CIS 668 Assignment #3

Sentiment Analysis of Amazon Product Reviews

Student: *Leah Luo (NetID lluojr)*

Date: *04/10/2020*

Content

CIS 668
Leah Luo (NetID lluojr)
Assignment #3
04/10/2020

Data Pre-processing

1. **Retrieve the data needed.**

a. Extract the text content from the text file. (rawtextSplit)
The second figure shows the first 20 words in the original file.

```
import nltk
import re
from nltk import FreqDist
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.corpus import wordnet as wn
from nltk.corpus import PlaintextCorpusReader
from nltk.tokenize import word_tokenize
from collections import defaultdict
from nltk.collocations import *

#Open the original file
path = "/Users/LXIN/Desktop/T/clothing_shoes_jewelry.txt"
textfile = open(path,"r")
rawtext = textfile.read()
textfile.close()
rawtextSplit = rawtext.splitlines()
#Original text content
print(rawtextSplit[:20])
```

```
['reviewerID:A1KLRMWW2FWPL4', 'asin:0000031887', 'reviewerName:Amazon Customer "cameramom"', 'helpful:[0, 0]', "rev
iewText:This is a great tutu and at a really great price. It doesn't look cheap at all. I'm so glad I looked on Ama
zon and found such an affordable tutu that isn't made poorly. A++", 'overall:5.0', 'summary:Great tutu-  not cheapl
y made', 'unixReviewTime:1297468800', 'reviewTime:02 12, 2011', '', 'reviewerID:A2G5TCU2WDFZ65', 'asin:0000031887',
'reviewerName:Amazon Customer', 'helpful:[0, 0]', 'reviewText:I bought this for my 4 yr old daughter for dance clas
s, she wore it today for the first time and the teacher thought it was adorable. I bought this to go with a light b
lue long sleeve leotard and was happy the colors matched up great. Price was very good too since some of these go f
or over $15.00 dollars.', 'overall:5.0', 'summary:Very Cute!!', 'unixReviewTime:1358553600', 'reviewTime:01 19, 201
3', '']
```

b. Create a document to save the extracted text, which contains only the reviews contents (reviews.txt)
The second figure shows the first 20 words in the new file.

```
#Extract only the reviewText from the file
def extract(files):
    f = open("/Users/LXIN/Desktop/T/reviews.txt",'w+')
    for var in files:
        if "reviewText" in var:
            varwrite = re.sub("reviewText:", "", var)
            f.write(varwrite+"\n")
    f.close()
extract(rawtextSplit)
```

```
textfile2 = open("/Users/LXIN/Desktop/T/reviews.txt")
reviewText = textfile2.read()
print(reviewText[:20])
```

```
This is a great tutu
```

2. **Pre-processing the data**

a. Tokenization. Open the new text file *reviews*. Separate the file content into tokens with sentence tokenizer.
The second figure shows the first 20 tokenized words in the file.

```
#Tokenize
from nltk import tokenize
tokensen = tokenize.sent_tokenize(reviewText)
print(len(tokensen))
print(tokensen[:20])
```

```
1140642
['This is a great tutu and at a really great price.', "It doesn't look cheap at all.", "I'm so glad I looked on Amazo
n and found such an affordable tutu that isn't made poorly.", 'A++\nI bought this for my 4 yr old daughter for dance
class, she wore it today for the first time and the teacher thought it was adorable.', 'I bought this to go with a li
ght blue long sleeve leotard and was happy the colors matched up great.', 'Price was very good too since some of thes
e go for over $15.00 dollars.', 'What can I say... my daughters have it in orange, black, white and pink and I am thi
nking to buy for they the fuccia one.', 'It is a very good way for exalt a dancer outfit: great colors, comfortable,
looks great, easy to wear, durables and little girls love it.', 'I think it is a great buy for costumer and play to
o.', 'We bought several tutus at once, and they are got high reviews.', 'Sturdy and seemingly well-made.', 'The girls
have been wearing them regularly, including out to play, and the tutus have stood up well.', 'Fits the 3-yr old & the
5-yr old well.', 'Clearly plenty of room to grow.', 'Only con is that when the kids pull off the tutus, the waste ban
d gets twisted, and an adult has to un-tangle.', 'But this is not difficult.', 'Thank you Halo Heaven great product f
or Little Girls.', "My Great Grand Daughters Love these Tutu's.", 'Will buy more from this seller.', 'Made well and c
ute on the girls.']
```

b. Note that since we are going to explore the sentiment of the comments at sentence level, we won't do much pre-processing for the Amazon reviews. Instead, we will have a new step for this assignment, which is creating the word feature for training and testing.

Word Set

We are going to download the corpus and do the pre-processing before defining feature sets.

a. Download and load the sentence_polarity corpus.
Note: this corpus of sentences are from the Movie Review corpus, and all the sentences are already labeled with tags ('positive' or 'negative').

```
import nltk
# nltk.download('sentence_polarity')
from nltk.corpus import sentence_polarity
import random
sentences = sentence_polarity.sents()

print(len(sentences))
print(sentence_polarity.categories())
```

```
10662
['neg', 'pos']
```

b. Create a list of documents and each document is a sentence with the words and the label ('positive' or 'negative').

```
from nltk.corpus import sentence_polarity
import random

documents = [(sent, cat) for cat in sentence_polarity.categories()
             for sent in sentence_polarity.sents(categories=cat)]
print(documents[:2])
```

```
[(['simplistic', ',', 'silly', 'and', 'tedious', '.'], 'neg'), (["it's", 'so', 'laddish', 'and', 'juvenile', ',', 'on
ly', 'teenage', 'boys', 'could', 'possibly', 'find', 'it', 'funny', '.'], 'neg')]
```

c. Import random and use it to mix up the documents.
Why: The documents are sorted by label, we will need to mix them up so that both training and test sets have sentences from two categories.

```
random.shuffle(documents)
print(documents[0])
```

```
(['if', 'hill', "isn't", 'quite', 'his', "generation's", 'don', 'siegel', '(', 'or', 'robert', 'aldrich', ')', ',',
"it's", 'because', "there's", 'no', 'discernible', 'feeling', 'beneath', 'the', 'chest', 'hair', ';', "it's", 'all',
'bluster', 'and', 'cliché', '.'], 'neg')
```

d. Create a list named "all_words_list", which includes all the words in the document collection we created in last step.

```
all_words_list = [word for (sent,cat) in documents for word in sent]

print(all_words_list[:10])
print(len(all_words_list))
```

```
['still', 'rapturous', 'after', 'all', 'these', 'years', ',', 'cinema', 'paradiso', 'stands']
224073
```

e. Apply the lower() and isalpha() functions to the word list, to lower all the characters that are alphabetic.

```
#Fileters: isalpha() and lower()
wordLower = [w for w in all_words_list if w.isalpha()]
print(len(wordLower))
print(wordLower[:20])
```

```
187486
['still', 'rapturous', 'after', 'all', 'these', 'years', 'cinema', 'paradiso', 'stands', 'as', 'one', 'of', 'the', 'g
reat', 'films', 'about', 'movie', 'love', 'be', 'left']
```

Sentiment Classification

Note: Two different feature sets will be defined in this section, for the sentiment classification.

1. **Feature Set #1 (Subjectivity Count Features)**

a. Remove all the stopwords and define it as a new words-set.
   Why: to prune the word features

```
# Set 1: Remove stopwords
stopwords = nltk.corpus.stopwords.words('english')
wordRmStop = [w for w in wordLower if not w in stopwords]
print("----------------------------------------------------")
print(len(wordRmStop))
print(wordRmStop[:20])
```

```
----------------------------------------------------
105085
['still', 'rapturous', 'years', 'cinema', 'paradiso', 'stands', 'one', 'great', 'films', 'movie', 'love', 'left', 'se
nsation', 'witnessed', 'great', 'performance', 'perhaps', 'give', 'urge', 'get']
```

b. Call FreqDist to limit the collection to 2000 most frequent words.

```
# word_features 1 (without stopwords)
all_words = nltk.FreqDist(wordRmStop)
word_items = all_words.most_common(2000)
word_features = [word for (word,count) in word_items]
print(word_features[:20])
```

```
['film', 'movie', 'one', 'like', 'story', 'much', 'even', 'good', 'comedy', 'time', 'characters', 'little', 'way', 'f
unny', 'make', 'enough', 'never', 'makes', 'may', 'us']
```

c. I'll choose to read in the subjectivity words from the subjectivity lexicon file first. I'll create two features that involve counting the positive and negative subjectivity words present.
   I'll use just the words (also called "unigram features").
   Create a path variable, copy and paste the definition of the readSbujectivity function from the Subjectivity.py module.
   Note: this subjectivity lexicon file is created by Janyce Wiebe and her group at the University of Pittsburgh in MPQA project.

```
# Feature 1
SLpath = 'subjclueslen1-HLTEMNLP05.tff'
def readSubjectivity(path):
    flexicon = open(path, 'r')
    # initialize an empty dictionary
    sldict = { }
    for line in flexicon:
        fields = line.split()   # default is to split on whitespace
        # split each field on the '=' and keep the second part as the value
        strength = fields[0].split("=")[1]
        word = fields[2].split("=")[1]
        posTag = fields[3].split("=")[1]
        stemmed = fields[4].split("=")[1]
        polarity = fields[5].split("=")[1]
        if (stemmed == 'y'):
            isStemmed = True
        else:
            isStemmed = False
        # put a dictionary entry with the word as the keyword
        #     and a list of the other values
        sldict[word] = [strength, posTag, isStemmed, polarity]
    return sldict
SL = readSubjectivity(SLpath)
```

d. Define a function named "SL_features" to extract all the words has two features "positivecount" and "negativecount". The "positivecount" and "negativecount" features counts for all the positive and negative subjectivity words. Counting method differs, depending on how strongly the subjective word is.

```python
def SL_features(document, word_features, SL):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    # count variables for the 4 classes of subjectivity
    weakPos = 0
    strongPos = 0
    weakNeg = 0
    strongNeg = 0
    for word in document_words:
        if word in SL:
            strength, posTag, isStemmed, polarity = SL[word]
            if strength == 'weaksubj' and polarity == 'positive':
                weakPos += 1
            if strength == 'strongsubj' and polarity == 'positive':
                strongPos += 1
            if strength == 'weaksubj' and polarity == 'negative':
                weakNeg += 1
            if strength == 'strongsubj' and polarity == 'negative':
                strongNeg += 1
            features['positivecount'] = weakPos + (2 * strongPos)
            features['negativecount'] = weakNeg + (2 * strongNeg)
    return features
```

e. Apply the feature extraction function to all the words.

```python
SL_featuresets = [(SL_features(d, word_features, SL), c) for (d, c) in documents]
```

f. Create the training and test sets, train the Naïve Bayes classifier, and get the accuracy. The length of the documents is around 10500.

```python
train_set, test_set = SL_featuresets[1000:], SL_featuresets[:1000]
classifier = nltk.NaiveBayesClassifier.train(train_set)
nltk.classify.accuracy(classifier, test_set)
# classifier.show_most_informative_features(30)
```

```
0.76
```

g. Call the show_most_informative_features function to show the top 30 ranked features, according to the ratio of one label ('pos' / 'neg') to another.

```python
classifier.show_most_informative_features(30)
```

```
Most Informative Features
         contains(boring) = True              neg : pos    =     18.9 : 1.0
     contains(engrossing) = True              pos : neg    =     18.4 : 1.0
         contains(stupid) = True              neg : pos    =     18.3 : 1.0
       contains(provides) = True              pos : neg    =     17.7 : 1.0
       contains(mediocre) = True              neg : pos    =     16.3 : 1.0
      contains(inventive) = True              pos : neg    =     15.7 : 1.0
           contains(flat) = True              neg : pos    =     13.8 : 1.0
        contains(generic) = True              neg : pos    =     13.6 : 1.0
     contains(refreshing) = True              pos : neg    =     13.0 : 1.0
        contains(routine) = True              neg : pos    =     13.0 : 1.0
           contains(warm) = True              pos : neg    =     12.6 : 1.0
      contains(wonderful) = True              pos : neg    =     11.8 : 1.0
       contains(haunting) = True              pos : neg    =     11.7 : 1.0
   contains(refreshingly) = True              pos : neg    =     11.7 : 1.0
       contains(captures) = True              pos : neg    =     11.4 : 1.0
      contains(realistic) = True              pos : neg    =     11.0 : 1.0
           contains(ages) = True              pos : neg    =     10.4 : 1.0
    contains(mesmerizing) = True              pos : neg    =     10.4 : 1.0
       contains(mindless) = True              neg : pos    =     10.3 : 1.0
      contains(offensive) = True              neg : pos    =     10.3 : 1.0
           contains(dull) = True              neg : pos    =     10.0 : 1.0
            contains(wry) = True              pos : neg    =      9.7 : 1.0
          contains(bears) = True              neg : pos    =      9.6 : 1.0
       contains(powerful) = True              pos : neg    =      9.2 : 1.0
        contains(playful) = True              pos : neg    =      9.0 : 1.0
       contains(intimate) = True              pos : neg    =      9.0 : 1.0
       contains(chilling) = True              pos : neg    =      9.0 : 1.0
     contains(unexpected) = True              pos : neg    =      9.0 : 1.0
       contains(tiresome) = True              neg : pos    =      9.0 : 1.0
           contains(loud) = True              neg : pos    =      9.0 : 1.0
```

h. Precision, Recall and F-measure score on test-set

```
# Build the reference and test lists from the classifier on the test set:
reflist = []
testlist = []
for (features, label) in test_set:
    reflist.append(label)
    testlist.append(classifier.classify(features))
```

```
reflist[:30]
testlist[:30]

reffemale = set([i for i,label in enumerate(reflist) if label == 'pos'])
refmale = set([i for i,label in enumerate(reflist) if label == 'neg'])

testfemale = set([i for i,label in enumerate(testlist) if label == 'pos'])
testmale = set([i for i,label in enumerate(testlist) if label == 'neg'])
```

```
from nltk.metrics import *

# compute precision, recall and F-measure for each label
def printmeasures(label, refset, testset):
    print(label, 'precision:', precision(refset, testset))
    print(label, 'recall:', recall(refset, testset))
    print(label, 'F-measure:', f_measure(refset, testset))

printmeasures('pos', reffemale, testfemale)
print("-------------------------------------")
printmeasures('neg', refmale, testmale)
```

```
pos precision: 0.7718940936863544
pos recall: 0.747534516765286
pos F-measure: 0.7595190380761523
-------------------------------------
neg precision: 0.7485265225933202
neg recall: 0.7728194726166329
neg F-measure: 0.7604790419161678
```

2. Feature Set #2 (Negation Features)

a. Create a list of negation words.

```
# Feature 2 (Not Feature)

# Negation Words
# this list of negation words includes some "approximate negators" like hardly and rarely
negationwords = ['no', 'not', 'never', 'none', 'nowhere', 'nothing', 'noone', 'rather', 'hardly', 'scarcely',
                 'rarely', 'seldom', 'neither', 'nor','ain', 'aren', 'couldn', 'didn', 'doesn', 'hadn', 'hasn',
                 'haven','isn', 'ma', 'mightn', 'mustn', 'needn', 'shan', 'shouldn', 'wasn', 'weren', 'won', 'wouldn']
```

b. Remove stopwords, but reserve the negation words (or parts of words)
Why: the negation filter will need them

```
# Remove stop words
newstopwords = [word for word in stopwords if word not in negationwords]
stop_words_list = [word for word in wordLower if word not in newstopwords]
len(stop_words_list)
```

```
106321
```

c. Call FreqDist to limit the collection to 2000 most frequent words.

```
all_words = nltk.FreqDist(stop_words_list)
word_items = all_words.most_common(2000)
word_features = [word for (word,count) in word_items]
print(word_features[:20])
```

```
['film', 'movie', 'not', 'one', 'like', 'story', 'no', 'much', 'even', 'good', 'comedy', 'time', 'characters', 'littl
e', 'way', 'funny', 'make', 'enough', 'never', 'makes']
```

d. Define the feature function named "NOT_features". Two feature sets are defined:
2000 word features and 2000 Not word features sets.

How it works: if a negation occurs, add the following word as a Not word feature; else add it as a regular feature word.

```python
def NOT_features(document, word_features, negationwords):
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = False
        features['contains(NOT{})'.format(word)] = False
    # go through document words in order
    for i in range(0, len(document)):
        word = document[i]
        if ((i + 1) < len(document)) and ((word in negationwords) or (word.endswith("n't"))):
            i += 1
            features['contains(NOT{})'.format(document[i])] = (document[i] in word_features)
        else:
            features['contains({})'.format(word)] = (word in word_features)
    return features
```

e. Apply the NOT_features extraction function to words.

```python
# Not Feature Sets
NOT_featuresets = [(NOT_features(d, word_features, negationwords), c) for (d, c) in documents]
NOT_featuresets[0][0]['contains(NOTlike)']
```

```
False
```

f. Create the training and test sets, train the Naïve Bayes classifier, and get the accuracy.

```python
# Accuracy
train_set3, test_set3 = NOT_featuresets[1000:], NOT_featuresets[:1000]
classifier3 = nltk.NaiveBayesClassifier.train(train_set3)
nltk.classify.accuracy(classifier3, test_set3)
```

```
0.791
```

