# 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 Score/Rating. A rating of 4 or 5 can be cosnidered as a positive review. A rating of 1 or 2 can be considered as negative one. A review of rating 3 is considered nuetral and such reviews are ignored from our analysis. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

In [4]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6
qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3aietf%3awg%3aoauth%3a2.0%
b&response_type=code&scope=email%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdocs.test%20https%3a%2
www.googleapis.com%2fauth%2fdrive%20https%3a%2f%2fwww.googleapis.com%2fauth%2fdrive.photos.readonly
ttps%3a%2f%2fwww.googleapis.com%2fauth%2fpeopleapi.readonly

Enter your authorization code:
··········
Mounted at /content/drive
```

# [1]. Reading Data

## [1.1] 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 is 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 is above 3, then the recommendation wil be set to "positive". Otherwise, it will be set to "negative".

In [0]:

```
%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.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
```

In [6]:

```
# using SQLite Table to read data.
con = sqlite3.connect('/content/drive/My Drive/Colab Notebooks/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""", co
n)
# for tsne assignment you can take 5k data points

filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 LIMIT 5000""", con)

# Give reviews with Score>3 a positive rating(1), and reviews with a score<3 a negative rating(0).
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 (5000, 10)
```

Out[6]:

| | 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 |
| **2** | 3 | B000LQOCH0 | ABXLMWJIXXAIN | Natalia Corres "Natalia Corres" | 1 | 1 | 1 | 1219017600 | "Delight" says it all |

In [0]:

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

In [8]:

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

```
(80668, 7)
```

Out[8]:

| | UserId | ProductId | ProfileName | Time | Score | Text | COUNT(*) |
|---|---|---|---|---|---|---|---|
| **0** | #oc-R115TNMSPFT9I7 | B007Y59HVM | Breyton | 1331510400 | 2 | Overall its just OK when considering the price... | 2 |
| **1** | #oc-R11D9D7SHXIJB9 | B005HG9ET0 | Louis E. Emory "hoppy" | 1342396800 | 5 | My wife has recurring extreme muscle spasms, u... | 3 |
| **2** | #oc-R11DNU2NBKQ23Z | B007Y59HVM | Kim Cieszykowski | 1348531200 | 1 | This coffee is horrible and unfortunately not ... | 2 |
| **3** | #oc-R11O5J5ZVQE25C | B005HG9ET0 | Penguin Chick | 1346889600 | 5 | This will be the bottle that you grab from the... | 3 |
| **4** | #oc-R12KPBODL2B5ZD | B007OSBE1U | Christopher P. Presta | 1348617600 | 1 | I didnt like this coffee. Instead of telling y... | 2 |

In [9]:

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

Out[9]:

| | UserId | ProductId | ProfileName | Time | Score | Text | COUNT(*) |
|---|---|---|---|---|---|---|---|
| **80638** | AZY10LLTJ71NX | B006P7E5ZI | undertheshrine "undertheshrine" | 1334707200 | 5 | I was recommended to try green tea extract to ... | 5 |

In [10]:

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

Out[10]:

```
393063
```

# [2] Exploratory Data Analysis

## [2.1] 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 [11]:

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

Out[11]:

| | Id | ProductId | UserId | ProfileName | HelpfulnessNumerator | HelpfulnessDenominator | Score | Time | Summ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 78445 | B000HDL1RQ | AR5J8UI46CURR | Geetha Krishnan | 2 | 2 | 5 | 1199577600 | LOACK QUADRAT VANI WAFE |
| 1 | 138317 | B000HDOPYC | AR5J8UI46CURR | Geetha Krishnan | 2 | 2 | 5 | 1199577600 | LOACK QUADRAT VANI WAFE |
| 2 | 138277 | B000HDOPYM | AR5J8UI46CURR | Geetha Krishnan | 2 | 2 | 5 | 1199577600 | LOACK QUADRAT VANI WAFE |
| 3 | 73791 | B000HDOPZG | AR5J8UI46CURR | Geetha Krishnan | 2 | 2 | 5 | 1199577600 | LOACK QUADRAT VANI WAFE |
| 4 | 155049 | B000PAQ75C | AR5J8UI46CURR | Geetha Krishnan | 2 | 2 | 5 | 1199577600 | LOACK QUADRAT VANI WAFE |

As it can be seen above that same user has multiple reviews with 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 [0]:

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

In [13]:

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

Out[13]:

```
(4986, 10)
```

In [14]:

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

Out[14]:

```
99.72
```

**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 [15]:

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

display.head()
```

Out[15]:

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

In [0]:

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

In [17]:

```
#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()
```

```
(4986, 10)
```

Out[17]:

```
1    4178
0     808
Name: Score, dtype: int64
```

# [3] Preprocessing

## [3.1]. Preprocessing Review Text

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 [18]:

```
# 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)
```

```
Why is this $[...] when the same product is available for $[...] here?<br
/>http://www.amazon.com/VICTOR-FLY-MAGNET-BAIT-REFILL/dp/B00004RBDY<br /><br />The Victor M380 and
M502 traps are unreal, of course -- total fly genocide. Pretty stinky, but only right nearby.
==================================================
I recently tried this flavor/brand and was surprised at how delicious these chips are.  The best t
hing was that there were a lot of "brown" chips in the bsg (my favorite), so I bought some more th
rough amazon and shared with family and friends.  I am a little disappointed that there are not, s
o far, very many brown chips in these bags, but the flavor is still very good.  I like them better
than the yogurt and green onion flavor because they do not seem to be as salty, and the onion flav
or is better.  If you haven't eaten Kettle chips before, I recommend that you try a bag before buy
ing bulk.  They are thicker and crunchier than Lays but just as fresh out of the bag.
==================================================
Wow.  So far, two two-star reviews.  One obviously had no idea what they were ordering; the other
wants crispy cookies.  Hey, I'm sorry; but these reviews do nobody any good beyond reminding us to
look  before ordering.<br /><br />These are chocolate-oatmeal cookies.  If you don't like that com
bination, don't order this type of cookie.  I find the combo quite nice, really.  The oatmeal sort
of "calms" the rich chocolate flavor and gives the cookie sort of a coconut-type consistency.  Now
let's also remember that tastes differ; so, I've given my opinion.<br /><br />Then, these are soft
, chewy cookies -- as advertised.  They are not "crispy" cookies, or the blurb would say "crispy,"
rather than "chewy."  I happen to like raw cookie dough; however, I don't see where these taste li
ke raw cookie dough.  Both are soft, however, so is this the confusion?  And, yes, they stick toge
ther.  Soft cookies tend to do that.  They aren't individually wrapped, which would add to the
cost.  Oh yeah, chocolate chip cookies tend to be somewhat sweet.<br /><br />So, if you want
something hard and crisp, I suggest Nabiso's Ginger Snaps.  If you want a cookie that's soft, chew
y and tastes like a combination of chocolate and oatmeal, give these a try.  I'm here to place my
second order.
==================================================
love to order my coffee on amazon.  easy and shows up quickly.<br />This k cup is great coffee.  d
caf is very good as well
==================================================
```

In [19]:

```
# 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)
```

```
Why is this $[...] when the same product is available for $[...] here?<br /> /><br />The Victor M3
80 and M502 traps are unreal, of course -- total fly genocide. Pretty stinky, but only right nearb
y.
```

In [20]:

```
# https://stackoverflow.com/questions/16206380/python-beautifulsoup-how-to-remove-all-tags-from-an
-element
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)
```

```
Why is this $[...] when the same product is available for $[...] here? />The Victor M380 and M502
traps are unreal, of course -- total fly genocide. Pretty stinky, but only right nearby.
==================================================
I recently tried this flavor/brand and was surprised at how delicious these chips are.  The best t
hing was that there were a lot of "brown" chips in the bsg (my favorite), so I bought some more th
rough amazon and shared with family and friends.  I am a little disappointed that there are not, s
o far, very many brown chips in these bags, but the flavor is still very good.  I like them better
than the yogurt and green onion flavor because they do not seem to be as salty, and the onion flav
or is better.  If you haven't eaten Kettle chips before, I recommend that you try a bag before buy
ing bulk.  They are thicker and crunchier than Lays but just as fresh out of the bag.
==================================================
Wow.  So far, two two-star reviews.  One obviously had no idea what they were ordering; the other
wants crispy cookies.  Hey, I'm sorry; but these reviews do nobody any good beyond reminding us to
look  before ordering.These are chocolate-oatmeal cookies.  If you don't like that combination, do
n't order this type of cookie.  I find the combo quite nice, really.  The oatmeal sort of "calms"
the rich chocolate flavor and gives the cookie sort of a coconut-type consistency.  Now let's also
remember that tastes differ; so, I've given my opinion.Then, these are soft, chewy cookies -- as
advertised.  They are not "crispy" cookies, or the blurb would say "crispy," rather than "chewy."
I happen to like raw cookie dough; however, I don't see where these taste like raw cookie dough.
Both are soft, however, so is this the confusion?  And, yes, they stick together.  Soft cookies te
nd to do that.  They aren't individually wrapped, which would add to the cost.  Oh yeah, chocolate
chip cookies tend to be somewhat sweet.So, if you want something hard and crisp, I suggest Nabiso'
s Ginger Snaps.  If you want a cookie that's soft, chewy and tastes like a combination of
chocolate and oatmeal, give these a try.  I'm here to place my second order.
==================================================
love to order my coffee on amazon.  easy and shows up quickly.This k cup is great coffee.  dcaf is
very good as well
```

In [0]:

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

def decontracted(phrase):
    # specific
```

```
    # 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 [22]:

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

```
Wow.  So far, two two-star reviews.  One obviously had no idea what they were ordering; the other
wants crispy cookies.  Hey, I am sorry; but these reviews do nobody any good beyond reminding us t
o look  before ordering.<br /><br />These are chocolate-oatmeal cookies.  If you do not like that
combination, do not order this type of cookie.  I find the combo quite nice, really.  The oatmeal
sort of "calms" the rich chocolate flavor and gives the cookie sort of a coconut-type consistency.
Now let is also remember that tastes differ; so, I have given my opinion.<br /><br />Then, these a
re soft, chewy cookies -- as advertised.  They are not "crispy" cookies, or the blurb would say "c
rispy," rather than "chewy."  I happen to like raw cookie dough; however, I do not see where these
taste like raw cookie dough.  Both are soft, however, so is this the confusion?  And, yes, they st
ick together.  Soft cookies tend to do that.  They are not individually wrapped, which would add t
o the cost.  Oh yeah, chocolate chip cookies tend to be somewhat sweet.<br /><br />So, if you want
something hard and crisp, I suggest Nabiso is Ginger Snaps.  If you want a cookie that is soft, ch
ewy and tastes like a combination of chocolate and oatmeal, give these a try.  I am here to place
my second order.
==================================================
```

In [23]:

```
#remove words with numbers python: https://stackoverflow.com/a/18082370/4084039
sent_0 = re.sub("\S*\d\S*", "", sent_0).strip()
print(sent_0)
```

```
Why is this $[...] when the same product is available for $[...] here?<br /> /><br />The Victor  a
nd  traps are unreal, of course -- total fly genocide. Pretty stinky, but only right nearby.
```

In [24]:

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

```
Wow So far two two star reviews One obviously had no idea what they were ordering the other wants
crispy cookies Hey I am sorry but these reviews do nobody any good beyond reminding us to look bef
ore ordering br br These are chocolate oatmeal cookies If you do not like that combination do not
order this type of cookie I find the combo quite nice really The oatmeal sort of calms the rich ch
ocolate flavor and gives the cookie sort of a coconut type consistency Now let is also remember th
at tastes differ so I have given my opinion br br Then these are soft chewy cookies as advertised
They are not crispy cookies or the blurb would say crispy rather than chewy I happen to like raw c
ookie dough however I do not see where these taste like raw cookie dough Both are soft however so
is this the confusion And yes they stick together Soft cookies tend to do that They are not indivi
dually wrapped which would add to the cost Oh yeah chocolate chip cookies tend to be somewhat
sweet br br So if you want something hard and crisp I suggest Nabiso is Ginger Snaps If you want a
cookie that is soft chewy and tastes like a combination of chocolate and oatmeal give these a try
I am here to place my second order
```

In [0]:

```
# 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
```

```
# 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', "y
ou'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', 'e
ach', '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', "do
esn'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 [26]:

```
# 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%|██████████| 4986/4986 [00:01<00:00, 3024.17it/s]
```

In [27]:

```
preprocessed_reviews[1500]
```

Out[27]:

```
'wow far two two star reviews one obviously no idea ordering wants crispy cookies hey sorry review
s nobody good beyond reminding us look ordering chocolate oatmeal cookies not like combination not
order type cookie find combo quite nice really oatmeal sort calms rich chocolate flavor gives
cookie sort coconut type consistency let also remember tastes differ given opinion soft chewy cook
ies advertised not crispy cookies blurb would say crispy rather chewy happen like raw cookie dough
however not see taste like raw cookie dough soft however confusion yes stick together soft cookies
tend not individually wrapped would add cost oh yeah chocolate chip cookies tend somewhat sweet wa
nt something hard crisp suggest nabiso ginger snaps want cookie soft chewy tastes like combination
chocolate oatmeal give try place second order'
```

## [3.2] Preprocessing Review Summary

In [0]:

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

# [4] Featurization

## [4.1] BAG OF WORDS

In [29]:

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

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

```
some feature names  ['aa', 'aahhhs', 'aback', 'abandon', 'abates', 'abbott', 'abby', 'abdominal',
'abiding', 'ability']
==================================================
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (4986, 12997)
the number of unique words  12997
```

## [4.2] Bi-Grams and n-Grams.

In [30]:

