Manish Bhardwaj

Roll no. 29

Section - 3CB

NLP Text Preprocessing Assignment

```
# Load Dataset
import pandas as pd
df = pd.read_csv('/content/New Text Document.txt')
df.head()
```

	text	label
0	Why do Java developers wear glasses? Because t...	1
1	I told my computer I needed a break… it froze.	1
2	Debugging is like being the detective in a cri...	1
3	Why did the neural network go to therapy? Too ...	1
4	My data went on a date… now it’s an outlier.	1

```
# Remove HTML Tags and URLs
import re
df['clean_text'] = df['text'].apply(lambda x: re.sub(r'<.*?>', '', x))
df['clean_text'] = df['clean_text'].apply(lambda x: re.sub(r'http\S+|www\S+', '', x))
df.head()
```

	text	label	clean_text
0	Why do Java developers wear glasses? Because t...	1	Why do Java developers wear glasses? Because t...
1	I told my computer I needed a break… it froze.	1	I told my computer I needed a break… it froze.
2	Debugging is like being the detective in a cri...	1	Debugging is like being the detective in a cri...
3	Why did the neural network go to therapy? Too ...	1	Why did the neural network go to therapy? Too ...
4	My data went on a date… now it’s an outlier.	1	My data went on a date… now it’s an outlier.

```
import sys
!{sys.executable} -m pip install emoji

# Emoji Handling
import emoji
df['clean_text'] = df['clean_text'].apply(lambda x: emoji.demojize(x))
df.head()
```

```
Collecting emoji
  Downloading emoji-2.15.0-py3-none-any.whl.metadata (5.7 kB)
Downloading emoji-2.15.0-py3-none-any.whl (608 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 608.4/608.4 kB 12.8 MB/s eta 0:00:00
Installing collected packages: emoji
Successfully installed emoji-2.15.0
```

	text	label	clean_text
0	Why do Java developers wear glasses? Because t...	1	Why do Java developers wear glasses? Because t...
1	I told my computer I needed a break… it froze.	1	I told my computer I needed a break… it froze.
2	Debugging is like being the detective in a cri...	1	Debugging is like being the detective in a cri...
3	Why did the neural network go to therapy? Too ...	1	Why did the neural network go to therapy? Too ...
4	My data went on a date… now it’s an outlier.	1	My data went on a date… now it’s an outlier.

```
# Lowercasing
df['clean_text'] = df['clean_text'].str.lower()
df.head()
```

	text	label	clean_text
0	Why do Java developers wear glasses? Because t...	1	why do java developers wear glasses? because t...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze.
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...
3	Why did the neural network go to therapy? Too ...	1	why did the neural network go to therapy? too ...
4	My data went on a date… now it’s an outlier.	1	my data went on a date… now it’s an outlier.

```
# Punctuation Removal
import string
df['clean_text'] = df['clean_text'].apply(lambda x: x.translate(str.maketrans('', '', string.punctuation)))
df.head()
```

	text	label	clean_text
0	Why do Java developers wear glasses? Because t...	1	why do java developers wear glasses because th...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...
3	Why did the neural network go to therapy? Too ...	1	why did the neural network go to therapy too m...
4	My data went on a date… now it’s an outlier.	1	my data went on a date… now it’s an outlier

```
# Chat Normalization
slang = {'u':'you','ur':'your','pls':'please','lol':'laughing','idk':'i don’t know'}

df['clean_text'] = df['clean_text'].apply(lambda x: ' '.join([slang.get(w, w) for w in x.split()]))
df.head()
```

	text	label	clean_text
0	Why do Java developers wear glasses? Because t...	1	why do java developers wear glasses because th...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...
3	Why did the neural network go to therapy? Too ...	1	why did the neural network go to therapy too m...
4	My data went on a date… now it’s an outlier.	1	my data went on a date… now it’s an outlier

```
import sys
!{sys.executable} -m pip install textblob

# Spelling Correction with TextBlob
from textblob import TextBlob

df['clean_text'] = df['clean_text'].apply(lambda x: str(TextBlob(x).correct()))
display(df.head())
```

```
Requirement already satisfied: textblob in /usr/local/lib/python3.12/dist-packages (0.19.0)
Requirement already satisfied: nltk>=3.9 in /usr/local/lib/python3.12/dist-packages (from textblob) (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packages (from nltk>=3.9->textblob) (8.3.0)
Requirement already satisfied: joblib in /usr/local/lib/python3.12/dist-packages (from nltk>=3.9->textblob) (1.5.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.12/dist-packages (from nltk>=3.9->textblob) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (from nltk>=3.9->textblob) (4.67.1)
```

	text	label	clean_text
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...
3	Why did the neural network go to therapy? Too ...	1	why did the neutral network go to therapy too ...
4	My data went on a date… now it’s an outlier.	1	my data went on a date… now it’s an outer

```
# Tokenization
import nltk
nltk.download('punkt')
nltk.download('punkt_tab') # Download the missing resource
from nltk.tokenize import word_tokenize
```

```
df['tokens'] = df['clean_text'].apply(word_tokenize)
df.head()
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

	text	label	clean_text	tokens
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...	[why, do, cava, developer, wear, glasses, beca...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze	[i, told, my, computer, i, needed, a, break…, ...
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...	[debugging, is, like, being, the, detective, i...

