Amazon Fine Food Reviews Analysis

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

EDA: https://nycdatascience.com/blog/student-works/amazon-fine-foods-visualization/

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.

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 [55]:

```
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")



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.model_selection import train_test_split
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

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

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

from tqdm import tqdm
import os
```

[1]. Reading Data

In [2]:

```
# using the SQLite Table to read data.
con = sqlite3.connect('../../dataset/database.sqlite')
#filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
# SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000, will give top 500000 data points
# you can change the number to any other number based on your computing power

# filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 LIMIT 500000""", con)
# for tsne assignment you can take 5k data points

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 0
    return 1

#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
print("Number of data points in our data", filtered_data.shape)
filtered_data.head(3)
```

```
Number of data points in our data (525814, 10)
```

Out[2]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	1	1303862400	Good Quality Dog Food
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	0	1346976000	Not as Advertised
				Natalia					

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
2	3	B000LQOCH0	ABXLMWJIXXAIN	"Natalia Corres"	1	1	1	1219017600	"Delight" says it all

In [3]:

```
display = pd.read_sql_query("""
SELECT UserId, ProductId, ProfileName, Time, Score, Text, COUNT(*)
FROM Reviews
GROUP BY UserId
HAVING COUNT(*)>1
""", con)
```

In [4]:

```
print(display.shape)
display.head()
```

```
(80668, 7)
```

Out[4]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
0	#oc-R115TNMSPFT9I7	B005ZBZLT4	Breyton	1331510400	2	Overall its just OK when considering the price...	2
1	#oc-R11D9D7SHXIJB9	B005HG9ESG	Louis E. Emory "hoppy"	1342396800	5	My wife has recurring extreme muscle spasms, u...	3
2	#oc-R11DNU2NBKQ23Z	B005ZBZLT4	Kim Cieszykowski	1348531200	1	This coffee is horrible and unfortunately not ...	2
3	#oc-R11O5J5ZVQE25C	B005HG9ESG	Penguin Chick	1346889600	5	This will be the bottle that you grab from the...	3
4	#oc-R12KPBODL2B5ZD	B007OSBEV0	Christopher P. Presta	1348617600	1	I didnt like this coffee. Instead of telling y...	2

In [5]:

```
display[display['UserId']=='AZY10LLTJ71NX']
```

Out[5]:

	UserId	ProductId	ProfileName	Time	Score	Text	COUNT(*)
80638	AZY10LLTJ71NX	B001ATMQK2	undertheshrine "undertheshrine"	1296691200	5	I bought this 6 pack because for the price tha...	5

In [6]:

```
display['COUNT(*)'].sum()
```

Out[6]:

```
393063
```

Exploratory Data Analysis

[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 [7]:

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

Out[7]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Sumn
0	78445	B000HDL1RQ	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACH QUADRA VANI WAFE
1	138317	B000HDOPYC	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACH QUADRA VANI WAFE
2	138277	B000HDOPYM	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACH QUADRA VANI WAFE
3	73791	B000HDOPZG	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACH QUADRA VANI WAFE
4	155049	B000PAQ75C	AR5J8UI46CURR	Geetha Krishnan	2	2	5	1199577600	LOACH QUADRA VANI WAFE

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 [8]:

```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')
```

In [9]:

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

Out[9]:

```
(364173, 10)
```

In [10]:

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

Out[10]:

```
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 [11]:

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

display.head()
```

Out[11]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
0	64422	B000MIDROQ	A161DK06JJMCYF	J. E. Stephens "Jeanne"	3	1	5	1224892800	Bough This fo My Son a College
1	44737	B001EQ55RW	A2V0I904FH7ABY	Ram	3	2	4	1212883200	Pure cocoa taste with crunchy almonds inside

In [12]:

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

In [13]:

```
#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[13]:

```
1    307061
0     57110
Name: Score, dtype: int64
```

[3]. Text Preprocessing.

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 [15]:

```
# printing some random reviews
sent_0 = final['Text'].values[0]
print(sent_0)
print("="*50)

sent_1000 = final['Text'].values[1000]
print(sent_1000)
print("="*50)

sent_1500 = final['Text'].values[1500]
print(sent_1500)
print("="*50)

sent_4900 = final['Text'].values[4900]
print(sent_4900)
print("="*50)
```

```
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 whales, India, drooping roses:  i love all the new word
s this book  introduces and the silliness of it all.  this is a classic book i am  willing to bet my so
n will STILL be able to recite from memory when he is  in college
==================================================
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back in 2005 fo
r gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, 2 year
s expired!!!  I'm hoping to find local San Diego area shoppe that carries pods so that I can try someth
ing different than starbucks.
==================================================
Great ingredients although, chicken should have been 1st rather than chicken broth, the only thing I do
not think belongs in it is Canola oil. Canola or rapeseed is not someting a dog would ever find in natu
re and if it did find rapeseed in nature and eat it, it would poison them. Today's Food industries have
convinced the masses that Canola oil is a safe and even better oil than olive or virgin coconut, facts
though say otherwise. Until the late 70's it was poisonous until they figured out a way to fix that. I
still like it but it could be better.
==================================================
Can't do sugar.  Have tried scores of SF Syrups.  NONE of them can touch the excellence of this product
.<br /><br />Thick, delicious.  Perfect.  3 ingredients: Water, Maltitol, Natural Maple Flavor.  PERIOD
.  No chemicals.  No garbage.<br /><br />Have numerous friends & family members hooked on this stuff.
My husband & son, who do NOT like "sugar free" prefer this over major label regular syrup.<br /><br />I
use this as my SWEETENER in baking: cheesecakes, white brownies, muffins, pumpkin pies, etc... Unbeliev
ably delicious...<br /><br />Can you tell I like it? :)
==================================================
```

In [16]:

```
# remove urls from text python: https://stackoverflow.com/a/40823105/4084039
sent_0 = re.sub(r"http\S+", "", sent_0)
sent_1000 = re.sub(r"http\S+", "", sent_1000)
sent_150 = re.sub(r"http\S+", "", sent_1500)
sent_4900 = re.sub(r"http\S+", "", sent_4900)

print(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 whales, India, drooping roses:  i love all the new word
```

s this book introduces and the silliness of it all. this is a classic book i am willing to bet my so
n will STILL be able to recite from memory when he is in college

In [17]:

```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an-elem
ent
from bs4 import BeautifulSoup

soup = BeautifulSoup(sent_0, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_1000, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_1500, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(sent_4900, 'lxml')
text = soup.get_text()
print(text)
```

```
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 whales, India, drooping roses:  i love all the new word
s this book  introduces and the silliness of it all.  this is a classic book i am  willing to bet my so
n will STILL be able to recite from memory when he is  in college
==================================================
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back in 2005 fo
r gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, 2 year
s expired!!!  I'm hoping to find local San Diego area shoppe that carries pods so that I can try someth
ing different than starbucks.
==================================================
Great ingredients although, chicken should have been 1st rather than chicken broth, the only thing I do
not think belongs in it is Canola oil. Canola or rapeseed is not someting a dog would ever find in natu
re and if it did find rapeseed in nature and eat it, it would poison them. Today's Food industries have
convinced the masses that Canola oil is a safe and even better oil than olive or virgin coconut, facts
though say otherwise. Until the late 70's it was poisonous until they figured out a way to fix that. I
still like it but it could be better.
==================================================
Can't do sugar.  Have tried scores of SF Syrups.  NONE of them can touch the excellence of this product
.Thick, delicious.  Perfect.  3 ingredients: Water, Maltitol, Natural Maple Flavor.  PERIOD.  No chemic
als.  No garbage.Have numerous friends & family members hooked on this stuff.  My husband & son, who do
NOT like "sugar free" prefer this over major label regular syrup.I use this as my SWEETENER in baking:
cheesecakes, white brownies, muffins, pumpkin pies, etc... Unbelievably delicious...Can you tell I like
it? :)
```

In [18]:

```
# https://stackoverflow.com/a/47091490/4084039
import re

def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can\'t", "can not", phrase)