```
#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/sklearn.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, max_features=5000)
final_bigram_counts = count_vect.fit_transform(preprocessed_reviews)
print("the type of count vectorizer ",type(final_bigram_counts))
print("the shape of out text BOW vectorizer ",final_bigram_counts.get_shape())
print("the number of unique words including both unigrams and bigrams ", final_bigram_counts.get_s
hape()[1])
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer  (4986, 3144)
the number of unique words including both unigrams and bigrams  3144
```

## [4.3] TF-IDF

In [31]:

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

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

```
some sample features(unique words in the corpus) ['ability', 'able', 'able find', 'able get',
'absolute', 'absolutely', 'absolutely delicious', 'absolutely love', 'absolutely no', 'according']
==================================================
```

```
the type of count vectorizer  <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer  (4986, 3144)
the number of unique words including both unigrams and bigrams  3144
```

## [4.4] Word2Vec

In [0]:

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

In [33]:

```
# 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=Tr
ue)
        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 ")
```

```
[('excellent', 0.9966709613800049), ('calorie', 0.99604731798172), ('wonderful',
0.9953616857528687), ('awesome', 0.9951024055480957), ('matter', 0.9950856566429138),
('everything', 0.9950630068778992), ('everyone', 0.9950519800186157), ('licorice',
0.9949842095375061), ('especially', 0.9949439167976379), ('similar', 0.9949134588241577)]
==================================================
[('realize', 0.9994642734527588), ('normal', 0.99945467710495), ('types', 0.9994537830352783),
('audio', 0.9994482398033142), ('fridge', 0.9994168877601624), ('already', 0.9994139075279236), ('
usual', 0.9994124174118042), ('uses', 0.9993946552276611), ('experience', 0.9993859529495239), ('f
ell', 0.9993650913238525)]
```

In [34]:

```
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  3817
sample words  ['product', 'available', 'course', 'total', 'pretty', 'stinky', 'right', 'nearby', '
```

```
used', 'ca', 'not', 'beat', 'great', 'received', 'shipment', 'could', 'hardly', 'wait', 'try', 'lo
ve', 'call', 'instead', 'removed', 'easily', 'daughter', 'designed', 'printed', 'use', 'car', 'win
dows', 'beautifully', 'shop', 'program', 'going', 'lot', 'fun', 'everywhere', 'like', 'tv',
'computer', 'really', 'good', 'idea', 'final', 'outstanding', 'window', 'everybody', 'asks',
'bought', 'made']
```

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

#### [4.4.1.1] Avg W2v

In [35]:

```
# 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 300 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%|██████████| 4986/4986 [00:03<00:00, 1290.04it/s]
```

```
4986
50
```

#### [4.4.1.2] TFIDF weighted W2v

In [0]:

```
# S = ["abc def pqr", "def def def abc", "pqr pqr def"]
model = TfidfVectorizer()
tf_idf_matrix = model.fit_transform(preprocessed_reviews)
# 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 [37]:

```
# 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 tqdm(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
```

```
        sent_vec /= weight_sum
    tfidf_sent_vectors.append(sent_vec)
    row += 1
```

```
100%|██████████| 4986/4986 [00:24<00:00, 205.23it/s]
```

# [5] Assignment 11: Truncated SVD

1. **Apply Truncated-SVD on only this feature set:**

   - SET 2:Review text, preprocessed one converted into vectors using (TFIDF)

   - **Procedure:**
     - Take top 2000 or 3000 features from tf-idf vectorizers using idf_ score.
     - You need to calculate the co-occurrence matrix with the selected features (Note: X.X^T doesn't give the co-occurrence matrix, it returns the covariance matrix, check these bolgs blog-1, blog-2 for more information)
     - You should choose the n_components in truncated svd, with maximum explained variance. Please search on how to choose that and implement them. (hint: plot of cumulative explained variance ratio)
     - After you are done with the truncated svd, you can apply K-Means clustering and choose the best number of clusters based on elbow method.
     - Print out wordclouds for each cluster, similar to that in previous assignment.
     - You need to write a function that takes a word and returns the most similar words using cosine similarity between the vectors(vector: a row in the matrix after truncatedSVD)

## Truncated-SVD

In [38]:

```
final['CleanedText']=preprocessed_reviews

conn = sqlite3.connect('final.sqlite')
c=conn.cursor()
conn.text_factory = str
final.to_sql('Reviews', conn, schema=None, if_exists='replace', index=True, index_label=None, chunk
size=None, dtype=None)
con = sqlite3.connect("final.sqlite")
cleaned_data = pd.read_sql_query("select * from Reviews", con)
cleaned_data.shape
```

Out[38]:

```
(4986, 12)
```

In [39]:

```
from sklearn.model_selection import cross_validate
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
import seaborn as sb
from sklearn.metrics import classification_report
from sklearn.feature_extraction.text import TfidfVectorizer
import numpy as np

positive_points = cleaned_data[cleaned_data['Score'] == 1]
negative_points = cleaned_data[cleaned_data['Score'] == 0]
total_points = pd.concat([positive_points, negative_points])
print(total_points.columns)
print(total_points.shape)
```

```
Index(['index', 'Id', 'ProductId', 'UserId', 'ProfileName',
       'HelpfulnessNumerator', 'HelpfulnessDenominator', 'Score', 'Time',
```

```
       'Summary', 'Text', 'CleanedText'],
      dtype='object')
(4986, 12)
```

In [40]:

```
total_points['Time'] = pd.to_datetime(total_points['Time'], origin='unix')
total_points = total_points.sort_values('Time')
sample_points = total_points['CleanedText']
labels = total_points['Score']

print(total_points.shape)
print(sample_points.shape)
print(labels.shape)
```

```
(4986, 12)
(4986,)
(4986,)
```

In [0]:

```
from sklearn.model_selection import cross_validate
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
import seaborn as sb
from sklearn.metrics import classification_report
from sklearn.feature_extraction.text import TfidfVectorizer
import numpy as np

from sklearn.metrics import classification_report
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.preprocessing import StandardScaler
import numpy as np

X=sample_points

y=labels
X_train=X
```

In [0]:

```
vec=TfidfVectorizer(max_features=2002)
vec.fit(X_train) # fit has to happen only on train data
scalar = StandardScaler(with_mean=False)
X_train_tf = vec.transform(X_train)
X_train_tf=scalar.fit_transform(X_train_tf)
```

