[7] Amazon Fine Food Reviews Analysis

Data Source: https://www.kaggle.com/snap/amazon-fine-food-reviews (https://www.kaggle.com/snap/amazon-fine-food-reviews)

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.

Number of reviews: 568,454
Number of users: 256,059
Number of products: 74,258
Timespan: Oct 1999 - Oct 2012
Number of Attributes/Columns in data: 10

Attribute Information:

1. Id
2. ProductId - unique identifier for the product
3. UserId - unqiue identifier for the user
4. ProfileName
5. HelpfulnessNumerator - number of users who found the review helpful
6. HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

Objective:

Given a review, determine whether the review is positive (Rating of 4 or 5) or negative (rating of 1 or 2).

[Q] How to determine if a review is positive or negative?

[Ans] We could use the Score/Rating. A rating of 4 or 5 could be cosnidered a positive review. A review of 1 or 2 could be considered negative. A review of 3 is nuetral and ignored. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

[7.1] Loading the data

The dataset is available in two forms

1. .csv file
2. SQLite Database

In order to load the data, We have used the SQLITE dataset as it easier to query the data and visualise the data efficiently.

Here as we only want to get the global sentiment of the recommendations (positive or negative), we will purposefully ignore all Scores equal to 3. If the score id above 3, then the recommendation wil be set to "positive". Otherwise, it will be set to "negative".

In [2]:

```
%matplotlib inline

import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer



# using the SQLite Table to read data.
con = sqlite3.connect('./amazon-fine-food-reviews/database.sqlite')



#filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
filtered_data = pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3
""", con)




# Give reviews with Score>3 a positive rating, and reviews with a score<3 a negative rating
def partition(x):
    if x < 3:
        return 'negative'
    return 'positive'

#changing reviews with score less than 3 to be positive and vice-versa
actualScore = filtered_data['Score']
positiveNegative = actualScore.map(partition)
filtered_data['Score'] = positiveNegative
```

In [2]:

```
filtered_data.shape #looking at the number of attributes and size of the data
filtered_data.head()
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenom
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	

Exploratory Data Analysis

[7.1.2] Data Cleaning: Deduplication

It is observed (as shown in the table below) that the reviews data had many duplicate entries. Hence it was necessary to remove duplicates in order to get unbiased results for the analysis of the data. Following is an example:

In [3]:

```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND UserId="AR5J8UI46CURR"
ORDER BY ProductID
""", con)
display
```

1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2	
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2	
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2	
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha	2	2	

As can be seen above the same user has multiple reviews of the with the same values for HelpfulnessNumerator, HelpfulnessDenominator, Score, Time, Summary and Text and on doing analysis it was found that

ProductId=B000HDOPZG was Loacker Quadratini Vanilla Wafer Cookies, 8.82-Ounce Packages (Pack of 8)

ProductId=B000HDL1RQ was Loacker Quadratini Lemon Wafer Cookies, 8.82-Ounce Packages (Pack of 8) and so on

It was inferred after analysis that reviews with same parameters other than ProductId belonged to the same product just having different flavour or quantity. Hence in order to reduce redundancy it was decided to eliminate the rows having same parameters.

The method used for the same was that we first sort the data according to ProductId and then just keep the first similar product review and delelte the others. for eg. in the above just the review for ProductId=B000HDL1RQ remains. This method ensures that there is only one representative for each product and deduplication without sorting would lead to possibility of different representatives still existing for the same product.

In [4]:

```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, k
```

In [5]:

```
#Deduplication of entries
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='firs
final.shape
```

Out[5]:

```
(364173, 10)
```

In [6]:

```
#Checking to see how much % of data still remains
(final['Id'].size*1.0)/(filtered_data['Id'].size*1.0)*100
```

Out[6]:

```
69.25890143662969
```

Observation:- It was also seen that in two rows given below the value of HelpfulnessNumerator is greater than HelpfulnessDenominator which is not practically possible hence these two rows too are removed from calcualtions

In [8]:

```
display= pd.read_sql_query("""
SELECT *
FROM Reviews
WHERE Score != 3 AND Id=44737 OR Id=64422
ORDER BY ProductID
""", con)
display
```

Out[8]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenc
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	

In [9]:

```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
```

In [10]:

```
#Before starting the next phase of preprocessing lets see the number of entries left
print(final.shape)

#How many positive and negative reviews are present in our dataset?
final['Score'].value_counts()
```

```
(364171, 10)
```

Out[10]:

```
positive    307061
negative     57110
Name: Score, dtype: int64
```

7.2.3 Text Preprocessing: Stemming, stop-word removal and Lemmatization.

Now that we have finished deduplication our data requires some preprocessing before we go on further with analysis and making the prediction model.

Hence in the Preprocessing phase we do the following in the order below:-

1. Begin by removing the html tags
2. Remove any punctuations or limited set of special characters like , or . or # etc.
3. Check if the word is made up of english letters and is not alpha-numeric
4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)
5. Convert the word to lowercase
6. Remove Stopwords
7. Finally Snowball Stemming the word (it was obsereved to be better than Porter Stemming)

After which we collect the words used to describe positive and negative reviews

In [30]:

```
# find sentences containing HTML tags
i=0;
for sent in final['Text'].values:
    if (len(re.findall('<.*?>', sent))):
        print(i)
        print(sent)
        break;
    i += 1;
```

```
6
I set aside at least an hour each day to read to my son (3 y/o). At this poi
nt, I consider myself a connoisseur of children's books and this is one of t
he best. Santa Clause put this under the tree. Since then, we've read it per
petually and he loves it.<br /><br />First, this book taught him the months
of the year.<br /><br />Second, it's a pleasure to read. Well suited to 1.5
y/o old to 4+.<br /><br />Very few children's books are worth owning. Most s
hould be borrowed from the library. This book, however, deserves a permanent
spot on your shelf. Sendak's best.
```

In [5]:

```
import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

stop = set(stopwords.words('english')) #set of stopwords
sno = nltk.stem.SnowballStemmer('english') #initialising the snowball stemmer

def cleanhtml(sentence): #function to clean the word of any html-tags
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', sentence)
    return cleantext
def cleanpunc(sentence): #function to clean the word of any punctuation or special characte
    cleaned = re.sub(r'[?|!|\'|"|#]',r'',sentence)
    cleaned = re.sub(r'[.|,|)|(|\|/]',r' ',cleaned)
    return  cleaned
print(stop)
print('************************************')
print(sno.stem('tasty'))
```

```
{'their', 'isn', 'such', 'where', 'this', 'they', 'while', 'about', 'there',
'myself', 'from', 'mightn', 'was', 'between', 'who', 'are', 'only', 'our',
'those', 'through', 'any', 'is', 'a', 'nor', 'mustn', 'shouldn', 'yourself',
'no', 'itself', 'that', 'himself', 'out', 'what', 'my', 'against', 'below',
's', 'for', 'be', 'into', 'few', 'needn', 'you', 'aren', 'when', 'all', 'hi
m', 'but', 've', 'yours', 'being', 'why', 'own', 'up', 'whom', 're', 'and',
'she', 'me', 'of', 'than', 'doesn', 'both', 'same', 'too', 'am', 'how', 'no
t', 'her', 'd', 'until', 'o', 'your', 'yourselves', 'by', 'other', 'once',
'an', 'just', 'to', 'these', 'don', 'its', 'haven', 'having', 'some', 'sha
n', 'theirs', 'under', 'we', 'ain', 'it', 'at', 'in', 'y', 'the', 'off', 'he
rself', 'down', 'because', 'i', 'now', 'themselves', 'each', 'or', 'were',
'if', 'can', 'did', 'm', 'which', 'couldn', 'ourselves', 'hadn', 'has', 'was
n', 'with', 'here', 'further', 'them', 'hasn', 'should', 'ma', 'then', 'he',
'very', 'above', 'been', 'didn', 'during', 'most', 'hers', 'will', 'have',
'doing', 'again', 'had', 'do', 'before', 'as', 'wouldn', 'his', 'after', 'ou
rs', 'does', 'so', 'on', 'more', 't', 'won', 'weren', 'over', 'll'}
************************************
tasti
```

In [12]:

```
#Code for implementing step-by-step the checks mentioned in the pre-processing phase
# this code takes a while to run as it needs to run on 500k sentences.
i=0
str1=' '
final_string=[]
all_positive_words=[] # store words from +ve reviews here
all_negative_words=[] # store words from -ve reviews here.
s=''
for sent in final['Text'].values:
    filtered_sentence=[]
    #print(sent);
    sent=cleanhtml(sent) # remove HTML tags
    for w in sent.split():
        for cleaned_words in cleanpunc(w).split():
            if((cleaned_words.isalpha()) & (len(cleaned_words)>2)):
                if(cleaned_words.lower() not in stop):
                    s=(sno.stem(cleaned_words.lower())).encode('utf8')
                    filtered_sentence.append(s)
                    if (final['Score'].values)[i] == 'positive':
                        all_positive_words.append(s) #list of all words used to describe po
                    if(final['Score'].values)[i] == 'negative':
                        all_negative_words.append(s) #list of all words used to describe ne
                else:
                    continue
            else:
                continue
    #print(filtered_sentence)
    str1 = b" ".join(filtered_sentence) #final string of cleaned words
    #print("*****************************************************************")

    final_string.append(str1)
    i+=1
```

In [31]:

```
final['CleanedText']=final_string #adding a column of CleanedText which displays the data a
```

In [55]:

```
final.head(3) #below the processed review can be seen in the CleanedText Column


# store final table into an SQLite table for future.
conn = sqlite3.connect('final.sqlite')
c=conn.cursor()
conn.text_factory = str
final.to_sql('Reviews', conn, flavor=None, schema=None, if_exists='replace', index=True, ir
```

[7.2.2] Bag of Words (BoW)

In [82]:

```
#BoW
count_vect = CountVectorizer() #in scikit-learn
final_counts = count_vect.fit_transform(final['Text'].values)
```

In [83]:

```
type(final_counts)
```

Out[83]:

```
scipy.sparse.csr.csr_matrix
```

In [84]:

```
final_counts.get_shape()
```

Out[84]:

```
(364171, 115281)
```

[7.2.4] Bi-Grams and n-Grams.

Motivation

Now that we have our list of words describing positive and negative reviews lets analyse them.

We begin analysis by getting the frequency distribution of the words as shown below

In [44]:

```
freq_dist_positive=nltk.FreqDist(all_positive_words)
freq_dist_negative=nltk.FreqDist(all_negative_words)
print("Most Common Positive Words : ",freq_dist_positive.most_common(20))
print("Most Common Negative Words : ",freq_dist_negative.most_common(20))
```

```
Most Common Positive Words :  [(b'like', 139429), (b'tast', 129047), (b'goo
d', 112766), (b'flavor', 109624), (b'love', 107357), (b'use', 103888), (b'gr
eat', 103870), (b'one', 96726), (b'product', 91033), (b'tri', 86791), (b'te
a', 83888), (b'coffe', 78814), (b'make', 75107), (b'get', 72125), (b'food',
64802), (b'would', 55568), (b'time', 55264), (b'buy', 54198), (b'realli', 52
715), (b'eat', 52004)]
Most Common Negative Words :  [(b'tast', 34585), (b'like', 32330), (b'produc
t', 28218), (b'one', 20569), (b'flavor', 19575), (b'would', 17972), (b'tri',
17753), (b'use', 15302), (b'good', 15041), (b'coffe', 14716), (b'get', 1378
6), (b'buy', 13752), (b'order', 12871), (b'food', 12754), (b'dont', 11877),
(b'tea', 11665), (b'even', 11085), (b'box', 10844), (b'amazon', 10073), (b'm
ake', 9840)]
```

Observation:- From the above it can be seen that the most common positive and the negative words overlap for eg. 'like' could be used as 'not like' etc.
So, it is a good idea to consider pairs of consequent words (bi-grams) or q sequnce of n consecutive words (n-grams)

In [79]:

```
#bi-gram, tri-gram and n-gram

#removing stop words like "not" should be avoided before building n-grams
count_vect = CountVectorizer(ngram_range=(1,2) ) #in scikit-learn
final_bigram_counts = count_vect.fit_transform(final['Text'].values)
```

In [81]:

```
final_bigram_counts.get_shape()
```

Out[81]:

```
(364171, 2910192)
```

[7.2.5] TF-IDF

In [91]:

```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2))
final_tf_idf = tf_idf_vect.fit_transform(final['Text'].values)
```

In [92]:

```
final_tf_idf.get_shape()
```

Out[92]:

```
(364171, 2910192)
```

In [106]:

```
features = tf_idf_vect.get_feature_names()
len(features)
```

Out[106]:

```
2910192
```

In [111]:

```
features[100000:100010]
```

Out[111]:

```
['ales until',
 'ales ve',
 'ales would',
 'ales you',
 'alessandra',
 'alessandra ambrosia',
 'alessi',
 'alessi added',
 'alessi also',
 'alessi and']
```

In [127]:

```
# covnert a row in saprsematrix to a numpy array
print(final_tf_idf[3,:].toarray()[0])
```

```
[ 0.  0.  0. ...,  0.  0.  0.]
```

In [123]:

```
# source: https://buhrmann.github.io/tfidf-analysis.html
def top_tfidf_feats(row, features, top_n=25):
    ''' Get top n tfidf values in row and return them with their corresponding feature name
    topn_ids = np.argsort(row)[::-1][:top_n]
    top_feats = [(features[i], row[i]) for i in topn_ids]
    df = pd.DataFrame(top_feats)
    df.columns = ['feature', 'tfidf']
    return df

top_tfidf = top_tfidf_feats(final_tf_idf[1,:].toarray()[0],features,25)
```

In [124]:

```
top_tfidf
```

Out[124]:

	feature	tfidf
0	sendak books	0.173437
1	rosie movie	0.173437
2	paperbacks seem	0.173437
3	cover version	0.173437
4	these sendak	0.173437
5	the paperbacks	0.173437
6	pages open	0.173437
7	really rosie	0.168074
8	incorporates them	0.168074
9	paperbacks	0.168074
10	however miss	0.164269
11	hard cover	0.164269
12	seem kind	0.161317
13	up reading	0.156867
14	that incorporates	0.155100
15	the pages	0.149737
16	sendak	0.149737
17	rosie	0.146786
18	of flimsy	0.146786
19	two hands	0.145130
20	movie that	0.144374
21	reading these	0.137184
22	too do	0.134491
23	incorporates	0.134147
24	flimsy and	0.132254

[7.2.6] Word2Vec

In [7]:

```
# Using Google News Word2Vectors
from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

# in this project we are using a pretrained model by google
# its 3.3G file, once you load this into your memory
# it occupies ~9Gb, so please do this step only if you have >12G of ram
# we will provide a pickle file wich contains a dict ,
# and it contains all our courpus words as keys and  model[word] as values
# To use this code-snippet, download "GoogleNews-vectors-negative300.bin"
# from https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
# it's 1.9GB in size.


model = KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin', binary=True
```

In [129]:

```
model.wv['computer']
```

Out[129]:

```
array([  1.07421875e-01,  -2.01171875e-01,   1.23046875e-01,
         2.11914062e-01,  -9.13085938e-02,   2.16796875e-01,
        -1.31835938e-01,   8.30078125e-02,   2.02148438e-01,
         4.78515625e-02,   3.66210938e-02,  -2.45361328e-02,
         2.39257812e-02,  -1.60156250e-01,  -2.61230469e-02,
         9.71679688e-02,  -6.34765625e-02,   1.84570312e-01,
         1.70898438e-01,  -1.63085938e-01,  -1.09375000e-01,
         1.49414062e-01,  -4.65393066e-04,   9.61914062e-02,
         1.68945312e-01,   2.60925293e-03,   8.93554688e-02,
         6.49414062e-02,   3.56445312e-02,  -6.93359375e-02,
        -1.46484375e-01,  -1.21093750e-01,  -2.27539062e-01,
         2.45361328e-02,  -1.24511719e-01,  -3.18359375e-01,
        -2.20703125e-01,   1.30859375e-01,   3.66210938e-02,
        -3.63769531e-02,  -1.13281250e-01,   1.95312500e-01,
         9.76562500e-02,   1.26953125e-01,   6.59179688e-02,
         6.93359375e-02,   1.02539062e-02,   1.75781250e-01,
        -1.68945312e-01,   1.21307373e-03,  -2.98828125e-01,
        -1.15234375e-01,   5.66406250e-02,  -1.77734375e-01,
        -2.08984375e-01,   1.76757812e-01,   2.38037109e-02,
        -2.57812500e-01,  -4.46777344e-02,   1.88476562e-01,
         5.51757812e-02,   5.02929688e-02,  -1.06933594e-01,
         1.89453125e-01,  -1.16210938e-01,   8.49609375e-02,
        -1.71875000e-01,   2.45117188e-01,  -1.73828125e-01,
        -8.30078125e-03,   4.56542969e-02,  -1.61132812e-02,
         1.86523438e-01,  -6.05468750e-02,  -4.17480469e-02,
         1.82617188e-01,   2.20703125e-01,  -1.22558594e-01,
        -2.55126953e-02,  -3.08593750e-01,   9.13085938e-02,
         1.60156250e-01,   1.70898438e-01,   1.19628906e-01,
         7.08007812e-02,  -2.64892578e-02,  -3.08837891e-02,
         4.06250000e-01,  -1.01562500e-01,   5.71289062e-02,
        -7.26318359e-03,  -9.17968750e-02,  -1.50390625e-01,
        -2.55859375e-01,   2.16796875e-01,  -3.63769531e-02,
         2.24609375e-01,   8.00781250e-02,   1.56250000e-01,
         5.27343750e-02,   1.50390625e-01,  -1.14746094e-01,
        -8.64257812e-02,   1.19140625e-01,  -7.17773438e-02,
         2.73437500e-01,  -1.64062500e-01,   7.29370117e-03,
         4.21875000e-01,  -1.12792969e-01,  -1.35742188e-01,
        -1.31835938e-01,  -1.37695312e-01,  -7.66601562e-02,
         6.25000000e-02,   4.98046875e-02,  -1.91406250e-01,
        -6.03027344e-02,   2.27539062e-01,   5.88378906e-02,
        -3.24218750e-01,   5.41992188e-02,  -1.35742188e-01,
         8.17871094e-03,  -5.24902344e-02,  -1.74713135e-03,
        -9.81445312e-02,  -2.86865234e-02,   3.61328125e-02,
         2.15820312e-01,   5.98144531e-02,  -3.08593750e-01,
        -2.27539062e-01,   2.61718750e-01,   9.86328125e-02,
        -5.07812500e-02,   1.78222656e-02,   1.31835938e-01,
        -5.35156250e-01,  -1.81640625e-01,   1.38671875e-01,
        -3.10546875e-01,  -9.71679688e-02,   1.31835938e-01,
        -1.16210938e-01,   7.03125000e-02,   2.85156250e-01,
         3.51562500e-02,  -1.01562500e-01,  -3.75976562e-02,
         1.41601562e-01,   1.42578125e-01,  -5.68847656e-02,
         2.65625000e-01,  -2.09960938e-01,   9.64355469e-03,
        -6.68945312e-02,  -4.83398438e-02,  -6.10351562e-02,
         2.45117188e-01,  -9.66796875e-02,   1.78222656e-02,
```

```
       -1.27929688e-01,  -4.78515625e-02,  -7.26318359e-03,
        1.79687500e-01,   2.78320312e-02,  -2.10937500e-01,
       -1.43554688e-01,  -1.27929688e-01,   1.73339844e-02,
       -3.60107422e-03,  -2.04101562e-01,   3.63159180e-03,
       -1.19628906e-01,  -6.15234375e-02,   5.93261719e-02,
       -3.23486328e-03,  -1.70898438e-01,  -3.14941406e-02,
       -8.88671875e-02,  -2.89062500e-01,   3.44238281e-02,
       -1.87500000e-01,   2.94921875e-01,   1.58203125e-01,
       -1.19628906e-01,   7.61718750e-02,   6.39648438e-02,
       -4.68750000e-02,  -6.83593750e-02,   1.21459961e-02,
       -1.44531250e-01,   4.54101562e-02,   3.68652344e-02,
        3.88671875e-01,   1.45507812e-01,  -2.55859375e-01,
       -4.46777344e-02,  -1.33789062e-01,  -1.38671875e-01,
        6.59179688e-02,   1.37695312e-01,   1.14746094e-01,
        2.03125000e-01,  -4.78515625e-02,   1.80664062e-02,
       -8.54492188e-02,  -2.48046875e-01,  -3.39843750e-01,
       -2.83203125e-02,   1.05468750e-01,  -2.14843750e-01,
       -8.74023438e-02,   7.12890625e-02,   1.87500000e-01,
       -1.12304688e-01,   2.73437500e-01,  -3.26171875e-01,
       -1.77734375e-01,  -4.24804688e-02,  -2.69531250e-01,
        6.64062500e-02,  -6.88476562e-02,  -1.99218750e-01,
       -7.03125000e-02,  -2.43164062e-01,  -3.66210938e-02,
       -7.37304688e-02,  -1.77734375e-01,   9.17968750e-02,
       -1.25000000e-01,  -1.65039062e-01,  -3.57421875e-01,
       -2.85156250e-01,  -1.66992188e-01,   1.97265625e-01,
       -1.53320312e-01,   2.31933594e-02,   2.06054688e-01,
        1.80664062e-01,  -2.74658203e-02,  -1.92382812e-01,
       -9.61914062e-02,  -1.06811523e-02,  -4.73632812e-02,
        6.54296875e-02,  -1.25732422e-02,   1.78222656e-02,
       -8.00781250e-02,  -2.59765625e-01,   9.37500000e-02,
       -7.81250000e-02,   4.68750000e-02,  -2.22167969e-02,
        1.86767578e-02,   3.11279297e-02,   1.04980469e-02,
       -1.69921875e-01,   2.58789062e-02,  -3.41796875e-02,
       -1.44042969e-02,  -5.46875000e-02,  -8.78906250e-02,
        1.96838379e-03,   2.23632812e-01,  -1.36718750e-01,
        1.75781250e-01,  -1.63085938e-01,   1.87500000e-01,
        3.44238281e-02,  -5.63964844e-02,  -2.27689743e-05,
        4.27246094e-02,   5.81054688e-02,  -1.07910156e-01,
       -3.88183594e-02,  -2.69531250e-01,   3.34472656e-02,
        9.81445312e-02,   5.63964844e-02,   2.23632812e-01,
       -5.49316406e-02,   1.46484375e-01,   5.93261719e-02,
       -2.19726562e-01,   6.39648438e-02,   1.66015625e-02,
        4.56542969e-02,   3.26171875e-01,  -3.80859375e-01,
        1.70898438e-01,   5.66406250e-02,  -1.04492188e-01,
        1.38671875e-01,  -1.57226562e-01,   3.23486328e-03,
       -4.80957031e-02,  -2.48046875e-01,  -6.20117188e-02], dtype=float32)
```

In [12]:

```
model.wv.similarity('woman', 'man')
```

Out[12]:

```
0.76640122309953518
```

In [131]:

```
model.wv.most_similar('woman')
```

Out[131]:

```
[('man', 0.7664012312889099),
 ('girl', 0.7494641542434692),
 ('teenage_girl', 0.7336830496788025),
 ('teenager', 0.6317086219787598),
 ('lady', 0.6288787126541138),
 ('teenaged_girl', 0.6141784191131592),
 ('mother', 0.607630729675293),
 ('policewoman', 0.6069462299346924),
 ('boy', 0.5975908041000366),
 ('Woman', 0.5770982503890991)]
```

In [14]:

```
model.wv.most_similar('tasti')  # "tasti" is the stemmed word for tasty, tastful
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
<ipython-input-14-275c3585c172> in <module>()
----> 1 model.wv.most_similar('tasti')  # "tasti" is the stemmed word for ta
sty, tastful

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site-package
s/gensim/models/keyedvectors.py in most_similar(self, positive, negative, to
pn, restrict_vocab, indexer)
    334                 mean.append(weight * word)
    335             else:
--> 336                 mean.append(weight * self.word_vec(word, use_norm=Tr
ue))
    337                 if word in self.vocab:
    338                     all_words.add(self.vocab[word].index)

/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site-package
s/gensim/models/keyedvectors.py in word_vec(self, word, use_norm)
    282                 return self.syn0[self.vocab[word].index]
    283         else:
--> 284             raise KeyError("word '%s' not in vocabulary" % word)
    285
    286     def most_similar(self, positive=None, negative=None, topn=10, re
strict_vocab=None, indexer=None):

KeyError: "word 'tasti' not in vocabulary"
```

In [155]:

```
model.wv.most_similar('tasty')
```

Out[155]:

```
[('delicious', 0.8730389475822449),
 ('scrumptious', 0.8007042407989502),
 ('yummy', 0.7856923341751099),
 ('flavorful', 0.7420164346694946),
 ('delectable', 0.7385422587394714),
 ('juicy_flavorful', 0.7114803791046143),
 ('appetizing', 0.701721727848053),
 ('crunchy_salty', 0.7012301087379456),
 ('flavourful', 0.6912214159965515),
 ('flavoursome', 0.6857703328132629)]
```

In [137]:

```
model.wv.similarity('tasty', 'tast')
```

Out[137]:

```
0.44035054190088901
```

In [185]:

```
# Train your own Word2Vec model using your own text corpus
import gensim
i=0
list_of_sent=[]
for sent in final['Text'].values:
    filtered_sentence=[]
    sent=cleanhtml(sent)
    for w in sent.split():
        for cleaned_words in cleanpunc(w).split():
            if(cleaned_words.isalpha()):
                filtered_sentence.append(cleaned_words.lower())
            else:
                continue
    list_of_sent.append(filtered_sentence)
```

In [186]:

```
print(final['Text'].values[0])
print("*****************************************************************")
print(list_of_sent[0])
```

```
this witty little book makes my son laugh at loud. i recite it in the car as
we're driving along and he always can sing the refrain. he's learned about w
hales, India, drooping roses:  i love all the new words this book  introduce
s and the silliness of it all.  this is a classic book i am  willing to bet
my son will STILL be able to recite from memory when he is  in college
*****************************************************************
['this', 'witty', 'little', 'book', 'makes', 'my', 'son', 'laugh', 'at', 'lo
ud', 'i', 'recite', 'it', 'in', 'the', 'car', 'as', 'were', 'driving', 'alon
g', 'and', 'he', 'always', 'can', 'sing', 'the', 'refrain', 'hes', 'learne
d', 'about', 'whales', 'india', 'drooping', 'i', 'love', 'all', 'the', 'ne
w', 'words', 'this', 'book', 'introduces', 'and', 'the', 'silliness', 'of',
'it', 'all', 'this', 'is', 'a', 'classic', 'book', 'i', 'am', 'willing', 't
o', 'bet', 'my', 'son', 'will', 'still', 'be', 'able', 'to', 'recite', 'fro
m', 'memory', 'when', 'he', 'is', 'in', 'college']
```

In [187]:

```
w2v_model=gensim.models.Word2Vec(list_of_sent,min_count=5,size=50, workers=4)
```

In [190]:

```
words = list(w2v_model.wv.vocab)
print(len(words))
```

```
33783
```

In [191]:

```
w2v_model.wv.most_similar('tasty')
```

Out[191]:

```
[('tastey', 0.909038245677948),
 ('satisfying', 0.8556904792785645),
 ('yummy', 0.8543208837509155),
 ('filling', 0.8233586549758911),
 ('delicious', 0.8229926228523254),
 ('flavorful', 0.8061250448226929),
 ('addicting', 0.771919846534729),
 ('delish', 0.7653154730796814),
 ('nutritious', 0.7626035213470459),
 ('tasteful', 0.7547359466552734)]
```

In [199]:

```
w2v_model.wv.most_similar('like')
```

Out[199]:

```
[('resemble', 0.7139369249343872),
 ('mean', 0.651788055896759),
 ('prefer', 0.646423876285553),
 ('dislike', 0.6413203477859497),
 ('overpower', 0.6197512745857239),
 ('think', 0.5993291735649109),
 ('overwhelm', 0.5929116606712341),
 ('enjoy', 0.5895069241523743),
 ('gross', 0.5853881239891052),
 ('alright', 0.5824472308158875)]
```

In [204]:

```
count_vect_feat = count_vect.get_feature_names() # list of words in the BoW
count_vect_feat.index('like')
print(count_vect_feat[64055])
```

```
like
```

[7.2.7] Avg W2V, TFIDF-W2V

In [215]:

```
# average Word2Vec
# compute average word2vec for each review.
sent_vectors = []; # the avg-w2v for each sentence/review is stored in this list
for sent in list_of_sent: # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        try:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
        except:
            pass
    sent_vec /= cnt_words
    sent_vectors.append(sent_vec)
print(len(sent_vectors))
print(len(sent_vectors[0]))
```

```
/Library/Frameworks/Python.framework/Versions/3.6/lib/python3.6/site-package
s/ipykernel_launcher.py:14: RuntimeWarning: invalid value encountered in tru
e_divide


364171
50
```

In [ ]:

```
# TF-IDF weighted Word2Vec
tfidf_feat = tf_idf_vect.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf

tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in list_of_sent: # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        try:
            vec = w2v_model.wv[word]
            # obtain the tf_idfidf of a word in a sentence/review
            tfidf = final_tf_idf[row, tfidf_feat.index(word)]
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
        except:
            pass
    sent_vec /= weight_sum
    tfidf_sent_vectors.append(sent_vec)
    row += 1
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

```
```