g. Call the show_most_informative_features function to show the top 30 ranked features, according to the ratio of one label (‘pos’ / ‘neg’) to another.

```python
# Most informative features
classifier3.show_most_informative_features(30)
```

```
Most Informative Features
    contains(engrossing) = True              pos : neg    =     19.1 : 1.0
      contains(captures) = True              pos : neg    =     17.7 : 1.0
      contains(mediocre) = True              neg : pos    =     16.3 : 1.0
       contains(generic) = True              neg : pos    =     14.9 : 1.0
       contains(routine) = True              neg : pos    =     14.9 : 1.0
          contains(flat) = True              neg : pos    =     13.7 : 1.0
          contains(imax) = True              pos : neg    =     13.1 : 1.0
    contains(refreshing) = True              pos : neg    =     13.1 : 1.0
      contains(powerful) = True              pos : neg    =     12.5 : 1.0
          contains(dull) = True              neg : pos    =     12.4 : 1.0
     contains(inventive) = True              pos : neg    =     12.4 : 1.0
        contains(boring) = True              neg : pos    =     12.4 : 1.0
     contains(wonderful) = True              pos : neg    =     12.3 : 1.0
          contains(warm) = True              pos : neg    =     12.3 : 1.0
  contains(refreshingly) = True              pos : neg    =     11.7 : 1.0
         contains(stale) = True              neg : pos    =     11.6 : 1.0
     contains(realistic) = True              pos : neg    =     11.0 : 1.0
        contains(stupid) = True              neg : pos    =     11.0 : 1.0
   contains(mesmerizing) = True              pos : neg    =     10.4 : 1.0
          contains(ages) = True              pos : neg    =     10.4 : 1.0
      contains(delicate) = True              pos : neg    =     10.4 : 1.0
     contains(NOTenough) = True              neg : pos    =     10.3 : 1.0
      contains(provides) = True              pos : neg    =     10.2 : 1.0
           contains(wry) = True              pos : neg    =      9.7 : 1.0
    contains(apparently) = True              neg : pos    =      9.6 : 1.0
        contains(unless) = True              neg : pos    =      9.6 : 1.0
      contains(mindless) = True              neg : pos    =      9.6 : 1.0
      contains(intimate) = True              pos : neg    =      9.0 : 1.0
      contains(chilling) = True              pos : neg    =      9.0 : 1.0
        contains(tender) = True              pos : neg    =      9.0 : 1.0
```

h. Precision, Recall and F-measure score on test-set

```python
# Build the reference and test lists from the classifier on the test set:
reflist = []
testlist = []
for (features, label) in test_set3:
    reflist.append(label)
    testlist.append(classifier3.classify(features))
```

```python
reflist[:30]
testlist[:30]

reffemale = set([i for i,label in enumerate(reflist) if label == 'pos'])
refmale = set([i for i,label in enumerate(reflist) if label == 'neg'])

testfemale = set([i for i,label in enumerate(testlist) if label == 'pos'])
testmale = set([i for i,label in enumerate(testlist) if label == 'neg'])
```

```python
from nltk.metrics import *

# compute precision, recall and F-measure for each label
def printmeasures(label, refset, testset):
    print(label, 'precision:', precision(refset, testset))
    print(label, 'recall:', recall(refset, testset))
    print(label, 'F-measure:', f_measure(refset, testset))

printmeasures('pos', reffemale, testfemale)
print("--------------------------------------")
printmeasures('neg', refmale, testmale)
```

```
pos precision: 0.8003992015968064
pos recall: 0.7862745098039216
pos F-measure: 0.7932739861523244
-------------------------------------
neg precision: 0.781563126252505
neg recall: 0.7959183673469388
neg F-measure: 0.788675429726997
```

Sentiment Analysis

*Note: From the previous section we find out that **Subjectivity Count Features is less accurate than Negation Features**. But I'll apply both the Subjectivity Count Feature and Negation Feature on the Amazon reviews data.*

1. **Subjectivity Count Feature**
   a. Running result.
   There are 463424 sentences tagged as negative and 444152 sentences tagged as positive.

```
print("neg--------------------------------")
print(len(neg2))
print(neg2[:5])

print("\npos--------------------------------")
print(len(pos2))
print(pos2[:5])
```

```
neg--------------------------------
463424
["It doesn't look cheap at all.", "I'm so glad I looked on Amazon and found such an affordable tutu that isn't made p
oorly.", 'Price was very good too since some of these go for over $15.00 dollars.,What can I say... my daughters have
it in orange, black, white and pink and I am thinking to buy for they the fuccia one.', 'I think it is a great buy fo
r costumer and play too.,We bought several tutus at once, and they are got high reviews.', 'Sturdy and seemingly well
-made.']

pos--------------------------------
444152
['This is a great tutu and at a really great price.', 'A++,I bought this for my 4 yr old daughter for dance class, sh
e wore it today for the first time and the teacher thought it was adorable.', 'I bought this to go with a light blue
long sleeve leotard and was happy the colors matched up great.', 'It is a very good way for exalt a dancer outfit: gr
eat colors, comfortable, looks great, easy to wear, durables and little girls love it.', 'The girls have been wearing
them regularly, including out to play, and the tutus have stood up well.']
```

   b. Store the two sets of sentences in two files.

```
# pos into file
posFile = open('/Users/LXIN/Desktop/posSL.txt', 'w')
for r in posSL:
    posFile.write(r + '\n')
posFile.close()

# neg into file
negFile = open('/Users/LXIN/Desktop/negSL.txt', 'w')
for r in negSL:
    negFile.write(r + '\n')
negFile.close()
```

   c. Create a table for sample result. (100 for each in this case).

```
# pos into file
posFile = open('/Users/LXIN/Desktop/posSL.txt', 'w')
for r in posSL:
    posFile.write(r + '\n')
posFile.close()

# neg into file
negFile = open('/Users/LXIN/Desktop/negSL.txt', 'w')
for r in negSL:
    negFile.write(r + '\n')
negFile.close()
```