    # general
    phrase = re.sub(r"n\'t", " not", phrase)
    phrase = re.sub(r"\'re", " are", phrase)
    phrase = re.sub(r"\'s", " is", phrase)
    phrase = re.sub(r"\'d", " would", phrase)
    phrase = re.sub(r"\'ll", " will", phrase)
    phrase = re.sub(r"\'t", " not", phrase)
    phrase = re.sub(r"\'ve", " have", phrase)
    phrase = re.sub(r"\'m", " am", phrase)
    return phrase
```

In [19]:

```
sent_1500 = decontracted(sent_1500)
print(sent_1500)
print("="*50)
```

```
Great ingredients although, chicken should have been 1st rather than chicken broth, the only thing I do
not think belongs in it is Canola oil. Canola or rapeseed is not someting a dog would ever find in natu
re and if it did find rapeseed in nature and eat it, it would poison them. Today is Food industries hav
e convinced the masses that Canola oil is a safe and even better oil than olive or virgin coconut, fact
s though say otherwise. Until the late 70 is it was poisonous until they figured out a way to fix that.
I still like it but it could be better.
==================================================
```

In [20]:

```
#remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(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 whales, India, drooping roses:  i love all the new word
s this book  introduces and the silliness of it all.  this is a classic book i am  willing to bet my so
n will STILL be able to recite from memory when he is  in college
```

In [21]:

```
#remove spacial character: https://stackoverflow.com/a/5843547/4084039
sent_1500 = re.sub('[^A-Za-z0-9]+', ' ', sent_1500)
print(sent_1500)
```

```
Great ingredients although chicken should have been 1st rather than chicken broth the only thing I do n
ot think belongs in it is Canola oil Canola or rapeseed is not someting a dog would ever find in nature
and if it did find rapeseed in nature and eat it it would poison them Today is Food industries have con
vinced the masses that Canola oil is a safe and even better oil than olive or virgin coconut facts thou
gh say otherwise Until the late 70 is it was poisonous until they figured out a way to fix that I still
like it but it could be better
```

In [22]:

```
# https://gist.github.com/sebleier/554280
# we are removing the words from the stop words list: 'no', 'nor', 'not'
# <br /><br /> ==> after the above steps, we are getting "br br"
# we are including them into stop words list
# instead of <br /> if we have <br/> these tags would have revmoved in the 1st step

stopwords= set(['br', 'the', 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",\
            "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', \
            'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their',\
            'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', \
            'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', \
            'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', \
            'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after',\
            'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further',\
            'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more',\
            'most', 'other', 'some', 'such', 'only', 'own', 'same', 'so', 'than', 'too', 'very', \
            's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', \
            've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', \
```

```
            "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn',\
            "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't",
'weren', "weren't", \
            'won', "won't", 'wouldn', "wouldn't"])
```

In [23]:

```
# Combining all the above stundents
from tqdm import tqdm
preprocessed_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(final['Text'].values):
    sentance = re.sub(r"http\S+", "", sentance)
    sentance = BeautifulSoup(sentance, 'lxml').get_text()
    sentance = decontracted(sentance)
    sentance = re.sub("\S*\d\S*", "", sentance).strip()
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not in stopwords)
    preprocessed_reviews.append(sentance.strip())
```

```
100%|██████████| 364171/364171 [02:29<00:00, 2435.63it/s]
```

In [24]:

```
preprocessed_reviews[1500]
```

Out[24]:

'great ingredients although chicken rather chicken broth thing not think belongs canola oil canola rape seed not someting dog would ever find nature find rapeseed nature eat would poison today food industrie s convinced masses canola oil safe even better oil olive virgin coconut facts though say otherwise late poisonous figured way fix still like could better'

[3.2] Preprocess Summary

In [25]:

```
## Similartly you can do preprocessing for review summary also.


