Lemmatization (Annapoornima S)

```
from zipfile import ZipFile
import glob
import pandas as pd
import nltk
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import linear_kernel
from nltk.corpus import stopwords
import warnings
warnings.filterwarnings('ignore')
```

```
file_name = "movies.zip"
with ZipFile(file_name, 'r') as zip:
    zip.printdir()
```

How many sentences$_{each}$

How many tokens$_{each}$

```
files = [file for file in glob.glob("movies/*")]
for file in files:
    with open(file, 'r', encoding='cp1252') as f:
        contents = f.readlines()
        for row1 in contents:
            words = nltk.word_tokenize(row1)
        print("word tokenize ", len(words))
```

```
word181
     word            119
word20
     word            276
word
     word            70
word49
     word            98
word242
     word            67
word131
     word
word69
     word            66
word39
     word
word50
     word            208
word100
     word            569
```

each

stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
stopwords
365365365365365365365365365stopwords365365365365365365365365365

D. How many unique stems each (Use PorterStemmer)

```
def port_stemSentence(sentence):
    tokenizer = nltk.tokenize.WhitespaceTokenizer()
    tok = tokenizer.tokenize(sentence)
    filtered_sentence = [w for w in tok if not w in stop_words]
    stem_sentence = []
    for word in filtered_sentence:
        stem_sentence.append(ps.stem(word))
    return len(stem_sentence)
```

```
    porter_stemming
    96
porter_stemming
83
    porter_stemming
    20
porter_stemming
138
    porter_stemming
    63
porter_stemming
64
    porter_stemming
    20
porter_stemming
51
    porter_stemming
    131
porter_stemming
27
    porter_stemming
    53
porter_stemming
87
    porter_stemming
    35
porter_stemming
93
    porter_stemming
    23
porter_stemming
34
    porter_stemming
    52
porter_stemming
38
    porter_stemming
    33
porter_stemming
282
```

each(Use

```
def lan_stemSentence(sentence):
    tokenizer = nltk.tokenize.WhitespaceTokenizer()
    tok = tokenizer.tokenize(sentence)
    filtered_sentence = [w for w in tok if not w in stop_words]
    stem_sentence = []
    for word in filtered_sentence:
        stem_sentence.append(ls.stem(word))
    return len(stem_sentence)
```

F. How many unique words each(Use lemmatization)WordNetLemmatizer)

```
files = [file for file in glob.glob("movies/*")]
for file in files:
    with open(file, 'r', encoding='cp1252') as f:
        contents = f.readline()
        print("lancaster_stemming ")
        print(port_stemSentence(contents))
```

```
96
83
20

63
64
20
51

27

87


23
34

38

282
```

```
import nltk
nltk.download('wordnet')
```

```
True
```

```
def lemmSentence(sentence):
    tokenizer = nltk.tokenize.WhitespaceTokenizer()
    tok = tokenizer.tokenize(sentence)
    filtered_sentence = [w for w in tok if not w in stop_words]
    lemm_sentence = []
    for word in filtered_sentence:
        lemm_sentence.append(lemmatizer.lemmatize(word))
    return len(lemm_sentence)
```

```
for file in files:
    with open(file, 'r', encoding='cp1252') as f:
        contents = f.readline()
        print("lemmatization ")
        print(lemmSentence(contents))
```

```
lemmatization
    96
lemmatization
    83
lemmatization
    20
lemmatization

lemmatization
    63
lemmatization
    64
lemmatization
    20
lemmatization
    51
lemmatization
```

```
lemmatization
    27
lemmatization

lemmatization
    87
lemmatization

lemmatization

lemmatization
    23
lemmatization
    34
lemmatization

lemmatization
    38
lemmatization

lemmatization
    282
```

Step-1: Foreach

Tokenize terms and of

lemmatized words from the tokens

```
True
```

Takeofanytwoand

0.57735026918962580.57735026918962580.57735026918962580.5773502691896258

```
                  (108,7)
                  (118,5)
                  (121,13)
                  (124,12)
                  (128,6)
                  (134,10)
                  (138,15)
                  (143,15)
                  (148,7)
                  (152,1)
                  (154,1)
                  (156,1)
                  (165,9)
                  (166,0)
                  (172,4)
                  (173,2)
                  (174,8)
                  (177,10)
                  (179,3)
                  (180,0)
                  (188,20)
                  (193,7)
0.5773502691896258(194,11)
```

```
with open(files[5],'r',encoding='cp1252')as f:
    contents = f.read()
    tok = tokenizer.tokenize(contents)
    filtered_sentence = [w for w in tok if not w in stop_words]
    tfidf = TfidfVectorizer(min_df=2,max_df=0.5,ngram_range=(1,2))
    movie1 = tfidf.fit_transform(filtered_sentence)
    print(movie1)
```

```
(196,
(203,
```

```
(384,
(385,
```

```
doc1 = movie1[0:10]
doc2 = movie1[:]
score = linear_kernel(doc1,doc2)
print(score)
```