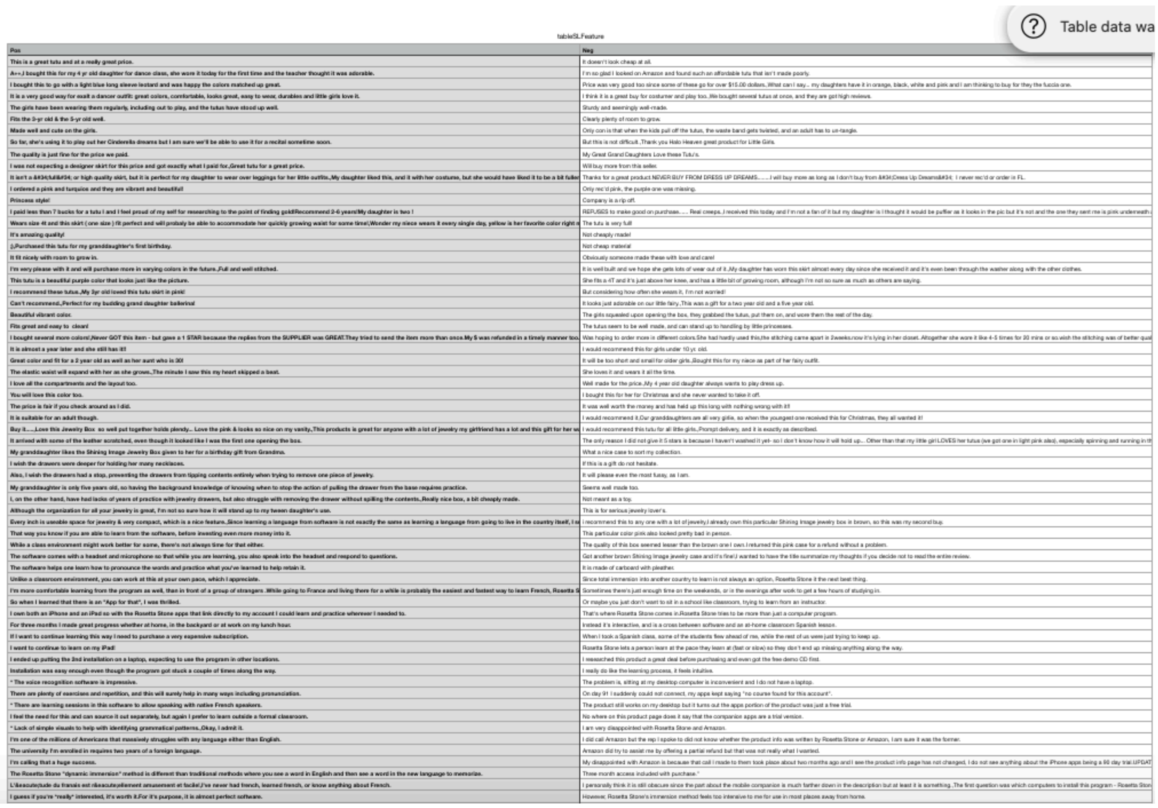

Table data wa

tableSL.Feature

Pos	Neg
This is a great tutu and at a really great price.	It doesn't look cheap at all.
A++,I bought this for my 4 yr old daughter for dance class, she wore it today for the first time and the teacher thought it was adorable.	I'm so glad I looked on Amazon and found such an affordable tutu that isn't made poorly.
I bought this to go with a light blue long sleeve leotard and was happy the colors matched up great.	Price was very good too since some of these go for over $15.00 dollars.,What can I say... my daughters have it in orange, black, white and pink and I am thinking to buy for they the fuccia one.
It is a very good way for exalt a dancer outfit: great colors, comfortable, looks great, easy to wear, durables and little girls love it.	I think it is a great buy for costumer and play too.,We bought several tutus at once, and they are got high reviews.
The girls have been wearing them regularly, including out to play, and the tutus have stood up well.	Sturdy and seemingly well-made.
Fits the 3-yr old & the 5-yr old well.	Clearly plenty of room to grow.
Made well and cute on the girls.	Only con is that when the kids pull off the tutus, the waste band gets twisted, and an adult has to un-tangle.
So far, she's using it to play out her Cinderella dreams but I am sure we'll be able to use it for a recital sometime soon.	But this is not difficult.,Thank you Halo Heaven great product for Little Girls.
The quality is just fine for the price we paid.	My Great Grand Daughters Love these Tutu's.
I was not expecting a designer skirt for this price and got exactly what I paid for.,Great tutu for a great price.	Will buy more from this seller.
It isn't a "full" or high quality skirt, but it is perfect for my daughter to wear over leggings for her little outfits.,My daughter liked this, and it with her costume, but she would have liked it to be a bit fuller	Thanks for a great product NEVER BUY FROM DRESS UP DREAMS.......I will buy more as long as I don't buy from "Dress Up Dreams" I never rec'd or order in FL.
I ordered a pink and turquioe and they are vibrant and beautiful!	Only rec'd pink, the purple one was missing.
Princess style!	Company is a rip off.
I paid less than 7 bucks for a tutu I and I feel proud of my self for researching to the point of finding gold!Recommend 2-6 years!My daughter is two !	REFUSES to make good on purchase..... Real creeps.,I received this today and I'm not a fan of it but my daughter is I thought it would be puffier as it looks in the pic but it's not and the one they sent me is pink underneath a
Wears size 4t and this skirt (one size) fit perfect and will probaly be able to accommodate her quickly growing waist for some time!,Wonder my niece wears it every single day, yellow is her favorite color right n	The tutu is very full!
It's amazing quality!	Not cheaply made!
:),Purchased this tutu for my granddaughter's first birthday.	Not cheap material
It fit nicely with room to grow in.	Obviously someone made these with love and care!
I'm very please with it and will purchase more in varying colors in the future.,Full and well stitched.	It is well built and we hope she gets lots of wear out of it.,My daughter has worn this skirt almost every day since she received it and it's even been through the washer along with the other clothes.
This tutu is a beautiful purple color that looks just like the picture.	She fits a 4T and it's just above her knee, and has a little bit of growing room, although I'm not so sure as much as others are saying.
I recommend these tutus.,My 3yr old loved this tutu skirt in pink!	But considering how often she wears it, I'm not worried!
Can't recommend.,Perfect for my budding grand daughter ballerina!	It looks just adorable on our little fairy.,This was a gift for a two year old and a five year old.
Beautiful vibrant color.	The girls squealed upon opening the box, they grabbed the tutus, put them on, and wore them the rest of the day.
Fits great and easy to clean!	The tutus seem to be well made, and can stand up to handling by little princesses.
I bought several more colors!,Never GOT this item - but gave a 1 STAR because the replies from the SUPPLIER was GREAT.They tried to send the item more than once.My $ was refunded in a timely manner too.	Was hoping to order more in different colors.She had hardly used this,the stitching came apart in 2weeks.now it's lying in her closet. Altogether she wore it like 4-5 times for 20 mins or so.wish the stitching was of better qua
It is almost a year later and she still has it!!	I would recommend this for girls under 10 yr. old.
Great color and fit for a 2 year old as well as her aunt who is 30!	It will be too short and small for older girls.,Bought this for my niece as part of her fairy outfit.
The elastic waist will expand with her as she grows.,The minute I saw this my heart skipped a beat.	She loves it and wears it all the time.
I love all the compartments and the layout too.	Well made for the price.,My 4 year old daughter always wants to play dress up.
You will love this color too.	I bought this for her for Christmas and she never wanted to take it off.
The price is fair if you check around as I did.	It was well worth the money and has held up this long with nothing wrong with it!!
It is suitable for an adult though.	I would recommend it,Our granddaughters are all very girlie, so when the youngest one received this for Christmas, they all wanted it!
Buy it.....,Love this Jewelry Box so well put together holds plendy... Love the pink & looks so nice on my vanity.,This products is great for anyone with a lot of jewelry my girlfriend has a lot and this gift for her w	I would recommend this tutu for all little girls.,Prompt delivery, and it is exactly as described.
It arrived with some of the leather scratched, even though it looked like I was the first one opening the box.	The only reason I did not give it 5 stars is because I haven't washed it yet- so I don't know how it will hold up... Other than that my little girl LOVES her tutus (we got one in light pink also), especially spinning and running in th
My granddaughter likes the Shining Image Jewelry Box given to her for a birthday gift from Grandma.	What a nice case to sort my collection.
I wish the drawers were deeper for holding her many necklaces.	If this is a gift do not hesitate.
Also, I wish the drawers had a stop, preventing the drawers from tipping contents entirely when trying to remove one piece of jewelry.	It will please even the most fussy, as I am.
My granddaughter is only five years old, so having the background knowledge of knowing when to stop the action of pulling the drawer from the base requires practice.	Seems well made too.
I, on the other hand, have had lacks of years of practice with jewelry drawers, but also struggle with removing the drawer without spilling the contents.,Really nice box, a bit cheaply made.	Not meant as a toy.
Although the organization for all your jewelry is great, I'm not so sure how it will stand up to my tween daughter's use.	This is for serious jewelry lover's.
Every inch is useable space for jewelry & very compact, which is a nice feature.,Since learning a language from software is not exactly the same as learning a language from going to live in the country itself, I sa	i recommend this to any one with a lot of jewelry,I already own this particular Shining Image jewelry box in brown, so this was my second buy.
That way you know if you are able to learn from the software, before investing even more money into it.	This particular color pink also looked pretty bad in person.
While a class environment might work better for some, there's not always time for that either.	The quality of this box seemed lesser than the brown one I own.I returned this pink case for a refund without a problem.
The software comes with a headset and microphone so that while you are learning, you also speak into the headset and respond to questions.	Got another brown Shining Image jewelry case and it's fine!,I wanted to have the title summarize my thoughts if you decide not to read the entire review.
The software helps one learn how to pronounce the words and practice what you've learned to help retain it.	It is made of carboard with pleather.
Unlike a classroom environment, you can work at this at your own pace, which I appreciate.	Since total immersion into another country to learn is not always an option, Rosetta Stone it the next best thing.
I'm more comfortable learning from the program as well, than in front of a group of strangers .While going to France and living there for a while is probably the easiest and fastest way to learn French, Rosetta S	Sometimes there's just enough time on the weekends, or in the evenings after work to get a few hours of studying in.
So when I learned that there is an "App for that", I was thrilled.	Or maybe you just don't want to sit in a school like classroom, trying to learn from an instructor.
I own both an iPhone and an iPad so with the Rosetta Stone apps that link directly to my account I could learn and practice wherever I needed to.	That's where Rosetta Stone comes in.Rosetta Stone tries to be more than just a computer program.
For three months I made great progress whether at home, in the backyard or at work on my lunch hour.	Instead it's interactive, and is a cross between software and an at-home classroom Spanish lesson.
If I want to continue learning this way I need to purchase a very expensive subscription.	When I took a Spanish class, some of the students flew ahead of me, while the rest of us were just trying to keep up.
I want to continue to learn on my iPad!	Rosetta Stone lets a person learn at the pace they learn at (fast or slow) so they don't end up missing anything along the way.
I ended up putting the 2nd installation on a laptop, expecting to use the program in other locations.	I researched this product a great deal before purchasing and even got the free demo CD first.
Installation was easy enough even though the program got stuck a couple of times along the way.	I really do like the learning process, it feels intuitive.
* The voice recognition software is impressive.	The problem is, sitting at my desktop computer is inconvenient and I do not have a laptop.
There are plenty of exercises and repetition, and this will surely help in many ways including pronunciation.	On day 91 I suddenly could not connect, my apps kept saying "no course found for this account".
* There are learning sessions in this software to allow speaking with native French speakers.	The product still works on my desktop but it turns out the apps portion of the product was just a free trial.
I feel the need for this and can source it out separately, but again I prefer to learn outside a formal classroom.	No where on this product page does it say that the companion apps are a trial version.
* Lack of simple visuals to help with identifying grammatical patterns.,Okay, I admit it.	I am very disappointed with Rosetta Stone and Amazon.
I'm one of the millions of Americans that massively struggles with any language either than English.	I did call Amazon but the rep I spoke to did not know whether the product info was written by Rosetta Stone or Amazon, I am sure it was the former.
The university I'm enrolled in requires two years of a foreign language.	Amazon did try to assist me by offering a partial refund but that was not really what I wanted.
I'm calling that a huge success.	My disappointed with Amazon is because that call I made to them took place about two months ago and I see the product info page has not changed, I do not see anything about the iPhone apps being a 90 day trial.UPDAT
The Rosetta Stone "dynamic immersion" method is different than traditional methods where you see a word in English and then see a word in the new language to memorize.	Three month access included with purchase."
L'étude du franais est réellement amusement et facile!,I've never had french, learned french, or know anything about French.	I personally think it is still obscure since the part about the mobile companion is much farther down in the description but at least it is something.,The first question was which computers to install this program - Rosetta Ston
I guess if you're *really* interested, it's worth it.For it's purpose, it is almost perfect software.	However, Rosetta Stone's immersion method feels too intensive to me for use in most places away from home.