# printing some random reviews
summary_sent_0 = final['Summary'].values[0]
print(summary_sent_0)
print("="*50)

summary_sent_1000 = final['Summary'].values[1000]
print(summary_sent_1000)
print("="*50)
```

```
EVERY book is educational
==================================================
Shipment was expired by 2 years
==================================================
```

In [26]:

```
# remove urls from text python: https://stackoverflow.com/a/40823105/4084039
summary_sent_0 = re.sub(r"http\S+", "", sent_0)
summary_sent_1000 = re.sub(r"http\S+", "", sent_1000)

print(summary_sent_0)
print(summary_sent_1000)
```

```
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 whales, India, drooping roses:  i love all the new word
s this book  introduces and the silliness of it all.  this is a classic book i am  willing to bet my so
n will STILL be able to recite from memory when he is  in college
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back in 2005 fo
r gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, 2 year
s expired!!!  I'm hoping to find local San Diego area shoppe that carries pods so that I can try someth
ing different than starbucks.
```

In [27]:

```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an-elem
ent
from bs4 import BeautifulSoup

soup = BeautifulSoup(summary_sent_0, 'lxml')
text = soup.get_text()
print(text)
print("="*50)

soup = BeautifulSoup(summary_sent_1000, 'lxml')
text = soup.get_text()
print(text)
print("="*50)
```

```
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 whales, India, drooping roses:  i love all the new word
s this book  introduces and the silliness of it all.  this is a classic book i am  willing to bet my so
n will STILL be able to recite from memory when he is  in college
==================================================
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back in 2005 fo
r gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, 2 year
s expired!!!  I'm hoping to find local San Diego area shoppe that carries pods so that I can try someth
ing different than starbucks.
==================================================
```

In [28]:

```
# https://stackoverflow.com/a/47091490/4084039
import re

def decontracted(phrase):
    # specific
    phrase = re.sub(r"won't", "will not", phrase)
    phrase = re.sub(r"can\'t", "can not", phrase)

    # general
    phrase = re.sub(r"n\'t", " not", phrase)
    phrase = re.sub(r"\'re", " are", phrase)
    phrase = re.sub(r"\'s", " is", phrase)
    phrase = re.sub(r"\'d", " would", phrase)
    phrase = re.sub(r"\'ll", " will", phrase)
    phrase = re.sub(r"\'t", " not", phrase)
    phrase = re.sub(r"\'ve", " have", phrase)
    phrase = re.sub(r"\'m", " am", phrase)
    return phrase
```

In [29]:

```
summary_sent_1000 = decontracted(summary_sent_1000)
print(summary_sent_1000)
print("="*50)
```

```
I was really looking forward to these pods based on the reviews.  Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back in 2005 fo
r gosh sakes.  I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, 2 year
s expired!!!  I am hoping to find local San Diego area shoppe that carries pods so that I can try somet
hing different than starbucks.
==================================================
```

In [30]:

```
#remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
summary_sent_1000 = re.sub("\S*\d\S*", "", summary_sent_1000).strip()
print(sent_1000)
```

I was really looking forward to these pods based on the reviews. Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered 2 boxes - both were expired! One expired back in 2005 fo
r gosh sakes. I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, 2 year
s expired!!! I'm hoping to find local San Diego area shoppe that carries pods so that I can try someth
ing different than starbucks.

In [31]:

```
summaey_sent_1000 = re.sub('[^A-Za-z0-9]+', ' ', summary_sent_1000)
print(summary_sent_1000)
```

I was really looking forward to these pods based on the reviews. Starbucks is good, but I prefer bolde
r taste.... imagine my surprise when I ordered boxes - both were expired! One expired back in for gos
h sakes. I admit that Amazon agreed to credit me for cost plus part of shipping, but geez, years expi
red!!! I am hoping to find local San Diego area shoppe that carries pods so that I can try something d
ifferent than starbucks.

In [32]:

```
from tqdm import tqdm
preprocessed_summary_reviews = []
# tqdm is for printing the status bar
for sentance in tqdm(final['Summary'].values):
    sentance = re.sub(r"http\S+", "", sentance)
    sentance = BeautifulSoup(sentance, 'lxml').get_text()
    sentance = decontracted(sentance)
    sentance = re.sub("\S*\d\S*", "", sentance).strip()
    sentance = re.sub('[^A-Za-z]+', ' ', sentance)
    # https://gist.github.com/sebleier/554280
    sentance = ' '.join(e.lower() for e in sentance.split() if e.lower() not in stopwords)
    preprocessed_summary_reviews.append(sentance.strip())
```