```
# Stopword Removal
nltk.download('stopwords')
from nltk.corpus import stopwords
sw = set(stopwords.words('english'))
df['tokens'] = df['tokens'].apply(lambda x: [w for w in x if w not in sw])
df.head()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

	text	label	clean_text	tokens
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...	[cava, developer, wear, glasses, ', c]
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze	[told, computer, needed, break…, froze]
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...	[debugging, like, detective, crime, movie, als...
3	Why did the neural network go to therapy? Too	1	why did the neutral network go to therapy too	[neutral, network, go, therapy, many,

```
# Stemming
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
df['stemmed'] = df['tokens'].apply(lambda x: [stemmer.stem(w) for w in x])
df.head()
```

	text	label	clean_text	tokens	stemmed
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...	[cava, developer, wear, glasses, ', c]	[cava, develop, wear, glass, ', c]
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze	[told, computer, needed, break…, froze]	[told, comput, need, break…, froze]
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...	[debugging, like, detective, crime, movie, als...	[debug, like, detect, crime, movi, also, murder]
3	Why did the neural network go to therapy? Too ...	1	why did the neutral network go to therapy too ...	[neutral, network, go, therapy, many, resolved...	[neutral, network, go, therapi, mani, resolv, ...
4	My data went on a date… now it's an outlier.	1	my data went on a date… now it's an outer	[data, went, date…, ', outer]	[data, went, date…, ', outer]

```
# Lemmatization
nltk.download('wordnet')
nltk.download('omw-1.4')
from nltk.stem import WordNetLemmatizer
lemm = WordNetLemmatizer()
df['lemmatized'] = df['tokens'].apply(lambda x: [lemm.lemmatize(w) for w in x])
df.head()
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
```

	text	label	clean_text	tokens	stemmed	lemmatized
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...	[cava, developer, wear, glasses, ', c]	[cava, develop, wear, glass, ', c]	[cava, developer, wear, glass, ', c]
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze	[told, computer, needed, break…, froze]	[told, comput, need, break…, froze]	[told, computer, needed, break…, froze]
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...	[debugging, like, detective, crime, movie, als...	[debug, like, detect, crime, movi, also, murder]	[debugging, like, detective, crime, movie, als...
3	Why did the neural network go to therapy? Too ...	1	why did the neutral network go to therapy too ...	[neutral, network, go, therapy, many, resolved...	[neutral, network, go, therapi, mani, resolv, ...	[neutral, network, go, therapy, many, resolved...
4	My data went on a date…	1	my data went on a date…	[data, went, date…, ',	[data, went, date…, ',	[data, went, date…, ',

```
# POS Tagging
import nltk
nltk.download('averaged_perceptron_tagger_eng') # Download the specific English tagger
from nltk.tokenize import word_tokenize
df['pos'] = df['clean_text'].apply(lambda x: nltk.pos_tag(word_tokenize(x)))
df.head()
```

```
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger_eng.zip.
```

	text	label	clean_text	tokens	stemmed	lemmatized	pos
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...	[cava, developer, wear, glasses, ', c]	[cava, develop, wear, glass, ', c]	[cava, developer, wear, glass, ', c]	[(why, WRB), (do, VBP), (cava, VB), (developer...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze	[told, computer, needed, break…, froze]	[told, comput, need, break…, froze]	[told, computer, needed, break…, froze]	[(i, NN), (told, VBD), (my, PRP$), (computer, ...
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...	[debugging, like, detective, crime, movie, als...	[debug, like, detect, crime, movi, also, murder]	[debugging, like, detective, crime, movie, als...	[(debugging, NN), (is, VBZ), (like, IN), (bein...
3	Why did the neural network go to	1	why did the neutral network go to therapy	[neutral, network, go, therapy, many,	[neutral, network, go, therapi, mani,	[neutral, network, go, therapy, many,	[(why, WRB), (did, VBD), (the, DT),

```
import spacy
nlp = spacy.load('en_core_web_sm')

def parse_sentence_to_list(text):
    doc = nlp(text)
    parsed_tokens = []
    for token in doc:
        parsed_tokens.append({
            'text': token.text,
            'dep': token.dep_,
            'head': token.head.text,
            'pos': token.pos_
        })
    return parsed_tokens

df['dependency_parsed'] = df['clean_text'].apply(parse_sentence_to_list)
display(df.head())
```

	text	label	clean_text	tokens	stemmed	lemmatized	pos	dependency_parsed
0	Why do Java developers wear glasses? Because t...	1	why do cava developer wear glasses because the...	[cava, developer, wear, glasses, ', c]	[cava, develop, wear, glass, ', c]	[cava, developer, wear, glass, ', c]	[(why, WRB), (do, VBP), (cava, VB), (developer...	[{'text': 'why', 'dep': 'advmod', 'head': 'wea...
1	I told my computer I needed a break… it froze.	1	i told my computer i needed a break… it froze	[told, computer, needed, break…, froze]	[told, comput, need, break…, froze]	[told, computer, needed, break…, froze]	[(i, NN), (told, VBD), (my, PRP$), (computer, ...	[{'text': 'i', 'dep': 'nsubj', 'head': 'told',...
2	Debugging is like being the detective in a cri...	1	debugging is like being the detective in a cri...	[debugging, like, detective, crime, movie, als...	[debug, like, detect, crime, movi, also, murder]	[debugging, like, detective, crime, movie, als...	[(debugging, NN), (is, VBZ), (like, IN), (bein...	[{'text': 'debugging', 'dep': 'nsubj', 'head':...
	Why did the				[neutral,			