In [0]:

```
```

## [5.1] Taking top features from TFIDF, SET 2

In [0]:

```
# Please write all the code with proper documentation
```

In [66]:

```
feat=vec.get_feature_names()
print(feat)
top_2000=feat
```

['able', 'absolute', 'absolutely', 'according', 'acid', 'acidic', 'across', 'active', 'actual', 'a
ctually', 'add', 'added', 'addicted', 'addicting', 'addictive', 'adding', 'addition',
'additional', 'additives', 'adds', 'admit', 'adult', 'adults', 'advertised', 'afford',
'affordable', 'afraid', 'afternoon', 'aftertaste', 'afterwards', 'agave', 'age', 'ago', 'agree', '
ahead', 'ahoy', 'aid', 'air', 'alcohol', 'alive', 'allergic', 'allergies', 'allergy', 'allow', 'al
lowed', 'allows', 'almond', 'almonds', 'almost', 'alone', 'along', 'alot', 'already', 'also', 'alt
ernative', 'although', 'always', 'amazed', 'amazing', 'amazon', 'american', 'among', 'amount', 'am
ounts', 'animal', 'animals', 'another', 'anti', 'anymore', 'anyone', 'anything', 'anytime',
'anyway', 'anywhere', 'apart', 'apparently', 'appears', 'apple', 'apples', 'appreciate', 'area', '
aroma', 'around', 'arrive', 'arrived', 'artificial', 'asian', 'ask', 'asked', 'ate', 'audio', 'aut
hentic', 'auto', 'available', 'average', 'avoid', 'aware', 'away', 'awesome', 'awful', 'babies', '
baby', 'back', 'backyard', 'bacon', 'bad', 'bag', 'bags', 'bake', 'baked', 'baking', 'balance', 'b
alanced', 'ball', 'banana', 'bar', 'barbecue', 'barbeque', 'barely', 'bargain', 'bars', 'base', 'b
ased', 'basic', 'basically', 'basil', 'basis', 'basket', 'batch', 'batches', 'batter', 'bbq', 'bea
n', 'beans', 'bears', 'beat', 'beautiful', 'became', 'become', 'beef', 'beer', 'began',
'beginning', 'behind', 'belgian', 'believe', 'belly', 'benefit', 'benefits', 'berry', 'besides', '
best', 'better', 'betty', 'beverage', 'beverages', 'beware', 'beyond', 'big', 'bigger', 'bill', 'b
ills', 'bird', 'birthday', 'biscotti', 'biscuit', 'biscuits', 'bisquick', 'bit', 'bite', 'bites',
'bits', 'bitter', 'bitterness', 'black', 'bland', 'blend', 'blends', 'blood', 'blue',
'blueberries', 'blueberry', 'bob', 'bodied', 'body', 'bold', 'bones', 'bonus', 'boost', 'bother',
'bottle', 'bottles', 'bottom', 'bought', 'bowl', 'box', 'boxes', 'boy', 'boyfriend', 'boys', 'bpa'
, 'brand', 'brands', 'bread', 'break', 'breakfast', 'breath', 'breed', 'brew', 'brewed', 'brewer',
'brewing', 'brews', 'bring', 'brings', 'broken', 'brother', 'brought', 'brown', 'brownie',
'brownies', 'bubble', 'bucks', 'buds', 'buffalo', 'bulk', 'bunch', 'burn', 'burnt', 'business', 'b
utter', 'buttermilk', 'buy', 'buyer', 'buying', 'ca', 'cable', 'cacao', 'cafe', 'caffeine', 'cajun
', 'cake', 'cakes', 'cakesters', 'cal', 'calcium', 'california', 'call', 'called', 'calls',
'calorie', 'calories', 'came', 'candies', 'candy', 'cane', 'canned', 'cannot', 'cans', 'cant', 'ca
r', 'caramel', 'carb', 'carbs', 'cardboard', 'care', 'careful', 'caribou', 'carried', 'carries',
'carrot', 'carrots', 'carry', 'carrying', 'case', 'cases', 'cat', 'cats', 'cause', 'caviar', 'celi
ac', 'center', 'cents', 'cereal', 'certain', 'certainly', 'certified', 'chance', 'change',
'changed', 'charge', 'charges', 'cheap', 'cheaper', 'cheapest', 'check', 'checked', 'cheddar', 'ch
eese', 'chemical', 'chemicals', 'cherry', 'chew', 'chewing', 'chewy', 'chicken', 'child',
'childhood', 'children', 'chili', 'china', 'chinese', 'chip', 'chips', 'chocolate', 'chocolates',
'choice', 'choices', 'cholesterol', 'choose', 'chopped', 'chose', 'chowder', 'christmas',
'chunks', 'cilantro', 'cinnamon', 'claim', 'claims', 'clam', 'clams', 'classic', 'clean', 'clear',
'clearly', 'close', 'closer', 'co', 'coat', 'coating', 'coats', 'cocker', 'coco', 'cocoa',
'coconut', 'coffee', 'coffees', 'cold', 'collection', 'color', 'colored', 'colors', 'com',
'combination', 'come', 'comes', 'coming', 'commented', 'commercial', 'common', 'community',
'companies', 'company', 'comparable', 'compare', 'compared', 'comparison', 'complain',
'complained', 'complaint', 'complaints', 'complete', 'completely', 'computer', 'concentrate',
'concerned', 'condenser', 'condition', 'consider', 'considered', 'considering', 'consistency', 'co
nsistent', 'consistently', 'consume', 'consumed', 'consuming', 'consumption', 'contact',
'contacted', 'contain', 'contained', 'container', 'containers', 'containing', 'contains',
'content', 'continue', 'control', 'convenience', 'convenient', 'cook', 'cooked', 'cookie',
'cookies', 'cooking', 'cooks', 'cool', 'corn', 'corner', 'cost', 'costco', 'costs', 'could', 'coun
t', 'counter', 'country', 'couple', 'coupon', 'course', 'cover', 'covered', 'cracker', 'crackers',
'crap', 'craving', 'cravings', 'crazy', 'cream', 'creamer', 'creamy', 'creme', 'crisp', 'crisps',
'crispy', 'crocker', 'crumbly', 'crumbs', 'crunch', 'crunchy', 'crust', 'cubes', 'cup', 'cups', 'c
urrent', 'currently', 'curry', 'customer', 'customers', 'cut', 'cute', 'cutting', 'dad', 'daily',
'dairy', 'damage', 'damaged', 'dark', 'date', 'dates', 'daughter', 'day', 'days', 'de', 'deal', 'd
ecaf', 'decaffeinated', 'decent', 'decide', 'decided', 'deep', 'definitely', 'delicate',
'delicious', 'delight', 'delighted', 'delightful', 'delivered', 'delivery', 'dense', 'dented', 'de
pending', 'describe', 'described', 'description', 'design', 'designed', 'desk', 'despite',
'dessert', 'device', 'dha', 'diagnosed', 'didnt', 'diet', 'difference', 'different', 'difficult',
'digest', 'digestive', 'dijon', 'dinner', 'dinners', 'dip', 'dipping', 'directions', 'directly', '
dirty', 'disappointed', 'disappointing', 'disappointment', 'discount', 'discovered', 'disease', 'd
isgusting', 'dish', 'dishes', 'distinct', 'doctor', 'dog', 'dogs', 'dollar', 'dollars', 'done', 'd
ont', 'donut', 'door', 'dop', 'double', 'doubt', 'dough', 'dr', 'drank', 'dressing', 'dried', 'dri
nk', 'drinker', 'drinking', 'drinks', 'drip', 'drive', 'driver', 'drop', 'drops', 'dry', 'due', 'd
umplings', 'dust', 'earlier', 'early', 'earth', 'easier', 'easily', 'easy', 'eat', 'eaten', 'eater
', 'eaters', 'eating', 'eats', 'economical', 'edible', 'effect', 'effective', 'effects', 'effort',
'egg', 'eggs', 'either', 'electrolytes', 'else', 'elsewhere', 'em', 'email', 'empty', 'end',
'ended', 'energy', 'england', 'english', 'enjoy', 'enjoyable', 'enjoyed', 'enjoying', 'enjoys', 'e
nough', 'entire', 'equally', 'escapes', 'especially', 'espresso', 'etc', 'european', 'even',
'evening', 'ever', 'every', 'everyday', 'everyone', 'everything', 'everywhere', 'evo', 'exact', 'e
xactly', 'example', 'excellent', 'except', 'excited', 'expect', 'expectations', 'expected',
'expecting', 'expensive', 'experience', 'expiration', 'extra', 'extract', 'extremely', 'eyes',
'fabulous', 'face', 'fact', 'fair', 'fairly', 'fake', 'fall', 'family', 'fan', 'fans',
'fantastic', 'far', 'farms', 'fast', 'faster', 'fat', 'father', 'fats', 'favor', 'favorite', 'favo
rites', 'fed', 'feed', 'feeding', 'feel', 'feeling', 'feels', 'feet', 'felidae', 'fell', 'felt', '
fence', 'fiber', 'figure', 'figured', 'fill', 'filled', 'fillers', 'filling', 'finally', 'find', '
finding', 'fine', 'fingers', 'finish', 'finished', 'firm', 'first', 'fish', 'fit', 'five', 'fix',
'flat', 'flavor', 'flavored', 'flavorful', 'flavoring', 'flavors', 'flavour', 'flax', 'flaxseed',
'flaxseeds', 'floor', 'flour', 'fluffy', 'folks', 'follow', 'followed', 'following', 'food', 'food
s', 'forever', 'forget', 'form', 'formula', 'forward', 'found', 'four', 'free', 'freeze',
'freezer', 'french', 'fresh', 'freshness', 'fridge', 'fried', 'friend', 'friendly', 'friends', 'fr
ont', 'frosting', 'frozen', 'fructose', 'fruit', 'fruits', 'full', 'fully', 'fun', 'funny', 'fur',