2. Negation Features

a. Running result.

There are 724380 sentences tagged as negative and 415262 sentences tagged as positive.

```
# Reviews NOT
posNOT = []
negNOT = []

for s in tokensen:
    wordToken = nltk.word_tokenize(was)
    getFeature = NOT_features(wordToken, word_features, negationwords)
    if classifier3.classify(getFeature) == 'pos':
        posNOT.append(s)
    elif classifier3.classify(getFeature) == 'neg':
        negNOT.append(s)

print("neg -------------------------")
print(len(negNOT))
print(negNOT[:5])

print("\npos -------------------------")
print(len(posNOT))
print(posNOT[:5])
```

```
neg -------------------------
725380
["I'm so glad I looked on Amazon and found such an affordable tutu that isn't made poorly.", 'Price was very good too since some of these go for over $15.00 dollars.', 'What can I say... my daughters have it in orange, black, white and pink and I am thinking to buy for they the fuccia one.', 'I think it is a great buy for costumer and play too.', 'The girls have been wearing them regularly, including out to play, and the tutus have stood up well.']

pos -------------------------
415262
['This is a great tutu and at a really great price.', "It doesn't look cheap at all.", 'A++\nI bought this for my 4 y
r old daughter for dance class, she wore it today for the first time and the teacher thought it was adorable.', 'I bo
ught this to go with a light blue long sleeve leotard and was happy the colors matched up great.', 'It is a very good
way for exalt a dancer outfit: great colors, comfortable, looks great, easy to wear, durables and little girls love i
t.']
```

b. Store the two sets of sentences in two files.

```
# pos into file
posFile = open('/Users/LXIN/Desktop/posNOT.txt', 'w')
for r in posNOT:
    posFile.write(r + '\n')
posFile.close()

# neg into file
negFile = open('/Users/LXIN/Desktop/negNOT.txt', 'w')
for r in negNOT:
    negFile.write(r + '\n')
negFile.close()
```

c. Create a table for sample result. (100 for each in this case).

```
import pandas as pd
import os
col = ['Positive', 'Negative']
define = pd.DataFrame(columns = col)

define['Positive'] = posNOT[:100]
define['Negative'] = negNOT[:100]

file = "/Users/LXIN/Desktop/table.csv"

if not os.path.isfile(file):
    define.to_csv(file, header = True, index = False, encoding = 'utf-8')
print(define)
```

```
                                              Positive  \
0    This is a great tutu and at a really great price.
1                          It doesn't look cheap at all.
2    A++\nI bought this for my 4 yr old daughter fo...
3    I bought this to go with a light blue long sle...
4    It is a very good way for exalt a dancer outfi...
..                                                 ...
95   This is a better way to learn to make the asso...
96   However, it is still nice to have some feedbac...
97   However, how much would it cost to take a fore...
98   Rosetta Stone Italiano Level 1Rosetta Stone It...
99   The included instructions are simple and strai...

                                              Negative
0    I'm so glad I looked on Amazon and found such ...
1    Price was very good too since some of these go...
2    What can I say... my daughters have it in oran...
3    I think it is a great buy for costumer and pla...
4    The girls have been wearing them regularly, in...
..                                                 ...
95   This wasn't a problem for me yet it's somethin...
96   * Lack of matching supplemental physical mater...
97                                    Okay, I admit it.
98   I have recently returned to college to finish ...
99                                                 Ugh.

[100 rows x 2 columns]
```

table

Positive	Negative
This is a great tutu and at a really great price.	I'm so glad I looked on Amazon and found such an affordable tutu that isn't made poorly.
It doesn't look cheap at all.	Price was very good too since some of these go for over $15.00 dollars.
A++ I bought this for my 4 yr old daughter for dance class, she wore it today for the first time and the teacher thought it was adorable.	What can I say... my daughters have it in orange, black, white and pink and I am thinking to buy for they the fuccia one.
I bought this to go with a light blue long sleeve leotard and was happy the colors matched up great.	I think it is a great buy for costumer and play too.
It is a very good way for exalt a dancer outfit: great colors, comfortable, looks great, easy to wear, durables and little girls love it.	The girls have been wearing them regularly, including out to play, and the tutus have stood up well.
We bought several tutus at once, and they are got high reviews.	Clearly plenty of room to grow.
Sturdy and seemingly well-made.	But this is not difficult.
Fits the 3-yr old & the 5-yr old well.	Thank you Halo Heaven great product for Little Girls.
Only con is that when the kids pull off the tutus, the waste band gets twisted, and an adult has to un-tangle.	My Great Grand Daughters Love these Tutu's.
Made well and cute on the girls.	Will buy more from this seller.
Thanks for a great product.NEVER BUY FROM DRESS UP DREAMS.......I will buy more as long as I don't buy from "Dress Up Dreams" I never rec'd or order in FL.	Only rec'd pink, the purple one was missing.
REFUSES to make good on purchase...... Real creeps.	Company is a rip off.
So far, she's using it to play out her Cinderella dreams but I am sure we'll be able to use it for a recital sometime soon.	I received this today and I'm not a fan of it but my daughter is I thought it would be puffer as it looks in the pic but it's not and the one they sent me is pink underneath and the waist band is pink which is not what I wanted d
Great tutu for a great price.	Bought this as a backup to the regular ballet outfit my daughter has to wear.
It isn't a "full" or high quality skirt, but it is perfect for my daughter to wear over leggings for her little outfits.	The quality is just fine for the price we paid.
My daughter liked this, and it with her costume, but she would have liked it to be a bit fuller.	I was not expecting a designer skirt for this price and got exactly what I paid for.
I ordered a pink and turquios and they are vibrant and beautiful!	For what I paid for two tutus is unbeatable anywhere!
Princess style!	The tutu is very full!
Wears size 4t and this skirt (one size) fit perfect and will probaly be able to accommodate her quickly growing waist for some time!	Not cheaply made!
It's amazing quality!	Not cheap material
It fit nicely with room to grow in.	Obviously someone made these with love and care!
I'm very please with it and will purchase more in varying colors in the future.	I paid less than 7 bucks for a tutu I and I feel proud of my self for researching to the point of finding gold!Recommend 2-6 years!My daughter is two !
Full and well stitched.	Wonder my niece wears it every single day, yellow is her favorite color right now an this cute little tutu made he da.
This tutu is a beautiful purple color that looks just like the picture.	It is well built and we hope she gets lots of wear out of it.
It looks just adorable on our little fairy.	My daughter has worn this skirt almost every day since she received it and it's even been through the washer along with the other clothes.
This was a gift for a two year old and a five year old.	She fits a 4T and it's just above her knee, and has a little bit of growing room, although I'm not so sure as much as others are saying.
The tutus seem to be well made, and can stand up to handling by little princesses.	But considering how often she wears it, I'm not worried!
My 3yr old loved this tutu skirt in pink!	:) Purchased this tutu for my granddaughter's first birthday.
Beautiful vibrant color.	The girls squealed upon opening the box, they grabbed the tutus, put them on, and wore them the rest of the day.
Fits great and easy to clean!	I recommend these tutus.
I bought several more colors!	Was hoping to order more in different colors.She had hardly used this,the stitching came apart in 2weeks.now it's lying in her closet. Altogether she wore it like 4-5 times for 20 mins or so.wish the stitching was of better qual
Nice and puffy tutu skirt.	Can't recommend.
Bought this for my niece as part of her fairy outfit.	Perfect for my budding grand daughter ballerina!
I bought this for her for Christmas and she never wanted to take it off.	Never GOT this item - but gave a 1 STAR because the replies from the SUPPLIER was GREAT.They tried to send the item more than once.My $ was refunded in a timely manner too.It was a shame I never got it for my daugh
It is almost a year later and she still has it!	I would recommend this for girls under 10 yr. old.
Great color and fit for a 2 year old as well as her aunt who is 30!	It will be too short and small for older girls.
The elastic waist will expand with her as she grows.	She loves it and wears it all the time.
You will love this color too.	Well made for the price.
If this is a gift do not hesitate.	My 4 year old daughter always wants to play dress up.
It is suitable for an adult though.	It was well worth the money and has held up this long with nothing wrong with it!
Buy it..... Love this Jewelry Box so well put together holds plendy... Love the pink & looks so nice on my vanity.	I would recommend it Our granddaughters are all very girlie, so when the youngest one received this for Christmas, they all wanted it!
My granddaughter likes the Shining Image Jewelry Box given to her for a birthday gift from Grandma.	I would recommend this tutu for all little girls.
Really nice box, a bit cheaply made.	Prompt delivery, and it is exactly as described.
Every inch is useable space for jewelry & very compact, which is a nice feature.	The only reason I did not give it 5 stars is because I haven't washed it yet- so I don't know how it will hold up... Other than that my little girl LOVES her tutus (we got one in light pink also), especially spinning and running in th
While a class environment might work better for some, there's not always time for that either.	The minute I saw this my heart skipped a beat.
The software comes with a headset and microphone so that while you are learning, you also speak into the headset and respond to questions.	What a nice case to sort my collection.
Unlike a classroom environment, you can work at this at your own pace, which I appreciate.	I love all the compartments and the layout too.
I'm more comfortable learning from the program as well, than in front of a group of strangers .While going to France and living there for a while is probably the easiest and fastest way to learn French, Rosetta S	It will please even the most fussy, as I am.
I own both an iPhone and an iPad so with the Rosetta Stone apps that link directly to my account I could learn and practice wherever I needed to.	The price is fair if you check around as I did.
For three months I made great progress whether at home, in the backyard or at work on my lunch hour.	Seems well made too.
If I want to continue learning this way I need to purchase a very expensive subscription.	Not meant as a toy.
I want to continue to learn on my iPad!	This is for serious jewelry lover's.
Three month access included with purchase.*	This products is great for anyone with a lot of jewelry my girlfriend has a lot and this gift for her was one of my best ideas!
The first question was which computers to install this program - Rosetta Stone has a two seat minimum.	i recommend this to any one with a lot of jewelry I already own this particular Shining Image jewelry box in brown, so this was my second buy.
* There are learning sessions in this software to allow speaking with native French speakers.	It arrived with some of the leather scratched, even though it looked like I was the first one opening the box.
For me as an individual, this is a better way to learn than a formal classroom.Cons: The limitation in the number of seats (2).	This particular color pink also looked pretty bad in person.
I feel the need for this and can source it out separately, but again I prefer to learn outside a formal classroom.	The quality of this box seemed lesser than the brown one I own.I returned this pink case for a refund without a problem.
* Lack of simple visuals to help with identifying grammatical patterns.	Got another brown Shining Image jewelry case and it's fine!
I'm one of the millions of Americans that massively struggles with any language either than English.	I wanted to have the title summarize my thoughts if you decide not to read the entire review.
The university I'm enrolled in requires two years of a foreign language.	I wish the drawers were deeper for holding her many necklaces.
I'm calling that a huge success.	Also, I wish the drawers had a stop, preventing the drawers from tipping contents entirely when trying to remove one piece of jewelry.
They had their lawyers contact the sites and have them remove the listing.Apparently in the tiny 4-point font disclaimers on the package it says that the license is not transferable, so you can't sell it as a used it	My granddaughter is only five years old, so having the background knowledge of knowing when to stop the action of pulling the drawer from the base requires practice.
The Rosetta Stone "dynamic immersion" method is different than traditional methods where you see a word in English and then see a word in the new language to memorize.	I, on the other hand, have had lacks of years of practice with jewelry drawers, but also struggle with removing the drawer without spilling the contents.
You will feel confident with the basics after this.	It is made of carboard with pleather.
I think that means the girl drinks.	Although the organization for all your jewelry is great, I'm not so sure how it will stand up to my tween daughter's use.
They really did one heck of a job making learning fun and easy.	Since learning a language from software is not exactly the same as learning a language from going to live in the country itself, I suggest picking up the Rosetta Stone software 1 level at a time, like this French level 1.