```
  0%|          | 236/364171 [00:00<02:34, 2359.95it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\sit
e-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should pro
bably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
  6%|▌         | 22361/364171 [00:07<01:47, 3175.92it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\s
ite-packages\bs4\__init__.py:273: UserWarning: "b'.'" looks like a filename, not markup. You should pro
bably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 10%|▉         | 35820/364171 [00:12<01:48, 3023.47it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\s
ite-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should p
robably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 10%|▉         | 36735/364171 [00:12<01:56, 2805.78it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\s
ite-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should p
robably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 14%|█▎        | 51637/364171 [00:17<01:40, 3122.45it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\s
ite-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should p
robably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 27%|██▍       | 98225/364171 [00:32<01:23, 3181.51it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\s
ite-packages\bs4\__init__.py:273: UserWarning: "b'.'" looks like a filename, not markup. You should pro
bably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 27%|██▍       | 98551/364171 [00:32<01:26, 3082.26it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\s
ite-packages\bs4\__init__.py:273: UserWarning: "b'.'" looks like a filename, not markup. You should pro
bably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 32%|██▊       | 115285/364171 [00:37<01:41, 2458.43it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
```

```
 40%|███▉      | 144746/364171 [00:52<01:13, 2968.95it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 41%|████      | 149774/364171 [00:53<01:01, 3465.05it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 60%|██████    | 216939/364171 [01:18<00:42, 3458.74it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'.'" looks like a filename, not markup. You should pr
obably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 68%|██████▊   | 248362/364171 [01:29<00:47, 2458.12it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 78%|███████▊  | 282358/364171 [01:39<00:30, 2674.95it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 89%|████████▉ | 323850/364171 [01:56<00:12, 3304.88it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 91%|█████████ | 329720/364171 [01:58<00:10, 3275.92it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 92%|█████████▏| 335064/364171 [02:00<00:09, 3145.01it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 97%|█████████▋| 354680/364171 [02:06<00:04, 2360.54it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'.'" looks like a filename, not markup. You should pr
obably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
 99%|█████████▉| 359104/364171 [02:08<00:01, 3000.50it/s]C:\Users\Srivas26\AppData\Local\anaconda3\lib\
site-packages\bs4\__init__.py:273: UserWarning: "b'...'" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  ' Beautiful Soup.' % markup)
C:\Users\Srivas26\AppData\Local\anaconda3\lib\site-packages\bs4\__init__.py:273: UserWarning: "b'...'"
looks like a filename, not markup. You should probably open this file and pass the filehandle into Beau
tiful Soup.
  ' Beautiful Soup.' % markup)
100%|██████████| 364171/364171 [02:10<00:00, 2796.24it/s]
```

In [34]:

```
preprocessed_reviews[1500]
```

Out[34]:

```
'great ingredients although chicken rather chicken broth thing not think belongs canola oil canola rape
seed not someting dog would ever find nature find rapeseed nature eat would poison today food industrie
s convinced masses canola oil safe even better oil olive virgin coconut facts though say otherwise late
poisonous figured way fix still like could better'
```

In [59]:

```
final['cleaned_text'] = preprocessed_reviews
```

In [62]:

```
final.to_pickle('../../dataset/final_pickle')
```

In [63]:

```
final_sample = final.sample(n=100000)
final_sample.to_pickle('../../dataset/final_100k_pickle')
```

In [64]:

```
final_sample['Score'].value_counts()
```

Out[64]:

```
1    84092
0    15908
Name: Score, dtype: int64
```

In [65]:

```
final_sample.sort_values(['Time'], inplace=True)
```

In [67]:

```
X_train, X_test, Y_train, Y_test = train_test_split(final_sample['cleaned_text'], final_sample['Score']
,
                                                    shuffle=False, test_size=0.3)
```

[4] Featurization

[4.1] BAG OF WORDS

[4.1.1] BAG OF WORDS TRAIN

In [78]:

```
#BoW
count_vect = CountVectorizer() #in scikit-learn
count_vect.fit(X_train)
print("some feature names ", count_vect.get_feature_names()[:10])
print('='*50)

X_train_bow = count_vect.transform(X_train)
print("the shape of out text BOW vectorizer ",X_train_bow.get_shape())
print("the number of unique words ", X_train_bow.get_shape()[1])
```