'future', 'gain', 'gallon', 'garbage', 'garlic', 'gas', 'gatorade', 'gave', 'general',
'generally', 'german', 'get', 'gets', 'getting', 'gevalia', 'gf', 'gift', 'gifts', 'ginger',
'gingerbread', 'girl', 'give', 'given', 'gives', 'giving', 'glad', 'glass', 'gluten', 'go',
'goes', 'going', 'gold', 'golden', 'gone', 'good', 'goodness', 'goods', 'got', 'gotten',
'gourmet', 'grab', 'graham', 'grain', 'grains', 'grainy', 'grams', 'grandson', 'granted', 'grape',
'grass', 'gravy', 'greasy', 'great', 'greatest', 'green', 'grew', 'grey', 'grind', 'grinding', 'gr
itty', 'grocery', 'gross', 'ground', 'grove', 'grow', 'growing', 'grown', 'guess', 'guests', 'guil
t', 'guilty', 'gum', 'gummy', 'guy', 'guys', 'habit', 'hair', 'half', 'hand', 'handle', 'hands', '
handy', 'happen', 'happened', 'happens', 'happier', 'happy', 'hard', 'harder', 'hardly',
'harmony', 'hate', 'hated', 'hazelnut', 'health', 'healthier', 'healthy', 'heard', 'heart', 'heat'
, 'heaven', 'heavy', 'help', 'helped', 'helpful', 'helps', 'herb', 'herbal', 'herbs', 'hershey', '
hey', 'high', 'higher', 'highly', 'hint', 'hit', 'hodgson', 'hold', 'hole', 'holes', 'home',
'homemade', 'honest', 'honestly', 'honey', 'hooked', 'hope', 'hoping', 'horrible', 'horses', 'hot'
, 'hour', 'hours', 'house', 'household', 'however', 'huge', 'hulls', 'human', 'hungry', 'husband',
'hydrogenated', 'iams', 'ice', 'iced', 'icicle', 'icing', 'idea', 'im', 'imagine', 'immediately',
'important', 'impossible', 'impressed', 'improved', 'include', 'included', 'including',
'increase', 'incredible', 'incredibly', 'individual', 'individually', 'inexpensive',
'information', 'ingredient', 'ingredients', 'inside', 'instant', 'instead', 'instructions',
'intake', 'intense', 'interested', 'interesting', 'internet', 'introduced', 'iron', 'issue',
'issues', 'italian', 'italy', 'item', 'items', 'jack', 'jalapeno', 'jam', 'japanese', 'jar', 'jars
', 'jelly', 'jerk', 'jerky', 'job', 'joe', 'juice', 'junk', 'kavli', 'keep', 'keeping', 'keeps', '
kept', 'kernels', 'ketchup', 'kettle', 'keurig', 'key', 'kibble', 'kick', 'kid', 'kids', 'kind', '
kinda', 'kinds', 'kit', 'kitchen', 'knew', 'knob', 'know', 'knowing', 'known', 'knows', 'kona', 'k
osher', 'lab', 'label', 'labels', 'lack', 'lamb', 'large', 'larger', 'last', 'lasts', 'late', 'lat
er', 'law', 'lay', 'lays', 'lb', 'lbs', 'leaf', 'learned', 'least', 'leave', 'leaves', 'left', 'le
mon', 'lemonade', 'less', 'let', 'level', 'levels', 'licorice', 'lid', 'life', 'light', 'lighter',
'lightly', 'like', 'liked', 'likely', 'likes', 'liking', 'lime', 'limited', 'line', 'liquid', 'lis
t', 'listed', 'literally', 'little', 'live', 'lived', 'liver', 'lives', 'living', 'loaded', 'loaf'
, 'lobster', 'local', 'locally', 'lock', 'lol', 'lollipops', 'long', 'longer', 'look', 'looked', '
looking', 'looks', 'loose', 'lose', 'loss', 'lost', 'lot', 'lots', 'love', 'loved', 'lovely', 'lov
er', 'lovers', 'loves', 'loving', 'low', 'lower', 'luck', 'lunch', 'lunches', 'machine', 'made', '
mahogany', 'mail', 'main', 'major', 'make', 'maker', 'makes', 'making', 'maltodextrin', 'man', 'ma
ngo', 'manner', 'manufactured', 'manufacturer', 'many', 'maple', 'marinade', 'market', 'marzano',
'match', 'mate', 'matter', 'may', 'maybe', 'mccann', 'meal', 'meals', 'mean', 'means', 'meat', 'me
ats', 'medium', 'meet', 'melitta', 'mellow', 'melt', 'melted', 'members', 'mention', 'mentioned',
'merrick', 'mess', 'messy', 'metal', 'mg', 'mic', 'microwave', 'mid', 'middle', 'might', 'mild', '
milk', 'mill', 'mind', 'mine', 'mini', 'mint', 'mints', 'minute', 'minutes', 'misleading', 'miss',
'missing', 'mistake', 'mix', 'mixed', 'mixes', 'mixing', 'modified', 'moist', 'molasses', 'mom', '
money', 'month', 'monthly', 'months', 'morning', 'mornings', 'mostly', 'mother', 'mountain', 'mout
h', 'move', 'moved', 'movie', 'mrs', 'msg', 'much', 'muffin', 'muffins', 'mug', 'multiple', 'must'
, 'mustard', 'nabisco', 'name', 'nasty', 'natural', 'naturally', 'nature', 'near', 'nearby',
'nearly', 'nectar', 'need', 'needed', 'needs', 'negative', 'neither', 'never', 'new', 'newman', 'n
ewmans', 'news', 'next', 'nice', 'nicely', 'night', 'no', 'noise', 'non', 'none', 'noodles',
'nor', 'normal', 'normally', 'nose', 'not', 'note', 'noted', 'nothing', 'notice', 'noticed',
'number', 'numerous', 'nustevia', 'nut', 'nutrients', 'nutrition', 'nutritional', 'nutritious', 'n
uts', 'oatmeal', 'oats', 'obviously', 'occasional', 'odd', 'offer', 'offered', 'offering',
'office', 'often', 'oh', 'oil', 'oils', 'oily', 'ok', 'okay', 'old', 'older', 'oldest', 'olive', '
omaha', 'omega', 'one', 'ones', 'onion', 'onions', 'online', 'open', 'opened', 'opening',
'opinion', 'opposed', 'option', 'options', 'orange', 'order', 'ordered', 'ordering', 'oreo',
'oreos', 'organic', 'organics', 'original', 'originally', 'others', 'otherwise', 'ounce',
'ounces', 'outside', 'outstanding', 'oven', 'overall', 'overly', 'overpowering', 'overwhelming', '
owner', 'oz', 'pack', 'package', 'packaged', 'packages', 'packaging', 'packed', 'packet',
'packets', 'packing', 'packs', 'page', 'paid', 'pain', 'palatable', 'pamela', 'pan', 'pancake', 'p
ancakes', 'panda', 'pantry', 'paper', 'part', 'partially', 'particular', 'particularly', 'party',
'pass', 'past', 'pasta', 'paste', 'pay', 'paying', 'peach', 'peanut', 'peanuts', 'pear', 'peas', '
people', 'pepper', 'peppermint', 'per', 'perfect', 'perfectly', 'perhaps', 'period', 'person',
'personal', 'personally', 'pet', 'pets', 'photo', 'pick', 'picked', 'picky', 'picture',
'pictures', 'pie', 'piece', 'pieces', 'pineapple', 'pizza', 'place', 'placed', 'plain', 'plan', 'p
lant', 'plants', 'plastic', 'play', 'pleasant', 'pleasantly', 'please', 'pleased', 'pleasing',
'pleasure', 'plenty', 'plug', 'plum', 'plus', 'pocket', 'pocky', 'pod', 'pods', 'point', 'points',
'pomegranate', 'poor', 'pop', 'popchips', 'popcorn', 'popped', 'popper', 'popping', 'pops',
'popular', 'pork', 'portable', 'portion', 'positive', 'possible', 'post', 'pot', 'potassium', 'pot
ato', 'potatoes', 'pouch', 'pouches', 'pound', 'pounds', 'pour', 'powder', 'powdered', 'power', 'p
re', 'prefer', 'preference', 'premium', 'prepare', 'prepared', 'present', 'preservatives',
'pressure', 'pretty', 'previous', 'previously', 'price', 'priced', 'prices', 'pricey', 'prime', 'p
robably', 'problem', 'problems', 'process', 'processed', 'produce', 'produced', 'product',