Bonus (Different Word Set & Additional Feature)

A brief description: A different dataset is used in this part. It is a dataset of sample tweets from NLTK package. I'll download it and apply some data cleaning methods on it (stopwords, etc..) I'll train a model on pre-classified tweets, and use this model to classify the Amazon reviews into positive and negative sentiments.
Reference: < https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk>

1. Download Data
   a. Download the sample tweets from NLTK package

```
import nltk

# Download the sample tweets from the NLTK package
nltk.download('twitter_samples')
```

```
[nltk_data] Downloading package twitter_samples to
[nltk_data]     /Users/LXIN/nltk_data...
[nltk_data]   Unzipping corpora/twitter_samples.zip.
```

```
True
```

   b. Download

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

```
[nltk_data] Downloading package punkt to /Users/LXIN/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
True
```

2. Pre-processing Data
   a. There are 3 datasets from NLTK which contain tweets to train and test the model:
   Negative_tweets.json: 5000 tweets with negative sentiments
   Positive_tweets.json: 5000 tweets with positive sentiments
   Tweets.20150430-223406: 20000 tweets with no sentiments

```
from nltk.corpus import twitter_samples

# negative_tweets.json: 5000 tweets with negative sentiments
# positive_tweets.json: 5000 tweets with positive sentiments
# tweets.20150430-223406.json: 20000 tweets with no sentiments

positive_tweets = twitter_samples.strings('positive_tweets.json')
negative_tweets = twitter_samples.strings('negative_tweets.json')
text = twitter_samples.strings('tweets.20150430-223406.json')
tweet_tokens = twitter_samples.tokenized('positive_tweets.json')
```

```
print(tweet_tokens[0])
```

```
['#FollowFriday', '@France_Inte', '@PKuchly57', '@Milipol_Paris', 'for', 'being', 'top', 'engaged', 'members', 'in',
'my', 'community', 'this', 'week', ':)']
```

   b. Tokenization

```
# Normalizing the Data
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')

from nltk.tag import pos_tag
from nltk.corpus import twitter_samples
tweet_tokens = twitter_samples.tokenized('positive_tweets.json')
print(pos_tag(tweet_tokens[0]))
```

```
[nltk_data] Downloading package wordnet to /Users/LXIN/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /Users/LXIN/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

```
[('#FollowFriday', 'JJ'), ('@France_Inte', 'NNP'), ('@PKuchly57', 'NNP'), ('@Milipol_Paris', 'NNP'), ('for', 'IN'),
('being', 'VBG'), ('top', 'JJ'), ('engaged', 'VBN'), ('members', 'NNS'), ('in', 'IN'), ('my', 'PRP$'), ('community',
'NN'), ('this', 'DT'), ('week', 'NN'), (':)', 'NN')]
```

   c. Normalizing the data

```python
from nltk.tag import pos_tag
from nltk.stem.wordnet import WordNetLemmatizer

# This code imports the WordNetLemmatizer class and initializes it to a variable, lemmatizer
def lemmatize_sentence(tokens):
    lemmatizer = WordNetLemmatizer()
    lemmatized_sentence = []
    for word, tag in pos_tag(tokens):
        if tag.startswith('NN'):
            pos = 'n'
        elif tag.startswith('VB'):
            pos = 'v'
        else:
            pos = 'a'
        lemmatized_sentence.append(lemmatizer.lemmatize(word, pos))
    return lemmatized_sentence

print(lemmatize_sentence(tweet_tokens[0]))
```

```
['#FollowFriday', '@France_Inte', '@PKuchly57', '@Milipol_Paris', 'for', 'be', 'top', 'engage', 'member', 'in', 'my', 'community', 'this', 'week', ':)']
```

d. Remove noise

Use regular expressions in Python to search for and remove these items:

Hyperlinks - All hyperlinks in Twitter are converted to the URL shortener t.co.

Twitter handles in replies

Punctuation and special characters

```python
# Removing Noise from the Data

# Use regular expressions in Python to search for and remove these items:
# Hyperlinks - All hyperlinks in Twitter are converted to the URL shortener t.co.
# Twitter handles in replies
# Punctuation and special characters

import re, string

def remove_noise(tweet_tokens, stop_words = ()):

    cleaned_tokens = []

    for token, tag in pos_tag(tweet_tokens):
        token = re.sub('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*\(\),]|'\
                       '(?:%[0-9a-fA-F][0-9a-fA-F]))+','', token)
        token = re.sub("(@[A-Za-z0-9_]+)","", token)

        if tag.startswith("NN"):
            pos = 'n'
        elif tag.startswith('VB'):
            pos = 'v'
        else:
            pos = 'a'

        lemmatizer = WordNetLemmatizer()
        token = lemmatizer.lemmatize(token, pos)

        if len(token) > 0 and token not in string.punctuation and token.lower() not in stop_words:
            cleaned_tokens.append(token.lower())
    return cleaned_tokens
```

e. Remove stopwords

```python
# Remove stopwords
from nltk.corpus import stopwords
stop_words = stopwords.words('english')

#print(remove_noise(tweet_tokens[0], stop_words))

positive_tweet_tokens = twitter_samples.tokenized('positive_tweets.json')
negative_tweet_tokens = twitter_samples.tokenized('negative_tweets.json')

positive_cleaned_tokens_list = []
negative_cleaned_tokens_list = []

for tokens in positive_tweet_tokens:
    positive_cleaned_tokens_list.append(remove_noise(tokens, stop_words))

for tokens in negative_tweet_tokens:
    negative_cleaned_tokens_list.append(remove_noise(tokens, stop_words))
```

3. Word set

a. Take a list of tweets as an argument to provide a list of words in all of the tweet tokens joined

