - `cleaned_text`.

- **Feature Engineering**

  Use the `cleaned_text` column you created in the previous step and extract following features as new column.

    1. number of words
    2. number of characters
    3. number of characters without space
    4. average word length
    5. count of numbers(37, 201, 20 etc.)

You will use the imdb moview review dataset. The details of the data can be found from this link : https://ai.stanford.edu/~amaas/data/sentiment/.

Description of the data from the above link : "*This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided. See the README file contained in the release for more details.*".

We extracted the data from text files and save the train and test data as csv files. **We will use train.csv file for this task. The file is availibale in 0_Data_folder in Course Home Page.**

Take a 10% subset of the data for the HW..

```
# Check if the code is running in a Colab environment
if 'google.colab' in str(get_ipython()):  # If the code is running in Colab
    # mount google drive
    from google.colab import drive
    drive.mount('/content/drive', force_remount=True)

    # set the base path to a Google Drive folder
    base_path = '/content/drive/MyDrive/NLP'
else:
    # If the code is not running in Colab, set the base path to a local folder
    base_path = '/home/harpreet/Insync/google_drive_shaannoor/data'


# Convert the base path to a Path object
base_folder = Path(base_path)

# Define the archive folder path
archive_folder = base_folder/'archive'

# Define the data folder path
data_folder = base_folder/'datasets'
```

```
Mounted at /content/drive
```

```
train_data = pd.read_csv(data_folder / 'aclImdb'/'train.csv', index_col=0)

train = train_data.sample(frac = 0.1)
```

```
train.head()
```

	Reviews	Labels
9387	This movie gets it right. As a former USAF Avi...	1

```
# load model
nlp = spacy.load('en_core_web_sm')
```

24748	This movie is really nothing besides an admitt...	0

```
import re
from bs4 import BeautifulSoup

def clean_text(input_text):
    # Remove HTML tags using BeautifulSoup
    soup = BeautifulSoup(input_text, "html.parser")
    text_without_tags = soup.get_text()

    # Remove newline characters using regular expressions
    text_without_newlines = re.sub(r'\n', ' ', text_without_tags)

    # Remove extra spaces
    cleaned_text = ' '.join(text_without_newlines.split())

    return cleaned_text



from spacy import tokens
# initialize an empty list to store tokens
tokens_method2 = []

# temporarily disable the named pipes of spaCy NLP processing pipeline
disabled = nlp.select_pipes(
    disable=['tok2vec', 'tagger', 'parser', 'attribute_ruler', 'lemmatizer', 'ner'])

# process multiple documents in parallel using the spaCy NLP library
for doc in nlp.pipe(train.Reviews.values, batch_size=1000, n_process=32):
    # extract text of each token in the document and create a list of tokens
    # creating tokens and removing email, urls and punctuations
    prepro1 = ' '.join([token.text for token in doc if not token.like_url if not token.like_email if not token.is_punct ])

    tokens = clean_text(prepro1)


    # add the list of tokens to the tokens_method2
    tokens_method2.append(tokens)

# add the tokens_method2 to the train_data dataframe as a new column 'tokens_method2'
train['cleaned_text'] = tokens_method2
```

```
<ipython-input-6-9bf6ecbde316>:6: MarkupResemblesLocatorWarning: The input looks more like a filename than markup. You may want to open
  soup = BeautifulSoup(input_text, "html.parser")
```

```
train.head()
```