'products', 'program', 'promptly', 'properly', 'protein', 'provide', 'provided', 'provides',
'pudding', 'pull', 'punch', 'puppy', 'purchase', 'purchased', 'purchasing', 'pure', 'puree', 'puri
na', 'purpose', 'purse', 'put', 'putting', 'quality', 'quantity', 'quick', 'quickly', 'quite', 'ra
men', 'ran', 'rancid', 'range', 'rare', 'rarely', 'raspberry', 'rate', 'rather', 'rating',
'ratio', 'raw', 'reach', 'reaction', 'read', 'readily', 'reading', 'ready', 'real', 'realize', 're
alized', 'really', 'reason', 'reasonable', 'reasonably', 'receive', 'received', 'receiving',
'recent', 'recently', 'recipe', 'recipes', 'recommend', 'recommended', 'record', 'recording', 'red
', 'reduced', 'refreshing', 'refund', 'regarding', 'regret', 'regular', 'regularly', 'remember', '
remind', 'reminds', 'remove', 'replace', 'replacement', 'require', 'required', 'requires',
'research', 'response', 'rest', 'restaurant', 'result', 'results', 'retail', 'return', 'returned',
'review', 'reviewer', 'reviewers', 'reviews', 'rice', 'rich', 'ridiculous', 'right', 'riviera', 'r

```
oast', 'roasted', 'roland', 'roll', 'rolls', 'room', 'root', 'rose', 'round', 'royal', 'run', 'run
ning', 'runny', 'runs', 'rye', 'sad', 'safe', 'said', 'salad', 'salads', 'sale', 'salmon',
'salsa', 'salt', 'salted', 'salty', 'sample', 'sampler', 'san', 'sandwich', 'sandwiches',
'sardines', 'satisfied', 'satisfy', 'satisfying', 'sauce', 'sauces', 'sausage', 'sausages',
'save', 'saved', 'saves', 'savor', 'savory', 'saw', 'say', 'saying', 'says', 'scent', 'school', 's
cience', 'scissors', 'scratch', 'sea', 'seal', 'sealed', 'search', 'searched', 'searching',
'season', 'seasoned', 'seasoning', 'seasonings', 'second', 'seconds', 'see', 'seed', 'seeds', 'see
m', 'seemed', 'seems', 'seen', 'select', 'sell', 'seller', 'selling', 'sells', 'send', 'sending',
'senior', 'sense', 'senseo', 'sensitive', 'sent', 'separate', 'serious', 'seriously', 'serve', 'se
rved', 'service', 'serving', 'servings', 'sesame', 'set', 'setting', 'several', 'shake', 'shame',
'shape', 'shaped', 'share', 'shared', 'shelf', 'shiny', 'ship', 'shipment', 'shipped', 'shipping',
'shop', 'shopping', 'short', 'shortening', 'shot', 'shots', 'show', 'shower', 'shown', 'shows', 's
ick', 'side', 'similac', 'similar', 'simple', 'simply', 'since', 'single', 'sinus', 'sip',
'sister', 'sit', 'site', 'sitting', 'six', 'size', 'sized', 'sizes', 'skeptical', 'skin', 'skip',
'sleep', 'slice', 'sliced', 'slices', 'slight', 'slightly', 'slowly', 'small', 'smaller', 'smell',
'smelled', 'smells', 'smooth', 'snack', 'snacking', 'snacks', 'soap', 'soda', 'sodium', 'soft', 's
ofter', 'software', 'sold', 'solid', 'solution', 'somehow', 'someone', 'something', 'sometimes', '
somewhat', 'somewhere', 'son', 'soon', 'sorry', 'sort', 'sound', 'sounds', 'soup', 'soups', 'sour'
, 'source', 'south', 'soy', 'special', 'specific', 'specifically', 'spend', 'spent', 'spice', 'spi
ces', 'spicy', 'splash', 'splenda', 'spoiled', 'spoon', 'sports', 'spot', 'spread', 'spring', 'spr
inkle', 'square', 'stage', 'stale', 'stand', 'standard', 'stands', 'staple', 'star', 'starbucks',
'starch', 'stars', 'start', 'started', 'starting', 'starts', 'stash', 'state', 'stated', 'states',
'stay', 'stays', 'steaks', 'steep', 'step', 'stevia', 'stick', 'sticks', 'sticky', 'still', 'stir'
, 'stock', 'stocking', 'stomach', 'stonewall', 'stools', 'stop', 'stopped', 'store', 'stores', 'st
ory', 'straight', 'strange', 'strawberry', 'strength', 'strong', 'stronger', 'stuck', 'stuff', 'st
yle', 'subscribe', 'subscription', 'substitute', 'subtle', 'success', 'sucralose', 'sugar',
'sugars', 'sugary', 'suggest', 'suggested', 'summer', 'sunflower', 'sunset', 'super', 'superb', 's
uperior', 'supermarket', 'supermarkets', 'supplement', 'supplier', 'supply', 'support',
'supposed', 'sure', 'surprise', 'surprised', 'surprisingly', 'sweet', 'sweetened', 'sweetener', 's
weeteners', 'sweeter', 'sweetner', 'sweetness', 'sweets', 'swiss', 'switch', 'switched',
'switching', 'syrup', 'system', 'table', 'tablespoon', 'take', 'taken', 'takes', 'taking', 'tangy'
, 'tape', 'target', 'tart', 'tassimo', 'taste', 'tasted', 'tasteless', 'tastes', 'tastier',
'tasting', 'tasty', 'tea', 'teas', 'teaspoon', 'teeth', 'tell', 'ten', 'tend', 'tender', 'terms',
'terrible', 'terrier', 'terrific', 'test', 'texture', 'thai', 'thank', 'thanks', 'therefore',
'thick', 'thicker', 'thin', 'thing', 'things', 'think', 'thinking', 'thinks', 'third', 'though', '
thought', 'three', 'threw', 'thrilled', 'throughout', 'throw', 'throwing', 'thrown', 'thus', 'thym
e', 'tic', 'tight', 'time', 'timely', 'times', 'tin', 'tiny', 'tired', 'toast', 'today', 'toddler'
, 'together', 'told', 'tomato', 'tomatoes', 'tongue', 'took', 'tooth', 'top', 'topping',
'tortilla', 'toss', 'total', 'totally', 'touch', 'town', 'traditional', 'training', 'trans', 'tras
h', 'travel', 'treat', 'treats', 'trick', 'tried', 'trip', 'trips', 'trouble', 'true', 'truffle',
'truly', 'trust', 'try', 'trying', 'tummy', 'tuna', 'turkey', 'turn', 'turned', 'turns', 'twice',
'two', 'type', 'types', 'typical', 'typically', 'unable', 'uncle', 'understand', 'unfortunately',
'unhealthy', 'unique', 'unit', 'unless', 'unlike', 'update', 'upon', 'ups', 'upset', 'us', 'usa',
'usb', 'use', 'used', 'uses', 'using', 'usual', 'usually', 'vacation', 'value', 'vanilla',
'varieties', 'variety', 'various', 'vegetable', 'vegetables', 'veggies', 'vendor', 'version',
'vet', 'via', 'vinegar', 'virtually', 'visit', 'vita', 'vitamin', 'vitamins', 'volume', 'vs',
'wafer', 'wafers', 'waffle', 'waffles', 'wait', 'walk', 'walmart', 'walnuts', 'want', 'wanted',
'wanting', 'wants', 'warm', 'wash', 'waste', 'wasted', 'watch', 'watchers', 'watching', 'water', '
watered', 'watery', 'way', 'ways', 'weak', 'weather', 'web', 'website', 'week', 'weekend',
'weeks', 'weight', 'weird', 'well', 'wellness', 'went', 'wet', 'whatever', 'wheat', 'whenever', 'w
hether', 'whim', 'whipped', 'white', 'whole', 'wholesome', 'wife', 'wild', 'windows', 'wine',
'winter', 'wise', 'wish', 'within', 'without', 'wonderful', 'wonderfully', 'wondering', 'wont', 'w
ord', 'work', 'worked', 'working', 'works', 'world', 'worried', 'worry', 'worse', 'worst', 'worth'
, 'would', 'wow', 'wrap', 'wrapped', 'wrappers', 'write', 'wrong', 'xlr', 'yeah', 'year', 'years',
'yeast', 'yellow', 'yes', 'yesterday', 'yet', 'yogurt', 'young', 'yuban', 'yum', 'yummy', 'zero',
'zip']
```