```python
# Determining Word Density
# Take a list of tweets as an argument to provide a list of words in all of the tweet tokens joined
def get_all_words(cleaned_tokens_list):
    for tokens in cleaned_tokens_list:
        for token in tokens:
            yield token

all_pos_words = get_all_words(positive_cleaned_tokens_list)
```

b. Get the most common words

```python
from nltk import FreqDist

freq_dist_pos = FreqDist(all_pos_words)
print(freq_dist_pos.most_common(10))
```

c. Creating training and test set for model
A label ("positive" or "negative") is labeled to each tweet. A dataset is created then by joining the positive and negative tweets

```python
# Preparing Data for the Model

# Converting Tokens to a Dictionary
def get_tweets_for_model(cleaned_tokens_list):
    for tweet_tokens in cleaned_tokens_list:
        yield dict([token, True] for token in tweet_tokens)

positive_tokens_for_model = get_tweets_for_model(positive_cleaned_tokens_list)
negative_tokens_for_model = get_tweets_for_model(negative_cleaned_tokens_list)


# Splitting the Dataset for Training and Testing the Model
import random

positive_dataset = [(tweet_dict, "Positive")
                     for tweet_dict in positive_tokens_for_model]

negative_dataset = [(tweet_dict, "Negative")
                     for tweet_dict in negative_tokens_for_model]

dataset = positive_dataset + negative_dataset

random.shuffle(dataset)

train_data = dataset[:7000]
test_data = dataset[7000:]
```

4. Building and training the model
   a. Build and train the model

```python
# Building and Testing the Model
from nltk import classify
from nltk import NaiveBayesClassifier

classifier = NaiveBayesClassifier.train(train_data)
```

b. Get the accuracy

```python
print("Accuracy is:", classify.accuracy(classifier, test_data))

print(classifier.show_most_informative_features(10))
```

```
Accuracy is: 0.9976666666666667
Most Informative Features
                      :( = True           Negati : Positi =   2085.4 : 1.0
                      :) = True           Positi : Negati =   1650.9 : 1.0
                     sad = True           Negati : Positi =     24.8 : 1.0
                    glad = True           Positi : Negati =     22.7 : 1.0
                     bam = True           Positi : Negati =     22.0 : 1.0
                follower = True           Positi : Negati =     21.9 : 1.0
                 welcome = True           Positi : Negati =     20.3 : 1.0
                     x15 = True           Negati : Positi =     17.2 : 1.0
                followed = True           Negati : Positi =     14.8 : 1.0
              appreciate = True           Positi : Negati =     14.8 : 1.0
None
```

5. Apply on the Amazon Review

a. Retrieve data and tokenize it

```
# Retrieve text file
textfile = open("/Users/LXIN/Desktop/reviews.txt")
reviewText = textfile.read()
print(reviewText[:20])

#Tokenize
from nltk import tokenize
tokensen = tokenize.sent_tokenize(reviewText)
print(len(tokensen))
print(tokensen[:20])
```

```
This is a great tutu
1140642
['This is a great tutu and at a really great price.', "It doesn't look cheap at all.", "I'm so glad I looked on Amazo
n and found such an affordable tutu that isn't made poorly.", 'A++\nI bought this for my 4 yr old daughter for dance
class, she wore it today for the first time and the teacher thought it was adorable.', 'I bought this to go with a li
ght blue long sleeve leotard and was happy the colors matched up great.', 'Price was very good too since some of thes
e go for over $15.00 dollars.', 'What can I say... my daughters have it in orange, black, white and pink and I am thi
nking to buy for they the fuccia one.', 'It is a very good way for exalt a dancer outfit: great colors, comfortable,
looks great, easy to wear, durables and little girls love it.', 'I think it is a great buy for costumer and play to
o.', 'We bought several tutus at once, and they are got high reviews.', 'Sturdy and seemingly well-made.', 'The girls
have been wearing them regularly, including out to play, and the tutus have stood up well.', 'Fits the 3-yr old & the
5-yr old well.', 'Clearly plenty of room to grow.', 'Only con is that when the kids pull off the tutus, the waste ban
d gets twisted, and an adult has to un-tangle.', 'But this is not difficult.', 'Thank you Halo Heaven great product f
or Little Girls.', "My Great Grand Daughters Love these Tutu's.", 'Will buy more from this seller.', 'Made well and c
ute on the girls.']
```

b. Perform the model on all comments

```
# Reviews
posBouns = []
negBouns = []

from nltk.tokenize import word_tokenize

for s in tokensen:
    wordToken = remove_noise(word_tokenize(s))

    if classifier.classify(dict([token, True] for token in wordToken)) == 'Positive':
        posBouns.append(s)
    elif classifier.classify(dict([token, True] for token in wordToken)) == 'Negative':
        negBouns.append(s)

print("neg -------------------------")
print(len(negBouns))
print(negBouns[:5])

print("\npos -------------------------")
print(len(posBouns))
print(posBouns[:5])
```

```
neg -------------------------
495485
['A++\nI bought this for my 4 yr old daughter for dance class, she wore it today for the first time and the teacher t
hought it was adorable.', 'Price was very good too since some of these go for over $15.00 dollars.', 'We bought sever
al tutus at once, and they are got high reviews.', 'Fits the 3-yr old & the 5-yr old well.', 'Clearly plenty of room
to grow.']

pos -------------------------
645157
['This is a great tutu and at a really great price.', "It doesn't look cheap at all.", "I'm so glad I looked on Amazo
n and found such an affordable tutu that isn't made poorly.", 'I bought this to go with a light blue long sleeve leot
ard and was happy the colors matched up great.', 'What can I say... my daughters have it in orange, black, white and
pink and I am thinking to buy for they the fuccia one.']
```

c. Save the two sets into files

```
# pos into file
posFile = open('/Users/LXIN/Desktop/posBonus.txt', 'w')
for r in posBouns:
    posFile.write(r + '\n')
posFile.close()

# neg into file
negFile = open('/Users/LXIN/Desktop/negBonus.txt', 'w')
for r in negBouns:
    negFile.write(r + '\n')
negFile.close()
```

d. Save the sample result in form of table to anther file (100 sentences from each set)

```python
import pandas as pd
import os
col = ['Positive', 'Negative']
define = pd.DataFrame(columns = col)

define['Positive'] = posBouns[:100]
define['Negative'] = negBouns[:100]

file = "/Users/LXIN/Desktop/tableBonus.csv"

if not os.path.isfile(file):
    define.to_csv(file, header = True, index = False, encoding = 'utf-8')
print(define)
```

```
                                              Positive  \
0   This is a great tutu and at a really great price.
1                          It doesn't look cheap at all.
2   I'm so glad I looked on Amazon and found such ...
3   I bought this to go with a light blue long sle...
4   What can I say... my daughters have it in oran...
..                                                 ...
95                               Okay, I admit it.
96  The university I'm enrolled in requires two ye...
97            I never thought that was a possibility.
98  I'm not ready to jet off to France or anything...
99                       I'm calling that a huge success.

                                              Negative
0   A++\nI bought this for my 4 yr old daughter fo...
1   Price was very good too since some of these go...
2   We bought several tutus at once, and they are ...
3             Fits the 3-yr old & the 5-yr old well.
4                    Clearly plenty of room to grow.
..                                                 ...
95                                      Neither do I.
96  I've taken college/high school and even junior...
97  But by far, the fastest way to a new language ...
98        Plus iphone apps and internet based learning.
99   I understand the need to prevent software theft.