	Reviews	Labels	cleaned_text
9387	This movie gets it right. As a former USAF Avi...	1	This movie gets it right As a former USAF Avia...
21353	Jonathan Rivers (Michael Keaton) suddenly beco...	0	Jonathan Rivers Michael Keaton suddenly become...
16267	This film, for an after school special, isn't ...	0	This film for an after school special is n't t...
24748	This movie is really nothing besides an admitt...	0	This movie is really nothing besides an admitt...
4833	The best Treasure Island ever made. They just ...	1	The best Treasure Island ever made They just d...

```
html_with_newlines = """
<html>
<head>
<title>Sample Page</title>
</head>
<body>
```

```
<p>This is a sample paragraph.</p>
<p>It has <b>HTML tags</b> and newline characters.</p>
</body>
</html>
"""

cleaned_text = clean_text(html_with_newlines)
print(cleaned_text)
```

```
     Sample Page This is a sample paragraph. It has HTML tags and newline characters.
```

```
col = train.columns.to_list()

ind_col = col.index('cleaned_text')
print(ind_col)
train.iloc[1, 2]
```

```
     2
     'Jonathan Rivers Michael Keaton suddenly becomes a widower when his wife dies Soon after he 's approached by a Dr. Price a
     n expert in Electronic Voice Phenomena EVP who claims he 's been receiving messages from Jonathan 's departed wife Anna vi
     a sundry electronic gadgets Is Anna trying to tell Jonathan something Is this merely a hint of something on a larger cosmi
     c or otherworldly scale It 's good to see Keaton in a leading role but the story he 's stuck with is convoluted and absurd
     at points it 's as if the movie does n't know how to answer any of the questions it brings up so it just distracts the vie
     wer with new unrelated questions Keaton himself is pretty good convincingly cast as the bereaving widower desperately tryi
```

```
from string import whitespace
m, n = train.shape

n_words = []
n_char = []
n_char_wo_space = []
avg_wrd_len = []
cnt_num = []

# temporarily disable the named pipes of spaCy NLP processing pipeline
disabled = nlp.select_pipes(
    disable=['tok2vec', 'tagger', 'parser', 'attribute_ruler', 'lemmatizer', 'ner'])

# # process multiple documents in parallel using the spaCy NLP library
# for doc in nlp.pipe(train.cleaned_text.values, batch_size=1000, n_process=32):
col = train.columns.to_list()

ind_col = col.index('cleaned_text')
nchar = 0

for i in range(m):
    text = train.iloc[i, ind_col]

    whitespace_tokens = text.split(' ')

    n_words.append(len([t for t in whitespace_tokens if not t.isdigit()]))

    n_char.append(len(text))

    n_char_wo_space.append(sum([len(t) for t in whitespace_tokens ]))

    avg_wrd_len.append(np.mean([len(t) for t in whitespace_tokens if not t.isdigit()]))

    cnt_num.append(len([t for t in whitespace_tokens if t.isdigit()]))



train['n_words'] = n_words
train['n_char'] = n_char
train['n_char_wo_space'] = n_char_wo_space
train['avg_wrd_len'] = avg_wrd_len
train['cnt_num'] = cnt_num
```

```
train.head()
```

	Reviews	Labels	cleaned_text	n_words	n_char	n_char_wo_space	avg_wrd_len	cnt_num
9387	This movie gets it right. As a former USAF Avi...	1	This movie gets it right As a former USAF Avia...	98	518	421	4.295918	0
21353	Jonathan Rivers (Michael Keaton) suddenly beco...	0	Jonathan Rivers Michael Keaton suddenly become...	310	1723	1413	4.551613	1

```
train.iloc[1, 0]
```

```
'Jonathan Rivers (Michael Keaton) suddenly becomes a widower when his wife dies. Soon after, he's approached by a Dr. Pric
e, an expert in Electronic Voice Phenomena (EVP), who claims he's been receiving messages from Jonathan's departed wife An
na via sundry electronic gadgets. Is Anna trying to tell Jonathan something? Is this merely a hint of something on a large
r cosmic or otherworldly scale? It's good to see Keaton in a leading role, but the story he's stuck with is convoluted and
absurd at points; it's as if the movie doesn't know how to answer any of the questions it brings up, so it just distracts
the viewer with new, unrelated questions <br /><br />Keaton himself is pretty good, convincingly cast as the bereaving wid
```

```
train.sort_values(['cnt_num'], ascending = False).iloc[0, 0]
```

```
'On Sunday July 27, 1997, the first episode of a new science fiction series called "Sta
rgate SG-1" was broadcast on Showtime. A spin-off of and sequel to the 1994 film "Starg
ate" starring Kurt Russell and James Spader, the series begins approximately one year a
fter the events portrayed in the film. For ten seasons, it chronicled the adventures an
d misadventures of an intrepid team of explorers known as SG-1. Originally, the series
starred Richard Dean Anderson as Colonel Jack O\'Neill (two "l"s!), Michael Shanks as D
```