```
some feature names  ['aa', 'aaa', 'aaaa', 'aaaaa', 'aaaaaa', 'aaaaaaaaaaa', 'aaaaaaaaaaaa', 'aaaaaahh
hhhyaaaaaa', 'aaaahhhhhh', 'aaaand']
==================================================
the shape of out text BOW vectorizer  (70000, 50780)
the number of unique words  50780
```

[4.1.2] BAG OF WORDS TEST

In [81]:

```
X_test_bow = count_vect.transform(X_test)
print("the shape of out text BOW vectorizer TEST",X_test_bow.get_shape())
print("the number of unique words TEST", X_test_bow.get_shape()[1])
```

```
the shape of out text BOW vectorizer TEST (30000, 50780)
the number of unique words TEST 50780
```

[4.2] Bi-Grams and n-Grams.

[4.2.1] Bi-Grams and n-Grams TRAIN.

In [90]:

```
#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))
# please do read the CountVectorizer documentation http://scikit-learn.org/stable/modules/generated/skl
earn.feature_extraction.text.CountVectorizer.html
# you can choose these numebrs min_df=10, max_features=5000, of your choice
count_vect = CountVectorizer(ngram_range=(1,2), min_df=10)
X_train_ngram = count_vect.fit_transform(X_train)
print("some feature names ", count_vect.get_feature_names()[:10])
print("the type of count vectorizer ",type(X_train_ngram))
print("the shape of out text BOW vectorizer ",X_train_ngram.get_shape())
print("the number of unique words including both unigrams and bigrams ", X_train_ngram.get_shape()[1])
```

```
some feature names  ['aa', 'aback', 'abandoned', 'abdominal', 'ability', 'ability make', 'able', 'able
buy', 'able chew', 'able continue']
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (70000, 40729)
the number of unique words including both unigrams and bigrams  40729
```

[4.2.2] Bi-Grams and n-Grams TEST.

In [91]:

```
X_test_ngram = count_vect.transform(X_test)
print("the type of count vectorizer ",type(X_test_ngram))
print("the shape of out text BOW vectorizer ",X_test_ngram.get_shape())
print("the number of unique words including both unigrams and bigrams ", X_test_ngram.get_shape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (30000, 40729)
the number of unique words including both unigrams and bigrams  40729
```

[4.3] TF-IDF

[4.3.1] TF-IDF TRAIN

In [94]:

```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2), min_df=10)
tf_idf_vect.fit(X_train)
print("some sample features(unique words in the corpus)",tf_idf_vect.get_feature_names()[0:10])
print('='*50)

X_train_tf_idf = tf_idf_vect.transform(preprocessed_reviews)
print("the type of count vectorizer ",type(X_train_tf_idf))
print("the shape of out text TFIDF vectorizer ",X_train_tf_idf.get_shape())
print("the number of unique words including both unigrams and bigrams ", X_train_tf_idf.get_shape()[1])
```

```
some sample features(unique words in the corpus) ['aa', 'aback', 'abandoned', 'abdominal', 'ability', '
ability make', 'able', 'able buy', 'able chew', 'able continue']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (364171, 40729)
the number of unique words including both unigrams and bigrams  40729
```

[4.3.2] TF-IDF TEST

In [95]:

```
X_test_tf_idf = tf_idf_vect.transform(X_test)
print("the type of count vectorizer ",type(X_test_tf_idf))
print("the shape of out text TFIDF vectorizer ",X_test_tf_idf.get_shape())
```

```
print("the number of unique words including both unigrams and bigrams ", X_test_tf_idf.get_shape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (30000, 40729)
the number of unique words including both unigrams and bigrams  40729
```

[4.4] Word2Vec

In [110]:

```
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sentance=[]
for sentance in X_train:
    list_of_sentance.append(sentance.split())
```

```
  0%|          | 132/100000 [00:27<2:01:35, 13.69it/s]
```

In [100]:

```
# Using Google News Word2Vectors

# 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.


# http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.W17SRFAzZPY
# you can comment this whole cell
# or change these varible according to your need

is_your_ram_gt_16g=False
want_to_use_google_w2v = False
want_to_train_w2v = True

if want_to_train_w2v:
    # min_count = 5 considers only words that occured atleast 5 times
    w2v_model=Word2Vec(list_of_sentance,min_count=5,size=50, workers=4)
    print(w2v_model.wv.most_similar('great'))
    print('='*50)
    print(w2v_model.wv.most_similar('worst'))

elif want_to_use_google_w2v and is_your_ram_gt_16g:
    if os.path.isfile('GoogleNews-vectors-negative300.bin'):
        w2v_model=KeyedVectors.load_word2vec_format('GoogleNews-vectors-negative300.bin', binary=True)
        print(w2v_model.wv.most_similar('great'))
        print(w2v_model.wv.most_similar('worst'))
    else:
        print("you don't have gogole's word2vec file, keep want_to_train_w2v = True, to train your own w2v ")
```

```
[('awesome', 0.8177148699760437), ('fantastic', 0.8127740025520325), ('good', 0.7955279350280762), ('excellent', 0.7850255966186523), ('wonderful', 0.7742749452590942), ('perfect', 0.7674868106842041), ('terrific', 0.7637571096420288), ('nice', 0.703532338142395), ('amazing', 0.6990883350372314), ('incredible', 0.6870638132095337)]
==================================================
[('best', 0.7580143809318542), ('greatest', 0.7565791010856628), ('disgusting', 0.6997318863868713), ('nastiest', 0.6935060620307922), ('closest', 0.6839133501052856), ('horrid', 0.6341369152069092), ('tastiest', 0.6048495173454285), ('awful', 0.6021798849105835), ('foul', 0.5980766415596008), ('blandest', 0.5850250720977783)]
```

In [101]:

```
w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 5 times ",len(w2v_words))
print("sample words ", w2v_words[0:50])
```

```
number of words that occured minimum 5 times  16100
sample words  ['twist', 'captured', 'film', 'michael', 'keaton', 'geena', 'davis', 'prime', 'tim', 'bur
ton', 'masterpiece', 'wonderfully', 'paced', 'point', 'not', 'dull', 'moment', 'beetlejuice', 'excellen
t', 'funny', 'movie', 'hilarious', 'great', 'special', 'effects', 'help', 'think', 'one', 'best', 'movi
es', 'ever', 'made', 'sure', 'agree', 'good', 'time', 'watch', 'always', 'enjoyed', 'entertaining', 'he
sitate', 'pick', 'edition', 'guess', 'marketing', 'plan', 'make', 'families', 'something', 'eliminated'
]
```

[4.4.1] Converting text into vectors using wAvg W2V, TFIDF-W2V

[4.4.1.1] Avg W2v TRAIN

[4.4.1.1.1] Avg W2v TRAIN

In [102]:

```
# 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 tqdm(list_of_sentance): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change this to 3
00 if you use google's w2v
    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
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors.append(sent_vec)
print(len(sent_vectors))
print(len(sent_vectors[0]))
```

```
100%|██████████| 70000/70000 [02:24<00:00, 484.00it/s]
```

```
70000
50
```

In [114]:

```
X_train_w2v = sent_vectors
```

[4.4.1.1.2] Avg W2v TEST

In [104]:

```
list_of_sentance_test=[]
for sentance in X_test:
    list_of_sentance_test.append(sentance.split())
w2v_model_test=Word2Vec(list_of_sentance_test,min_count=5,size=50, workers=4)
sent_vectors_test = []; # the avg-w2v for each sentence/review is stored in this list
for sent in tqdm(list_of_sentance_test): # for each review/sentence
    sent_vec = np.zeros(50) # as word vectors are of zero length 50, you might need to change this to 3
00 if you use google's w2v
    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
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
```

```
    if cnt_words != 0:
        sent_vec /= cnt_words
    sent_vectors_test.append(sent_vec)
print(len(sent_vectors_test))
print(len(sent_vectors_test[0]))
```

```
100%|██████████| 30000/30000 [01:17<00:00, 387.94it/s]
```

```
30000
50
```

In [115]:

```
X_test_w2v = sent_vectors_test
```

[4.4.1.2] TFIDF weighted W2v

In [111]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
model.fit(X_train)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))
```

