In [7]:

```
import pandas as pd
import sklearn as sklearn
import numpy as np
import nltk
```

In [29]:

```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/pranavbhadharla/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

Out[29]:

```
True
```

In [8]:

```
# reading in reddit data
df=pd.read_csv('REDDIT_COMMENTS_2022-09-10T16-48-26-0400 (2).csv')
```

In [11]:

```
# df of the comments which holds the most text data
comments=df["comment_body"]
```

In [21]:

```
# making sure the data types are correct
type(comments[0])
```

Out[21]:

```
str
```

In [32]:

```
# tokenizing the comments by word
words=[]
for i in range(len(comments)):
    words.append(nltk.word_tokenize(comments[i]))
```

In [54]:

```
from nltk.stem import *
from nltk.stem import *
```

In [55]:

```
# setting the function
stemmer = PorterStemmer()
```

In [61]:

```
# stemming the words one by one
stemmed_words=[]
for i in range(len(words)):
    for j in range(len(words[i])):
        stemmed_words.append(stemmer.stem(words[i][j]))
```

In [44]:

```
from nltk.stem import WordNetLemmatizer
import nltk
nltk.download('wordnet')
lemmatizer = WordNetLemmatizer()
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/pranavbhadharla/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     /Users/pranavbhadharla/nltk_data...
```

Out[44]:

```
True
```

In [73]:

```
# lemmatizing words one by one
lemmatized_words=[]
for i in range(len(stemmed_words)):
        lemmatized_words.append(lemmatizer.lemmatize(stemmed_words[i]))
```

In [77]:

```
# checking output and turning it into a csv
type(lemmatized_words)
np.savetxt("nlp.csv", lemmatized_words, delimiter=", ", fmt="% s")
```

In [80]:

```
# write up in seperate pdf file
```

In [ ]: