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
```

```
2023-08-26 20:51:36.969035: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is optimiz
To enable the following instructions: AVX2 FMA, in other operations, rebuild TensorFlow with the appropriate compi
2023-08-26 20:51:38.347730: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find
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.
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 28.6 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
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy<
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy<
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=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
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=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.
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.10/dist-packages (fr
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en-cor
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.0->en
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.0,>=3.6.
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.7.
Requirement already satisfied: annotated-types>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1
Requirement already satisfied: pydantic-core==2.6.1 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8
Requirement already satisfied: typing-extensions>=4.6.1 in /usr/local/lib/python3.10/dist-packages (from pydantic!
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2.0,>=8
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.2
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.10.0,>
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<3.7.
✓ 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..

SOLUTION-

Mount Google Drive and read the files.

```
# mount google drive
if 'google.colab' in str(get_ipython()):
    from google.colab import drive
    drive.mount('/content/drive')
```

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

```
if 'google.colab' in str(get_ipython()):
  base_folder = Path('/content/drive/MyDrive/NLP')


df = base_folder /'train.csv'


train_data = pd.read_csv(df , index_col = 0)


train_data.head()
```

	Reviews	Labels
0	Ever wanted to know just how much Hollywood co...	1
1	The movie itself was ok for the kids. But I go...	1
2	You could stage a version of Charles Dickens' ...	1
3	this was a fantastic episode. i saw a clip fro...	1
4	and laugh out loud funny in many scenes. ...	1

Taking 10% subset of the data for the analysis

```
ten_perc_n = int(0.1*len(train_data))
train_final = train_data.sample(n = ten_perc_n, random_state = 0)
```

Cleaning the data - Removing HTML tags, new line characters, URLs, Emails and punctuations.

```
def cleaned_text(text):
  #Remove HTML tags
  clean_text = BeautifulSoup(text , "html.parser").get_text()
  #Remove new line characters and replace them with space
  clean_text = re.sub(r'[\n\r]',' ', clean_text)
  #Remove URLs:
  clean_text = re.sub(r'http\S+' , '' , clean_text)
  #Remove Emails:
  clean_text = re.sub(r'\S+@\S+' , '' , clean_text)
  #Remove punctuation:
  clean_text = re.sub(r'[^\w\s.]' , '' , clean_text)
  return clean_text



train_final['cleaned_text'] = train_final['Reviews'].apply(cleaned_text)
```

```
<ipython-input-21-a30455e1d274>:3: MarkupResemblesLocatorWarning: The input looks more like a filename than markup
  clean_text = BeautifulSoup(text , "html.parser").get_text()
```

Load the NLP model.

```
#Load the model:
nlp = spacy.load("en_core_web_sm")
```

Find number of words in each review

```
#Number of words:
def word_count(review):
  doc = nlp(review)
  word_num = sum(1 for token in doc if not token.is_space)
  return word_num
train_final['Number of words'] = train_final['cleaned_text'].apply(word_count)
```

Find number of characters in each review

```
#Number of characters:
def char_count(review):
  doc = nlp(review)
  char_num = len(doc.text)
  return char_num
train_final['Number of characters'] = train_final['cleaned_text'].apply(char_count)
```

Find number of characters without spaces in each review

```
#Number of characters without spaces:
def char_wo_space(review):
  doc = nlp(review)
  count = sum(len(token.text) for token in doc if not token.is_space and token.is_alpha)
  return count
train_final['Number of characters without spaces'] = train_final['cleaned_text'].apply(char_wo_space)
```

Find average word length in each review

```
#Average Word length:
def avg_word_len(review):
  doc = nlp(review)
  summ = sum(len(token.text) for token in doc if not token.is_space)
  value = sum(1 for token in doc if not token.is_space)
  average = int(summ/value)
  return average
train_final['Average Word Length'] = train_final['cleaned_text'].apply(avg_word_len)
```

Find count of numbers in each review

```
#Count of numbers:
def number_count(review):
  doc = nlp(review)
  num_count = sum(1 for token in doc if token.is_digit)
  return num_count
train_final['Count of numbers'] = train_final['cleaned_text'].apply(number_count)
```

Preview of the final cleaned and analysed data.

```
train_final.head()
```

	Reviews	Labels	cleaned_text	Number of words	Number of characters	Number of characters without spaces	Average Word Length
14149	I saw the MST3K version of "Deathstalker III" ...	0	I saw the MST3K version of Deathstalker III an...	170	816	632	3
8946	The visuals and effects are up to par with the...	1	The visuals and effects are up to par with the...	77	381	307	4
22378	Poor Paul Mercurio. After landing the role of ...	0	Poor Paul Mercurio. After landing the role of ...	70	376	306	4
	I was surprised at		I was surprised at				