In [113]:

```
# TF-IDF weighted Word2Vec
tfidf_feat = model.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_sentance: # 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
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
            #tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors.append(sent_vec)
    row += 1
    if row%5000 is 0:
        print(row, " completed")
```

```
5000  completed
10000  completed
15000  completed
20000  completed
25000  completed
30000  completed
35000  completed
40000  completed
45000  completed
50000  completed
55000  completed
60000  completed
65000  completed
70000  completed
```

In [126]:

```
X_train_tf_idf_w2v = tfidf_sent_vectors
```

In [128]:

```
len(X_train_tf_idf_w2v)
```

Out[128]:

```
70000
```

In [121]:

```
model.transform(X_test)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(model.get_feature_names(), list(model.idf_)))
```

In [123]:

```
# TF-IDF weighted Word2Vec
tfidf_feat = model.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf
list_of_sentance_test=[]
for sentance in X_test:
    list_of_sentance_test.append(sentance.split())
tfidf_sent_vectors_test = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in list_of_sentance_test: # 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
        if word in w2v_words and word in tfidf_feat:
            vec = w2v_model.wv[word]
            #tf_idf = tf_idf_matrix[row, tfidf_feat.index(word)]
            # to reduce the computation we are
            # dictionary[word] = idf value of word in whole courpus
            # sent.count(word) = tf valeus of word in this review
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors_test.append(sent_vec)
    row += 1
    if row%5000 is 0:
        print(row, " completed")
```

```
5000  completed
10000  completed
15000  completed
20000  completed
25000  completed
30000  completed
```

In [129]:

```
X_test_tf_idf_w2v = tfidf_sent_vectors_test
```

In [131]:

```
len(tfidf_sent_vectors_test[0])
```

Out[131]:

```
50
```

In [132]:

```
def namestr(obj, namespace):
    return [name for name in namespace if namespace[name] is obj]
```

In [141]:

```
vals=[X_test_tf_idf_w2v,
    X_train_tf_idf_w2v,
    X_test_w2v,
    X_train_w2v,
    X_test_tf_idf,
    X_train_tf_idf,
    X_test_ngram,
    X_train_ngram,
    X_test_bow,
    X_train_bow]

names = ["X_test_tf_idf_w2v",
        "X_train_tf_idf_w2v",
        "X_test_w2v",
        "X_train_w2v",
        "X_test_tf_idf",
        "X_train_tf_idf",
        "X_test_ngram",
        "X_train_ngram",
        "X_test_bow",
        "X_train_bow"]
```

In [142]:

```
for i in range(10):
    with open('../../dataset/'+names[i]+'.pickle', 'wb') as handle:
        pickle.dump(vals[i], handle, protocol=pickle.HIGHEST_PROTOCOL)
```

In [149]:

```
!dir C:\Users\Srivas26\appliedai_assignments\dataset
```

```
 Volume in drive C is OS
 Volume Serial Number is 409A-2FCF

 Directory of C:\Users\Srivas26\appliedai_assignments\dataset

02/22/2019  12:16 AM    <DIR>          .
02/22/2019  12:16 AM    <DIR>          ..
05/01/2017  06:55 PM       372,798,464 database.sqlite
02/21/2019  10:02 PM        81,389,388 final_100k_pickle
02/21/2019  10:02 PM       295,457,202 final_pickle
05/01/2017  06:55 PM       300,904,694 Reviews.csv
02/22/2019  12:15 AM        12,022,109 X_test_bow.pickle
02/22/2019  12:15 AM        16,287,845 X_test_ngram.pickle
02/22/2019  12:15 AM        16,287,822 X_test_tf_idf.pickle
02/22/2019  12:15 AM        12,991,954 X_test_tf_idf_w2v.pickle
02/22/2019  12:15 AM        12,991,954 X_test_w2v.pickle
02/22/2019  12:15 AM        27,829,441 X_train_bow.pickle
02/22/2019  12:15 AM        37,896,242 X_train_ngram.pickle
02/22/2019  12:15 AM       195,434,942 X_train_tf_idf.pickle
02/22/2019  12:15 AM        30,314,401 X_train_tf_idf_w2v.pickle
02/22/2019  12:15 AM        30,314,401 X_train_w2v.pickle
              14 File(s)  1,442,920,859 bytes
               2 Dir(s)  98,713,804,800 bytes free
```

In [ ]:

```
```