### [5.2] Calulation of Co-occurrence matrix

In [0]:

```python
# Please write all the code with proper documentation
```

In [0]:

```python
window=2
co_occur_matrix = np.zeros([2002,2002])
```

In [0]:

```python
for sentence in X_train:
    words_in_sent=sentance.split()
    for i,word in enumerate(words_in_sent):
```

```
    for i,word in enumerate(words_in_sent):
        if word in top_2000:
            for j in range(max(i-window,0),min(i+window,len(words_in_sent)-1)+1):
                #if preprocessed_reviews[i] in top_n and X_train[i]!=word:
                if words_in_sent[j] in top_2000 and words_in_sent[j]!=word:
                    co_occur_matrix[top_2000.index(words_in_sent[j]),top_2000.index(word)]+=1
```

In [69]:

```
print(co_occur_matrix)
```

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

In [0]:

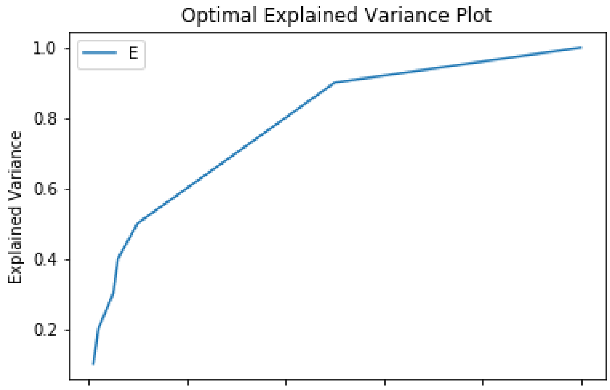
```

```

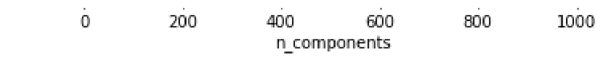
### [5.3] Finding optimal value for number of components (n) to be retained.