[100 rows x 2 columns]
```

tableBonus

Positive	Negative
This is a great tutu and at a really great price.	A++ I bought this for my 4 yr old daughter for dance class, she wore it today for the first time and the teacher thought it was adorable.
It doesn't look cheap at all.	Price was very good too since some of these go for over $15.00 dollars.
I'm so glad I looked on Amazon and found such an affordable tutu that isn't made poorly.	We bought several tutus at once, and they are got high reviews.
I bought this to go with a light blue long sleeve leotard and was happy the colors matched up great.	Fits the 3-yr old & the 5-yr old well.
What can I say... my daughters have it in orange, black, white and pink and I am thinking to buy for they the fuccia one.	Clearly plenty of room to grow.
It is a very good way for exalt a dancer outfit: great colors, comfortable, looks great, easy to wear, durables and little girls love it.	Only con is that when the kids pull off the tutus, the waste band gets twisted, and an adult has to un-tangle.
I think it is a great buy for costumer and play too.	Will buy more from this seller.
Sturdy and seemingly well-made.	Only rec'd pink, the purple one was missing.
The girls have been wearing them regularly, including out to play, and the tutus have stood up well.	Company is a rip off.
But this is not difficult.	Bought this as a backup to the regular ballet outfit my daughter has to wear.
Thank you Halo Heaven great product for Little Girls.	So far, she's using it to play out her Cinderella dreams but I am sure we'll be able to use it for a recital sometime soon.
My Great Grand Daughters Love these Tutu's.	I was not expecting a designer skirt for this price and got exactly what I paid for.
Made well and cute on the girls.	For what I paid for two tutus is unbeatable anywhere!
Thanks for a great product.NEVER BUY FROM DRESS UP DREAMS.......I will buy more as long as I don't buy from "Dress Up Dreams" I never rec'd or order in FL.	Obviously someone made these with love and care!
REFUSES to make good on purchase...... Real creeps.	I paid less than 7 bucks for a tutu I and I feel proud of my self for researching to the point of finding gold!Recommend 2-6 years!My daughter is two !
I received this today and I'm not a fan of it but my daughter is I thought it would be puffier as it looks in the pic but it's not and the one they sent me is pink underneath and the waist band is pink which is not wh	Wears size 4t and this skirt (one size) fit perfect and will probaly be able to accommodate her quickly growing waist for some time!
The quality is just fine for the price we paid.	My daughter has worn this skirt almost every day since she received it and it's even been through the washer along with the other clothes.
Great tutu for a great price.	She fits a 4T and it's just above her knee, and has a little bit of growing room, although I'm not so sure as much as others are saying.
It isn't a "full" or high quality skirt, but it is perfect for my daughter to wear over leggings for her little outfits.	But considering how often she wears it, I'm not worried!
My daughter liked this, and it with her costume, but she would have liked it to be a bit fuller.	It fit nicely with room to grow in.
I ordered a pink and turquice and they are vibrant and beautiful!	This was a gift for a two year old and a five year old.
The tutu is very full!	The girls squealed upon opening the box, they grabbed the tutus, put them on, and wore them the rest of the day.
Princess style!	Was hoping to order more in different colors.She had hardly used this,the stitching came apart in 2weeks.now it's lying in her closet. Altogether she wore it like 4-5 times for 20 mins or so.wish the stitching was of better qual
Not cheaply made!	It will be too short and small for older girls.
Not cheap material!	Bought this for my niece as part of her fairy outfit.
Wonder my niece wears it every single day, yellow is her favorite color right now an this cute little tutu made he da.	I bought this for her for Christmas and she never wanted to take it off.
It is well built and we hope she gets lots of wear out of it.	It is almost a year later and she still has it!
It's amazing quality!	It was well worth the money and has held up this long with nothing wrong with it!
:) Purchased this tutu for my granddaughter's first birthday.	I would recommend it Our granddaughters are all very girlie, so when the youngest one received this for Christmas, they all wanted it!
I'm very please with it and will purchase more in varying colors in the future.	Prompt delivery, and it is exactly as described.
Full and well stitched.	The minute I saw this my heart skipped a beat.
This tutu is a beautiful purple color that looks just like the picture.	If this is a gift do not hesitate.
It looks just adorable on our little fairy.	It will please even the most fussy, as I am.
The tutus seem to be well made, and can stand up to handling by little princesses.	The price is fair if you check around as I did.
I recommend these tutus.	Not meant as a toy.
My 3yr old loved this tutu skirt in pink!	i recommend this to any one with a lot of jewelry I already own this particular Shining Image jewelry box in brown, so this was my second buy.
Can't recommend.	It arrived with some of the leather scratched, even though it looked like I was the first one opening the box.
Perfect for my budding grand daughter ballerina!	The quality of this box seemed lesser than the brown one I own.I returned this pink case for a refund without a problem.
Beautiful vibrant color.	Got another brown Shining Image jewelry case and it's fine!
Fits great and easy to clean!	My granddaughter likes the Shining Image Jewelry Box given to her for a birthday gift from Grandma.
I bought several more colors!	I wish the drawers were deeper for holding her many necklaces.
Never GOT this item - but gave a 1 STAR because the replies from the SUPPLIER was GREAT.They tried to send the item more than once.My $ was refunded in a timely manner too.It was a shame I never got it t	Also, I wish the drawers had a stop, preventing the drawers from tipping contents entirely when trying to remove one piece of jewelry.
Nice and puffy tutu skirt.	My granddaughter is only five years old, so having the background knowledge of knowing when to stop the action of pulling the drawer from the base requires practice.
I would recommend this for girls under 10 yr. old.	That way you know if you are able to learn from the software, before investing even more money into it.
She loves it and wears it all the time.	Or maybe you just don't want to sit in a school like classroom, trying to learn from an instructor.
Well made for the price.	Instead it's interactive, and is a cross between software and an at-home classroom Spanish lesson.
My 4 year old daughter always wants to play dress up.	When I took a Spanish class, some of the students flew ahead of me, while the rest of us were just trying to keep up.
Great color and fit for a 2 year old as well as her aunt who is 30!	Rosetta Stone lets a person learn at the pace they learn at (fast or slow) so they don't end up missing anything along the way.
I would recommend this tutu for all little girls.	For three months I made great progress whether at home, in the backyard or at work on my lunch hour.
The only reason I did not give it 5 stars is because I haven't washed it yet- so I don't know how it will hold up... Other than that my little girl LOVES her tutus (we got one in light pink also), especially spinning an	If I want to continue learning this way I need to purchase a very expensive subscription.
The elastic waist will expand with her as she grows.	No where on this product page does it say that the companion apps are a trial version.
What a nice case to sort my collection.	I am very disappointed with Rosetta Stone and Amazon.
I love all the compartments and the layout too.	Amazon did try to assist me by offering a partial refund but that was not really what I wanted.
You will love this color too.	I want to continue to learn on my iPad!
Seems well made too.	My disappointed with Amazon is because that call I made to them took place about two months ago and I see the product info page has not changed, I do not see anything about the iPhone apps being a 90 day trial.UPDAT
It is suitable for an adult though.	I personally think it is still obscure since the part about the mobile companion is much farther down in the description but at least it is something.
This is for serious jewelry lover's.	I ended up putting the 2nd installation on a laptop, expecting to use the program in other locations.
Buy it...... Love this Jewelry Box so well put together holds plendy... Love the pink & looks so nice on my vanity.	The gear, headset with microphone, is sufficient.Pros:" For total immersion-type instruction without the benefit of communicating with other people, this seemed very good.
This products is great for anyone with a lot of jewelry my girlfriend has a lot and this gift for her was one of my best ideas!	This wasn't a problem for me yet it's something to know about.
This particular color pink also looked pretty bad in person.	I feel the need for this and can source it out separately, but again I prefer to learn outside a formal classroom.
I wanted to have the title summarize my thoughts if you decide not to read the entire review.	I'm one of the millions of Americans that massively struggles with any language either than English.
I, on the other hand, have had lacks of years of practice with jewelry drawers, but also struggle with removing the drawer without spilling the contents.	I have recently returned to college to finish out my Bachelors.
Really nice box, a bit cheaply made.	Ugh.
It is made of carboard with pleather.	Kill me.
[illegible]	[illegible]

Conclusion

1. **A brief description of how I conducted this assignment:** I used tokenization (sentence-level) to split the reviews into single sentence first. I believe no more cleaning or pre-processing is needed for the data in this assignment. I downloaded the sentence_poliarity to create a word_list for training and test then, which needs to eliminate non-alphabet characters and stopwords. Note that some stopwords need to be kept for the Not_feature.
   An additional step for Not_feature is a list of negation words which is defined by ourselves.
   These two features are implemented to train the unigram word_features.
   A comparison of accuracy between these two features was conducted then, and both were used to analyze the review contents from Amazon Product Data. In other word, exploring the sentiment of the comments at sentence level.
   The results of the sentiment analysis are very different, which will be discussed later.
   I also downloaded another dataset from NLTK and applied different features on it, in order to build a different model (Bonus part).
2. **Result**
   a. **Not_feature and SL_feature.** Both Not_feature and SL_feature have an accuracy lower than 0.8, an improvement should be considered therefore. For Not_feature, there are 415262 positive comments and 725380 negative comments, the ratio of positive comments is only 36%. This indicates that many customers who left comments on Amazon have negative attitude towards what they bought, in other word, they are not that satisfied. The conclusion implies that Amazon shall try to find out why most customers were not happy with the products bought and improve more on the quality.
   b. Another feature I applied on the different dataset has an accuracy of 0.997, which is much more higher than the Not_feature and SL_feature. Using the classifier trained by this feature, the result shows that there are 495485 negative comments and 645157 positive sentences, the ratio of positive comments is 56%. We could conclude that the comments are half positive and half negative, but more customers are satisfied with the products overall.
   c. The accuracy of Not_feature and SL_feature is less than 80%, while the other is 99.7%.
3. **Improvement**
   Though the feature I chose is the higher one, the accuracy is still lower than 0.8, and it's much lower than the feature I chose in the Bonus part.
   Possible measures to improve the accuracy include improving the accuracy of Not_feature by adding more (possible) negation words into the list and trying other advanced features.