In [88]:

```
from sklearn.decomposition import TruncatedSVD

n_components=[10,20,50,60,100,200,300,400,500,1000]
explained_variance=[]
for nn in n_components:
    tsvd=TruncatedSVD(n_components=nn)
    tsvd.fit(co_occur_matrix)
    expvar=tsvd.explained_variance_ratio_.sum()
    explained_variance.append(expvar)
    print('n_components=',nn,'Explained variance=',expvar)
ax = plt.subplot(111)
explained_var=np.arange(0.1,1.1,0.1)
plt.plot(n_components,explained_var)
plt.title('Optimal Explained Variance Plot')
plt.xlabel('n_components')
plt.ylabel('Explained Variance')
ax.legend("Explained Variance")
plt.show()
```

```
n_components= 10 Explained variance= 0.9119075385115525
n_components= 20 Explained variance= 0.9998246926419481
n_components= 50 Explained variance= 1.0000000000000024
n_components= 60 Explained variance= 1.0000000000000033
n_components= 100 Explained variance= 1.0000000000000027
n_components= 200 Explained variance= 1.000000000000002
n_components= 300 Explained variance= 1.000000000000002
n_components= 400 Explained variance= 1.0000000000000024
n_components= 500 Explained variance= 1.0000000000000016
n_components= 1000 Explained variance= 1.0000000000000027
```






### [5.4] Applying k-means clustering

In [0]:

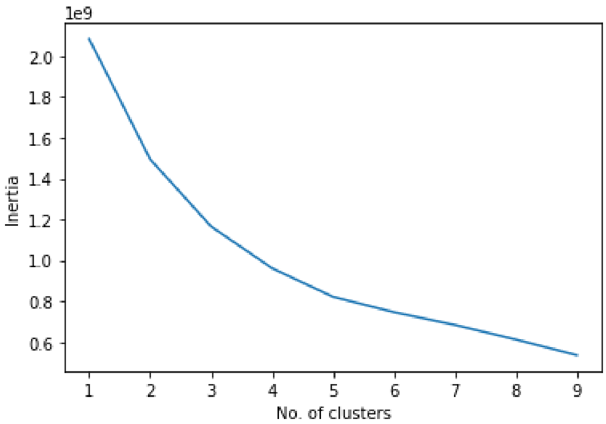
```
tsvd=TruncatedSVD(n_components=150)
co_occur_mat=tsvd.fit_transform(co_occur_matrix)
```

In [0]:

```
# Please write all the code with proper documentation
from sklearn.cluster import KMeans
```

In [109]:

```
inertias = []
K = range(1,10)
dic = {}
for i in K:
    model = KMeans(n_clusters = i)
    model.fit(co_occur_mat)
    dic[i] = model.inertia_
plt.plot(list(dic.keys()), list(dic.values()))
plt.xlabel("No. of clusters")
plt.ylabel("Inertia")
plt.show()
```




In [0]:

```
optimal_k = KMeans(n_clusters = 4)
pp = optimal_k.fit(co_occur_mat)
```

In [0]:

```
cluster1,cluster2,cluster3,cluster4=[],[],[],[]
for i in range(pp.labels_.shape[0]):
    if pp.labels_[i] == 0:
        cluster1.append(preprocessed_reviews[i])
    elif pp.labels_[i] == 1:
        cluster2.append(preprocessed_reviews[i])
    elif pp.labels_[i] == 2:
        cluster3.append(preprocessed_reviews[i])
    else:
        cluster4.append(preprocessed_reviews[i])
```

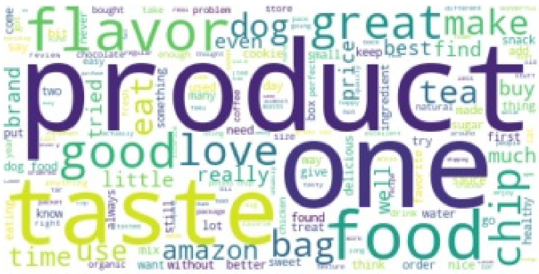
### [5.5] Wordclouds of clusters obtained in the above section

In [0]:

```
# Please write all the code with proper documentation
```

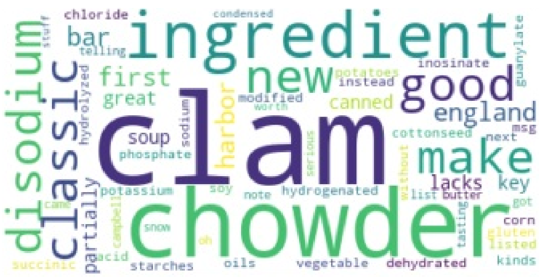
In [113]:

```
data=''
for i in cluster1:
    data+=str(i)
from wordcloud import WordCloud
wordcloud = WordCloud(background_color="white").generate(data)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```




In [114]:

```
data=''
for i in cluster2:
    data+=str(i)
from wordcloud import WordCloud
wordcloud = WordCloud(background_color="white").generate(data)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```




In [115]:

```
data=''
for i in cluster3:
    data+=str(i)
from wordcloud import WordCloud
wordcloud = WordCloud(background_color="white").generate(data)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```


In [116]:

```
data=''
for i in cluster4:
    data+=str(i)
from wordcloud import WordCloud
wordcloud = WordCloud(background_color="white").generate(data)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```


### [5.6] Function that returns most similar words for a given word.

In [0]:

```
# Please write all the code with proper documentation
```

In [118]:

```
from sklearn.metrics.pairwise import cosine_similarity

def similarity(word):
    sim=cosine_similarity(co_occur_matrix)
    word_vector=sim[top_2000.index(word)]
    print("Similar Word to",word)
    index = word_vector.argsort()[::-1][1:11]
    for j in range(len(index)):
        print((j+1),"Word",top_2000[index[j]] ,"is similar to",word,"\n")

similarity(top_2000[2])
```

```
Similar Word to absolutely
1 Word flat is similar to absolutely

2 Word finally is similar to absolutely

3 Word find is similar to absolutely

4 Word finding is similar to absolutely

5 Word fine is similar to absolutely

6 Word fingers is similar to absolutely

7 Word finish is similar to absolutely

8 Word finished is similar to absolutely

9 Word firm is similar to absolutely

10 Word first is similar to absolutely
```

In [0]:

```

```

In [0]:

```

```

# [6] Conclusions

In [0]:

```
# Please write down few lines about what you observed from this assignment.
# Also please do mention the optimal values that you obtained for number of components & number of clusters.
```

**1)In this assignment we took TfIdf Vectorizer and we have reduced the complexity by reducing the dimensions to 2000.**
**2)We have found the "Explained Variance" using SVD and got the optimal number of n_components(reduced dimensions).**
**3)We can say that we computed the TfIdf Vectorizer's(or any) optimal number of dimensions using SVD. The dimensionality can be reduced to that extent.**
**4)After that we have fitted the data by the optimal dimension we got.**
**5)Later we've found the optimal clusters using KMeans and finally wrote a similarity function.**