Assignment - 10.b (Apply Hierarchical clustering on Amazon Food Reviews)

September 11, 2018

1 OBJECTIVE :- Apply Hierarchical clustering on Amazon Food Reviews

In [2]:
```
# Importing libraries
import warnings
warnings.filterwarnings("ignore")

import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
%matplotlib inline
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 nltk.stem.porter import PorterStemmer

import 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
```

2 Loading Data

In [3]:
```
# using the SQLite Table to read data.
con1 = sqlite3.connect('database.sqlite')

# Eliminating neutral reviews i.e. those reviews with Score = 3
filtered_data = pd.read_sql_query(" SELECT * FROM Reviews WHERE Score != 3 ", con1)

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

# Applying polarity function on Score column of filtered_data
filtered_data['Score'] = filtered_data['Score'].map(polarity)

print(filtered_data.shape)
filtered_data.head()
```

```
(525814, 10)
```

Out[3]:

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

	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	\
0	1	1	positive	1303862400	
1	0	0	negative	1346976000	
2	1	1	positive	1219017600	
3	3	3	negative	1307923200	
4	0	0	positive	1350777600	

	Summary	Text
0	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	"Delight" says it all	This is a confection that has been around a fe...
3	Cough Medicine	If you are looking for the secret ingredient i...
4	Great taffy	Great taffy at a great price. There was a wid...

3 Data Cleaning: Deduplication

In [4]:
```
#Sorting data according to ProductId in ascending order
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=Fals
```

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

        #Checking to see how much % of data still remains
        ((final.shape[0]*1.0)/(filtered_data.shape[0]*1.0)*100)
```

```
(364173, 10)
```

Out[4]: 69.25890143662969

In [5]:
```
        # Removing rows where HelpfulnessNumerator is greater than HelpfulnessDenominator
        final = final[final.HelpfulnessNumerator <= final.HelpfulnessDenominator]

        print(final.shape)
        final[30:50]
```

```
(364171, 10)
```

Out[5]:

	Id	ProductId	UserId	ProfileName
138683	150501	0006641040	AJ46FKXOVC7NR	Nicholas A Mesiano
138676	150493	0006641040	AMXOPJKV4PPNJ	E. R. Bird "Ramseelbird"
138682	150500	0006641040	A1IJKK6Q1GTEAY	A Customer
138681	150499	0006641040	A3E7R866M94LOC	L. Barker "simienwolf"
476617	515426	141278509X	AB1A5EGHHVA9M	
22621	24751	2734888454	A1C298ITT645B6	
22620	24750	2734888454	A13ISQV0U9GZIC	
284375	308077	2841233731	A3QD68O22M2XHQ	
157850	171161	7310172001	AFXMWPNS1BLU4	
157849	171160	7310172001	A74C7IARQEM1R	
157833	171144	7310172001	A1V5MY8V9AWUQB	
157832	171143	7310172001	A2SWO6OIW01VPX	
157837	171148	7310172001	A3TFTWTG2CC1GA	
157831	171142	7310172001	A2ZO1AYFVQYG44	
157830	171141	7310172001	AZ4O270J4JBZN	
157829	171140	7310172001	ADXXVGRCGQQUO	
157828	171139	7310172001	A13MS1JQG2ADOJ	
157827	171138	7310172001	A13LAE0YTXA11B	
157848	171159	7310172001	A16GY2RCF410DT	
157834	171145	7310172001	A1L8DNQYY69L2Z	

```
476617                                        CHelmic
22621                                 Hugh G. Pritchard
22620                                        Sandikaye
284375                                         LABRNTH
157850                                      H. Sandler
157849                                         stucker
157833                   Cheryl Sapper "champagne girl"
157832                                             Sam
157837                                      J. Umphress
157831                           Cindy Rellie "Rellie"
157830  Zhinka Chunmee "gamer from way back in the 70's"
157829                             Richard Pearlstein
157828                                      C. Perrone
157827                        Dita Vyslouzilova "dita"
157848                                             LB
157834                                      R. Flores

        HelpfulnessNumerator  HelpfulnessDenominator     Score        Time  \
138683                     2                       2  positive   940809600
138676                    71                      72  positive  1096416000
138682                     2                       2  positive  1009324800
138681                     2                       2  positive  1065830400
476617                     1                       1  positive  1332547200
22621                      0                       0  positive  1195948800
22620                      1                       1  negative  1192060800
284375                     0                       0  positive  1345852800
157850                     0                       0  positive  1229385600
157849                     0                       0  positive  1230076800
157833                     0                       0  positive  1244764800
157832                     0                       0  positive  1252022400
157837                     0                       0  positive  1240272000
157831                     0                       0  positive  1254960000
157830                     0                       0  positive  1264291200
157829                     0                       0  positive  1264377600
157828                     0                       0  positive  1265760000
157827                     0                       0  positive  1269216000
157848                     0                       0  positive  1231718400
157834                     0                       0  positive  1243728000

                                                  Summary  \
138683  This whole series is great way to spend time w...
138676  Read it once. Read it twice. Reading Chicken S...
138682                                It Was a favorite!
138681                                 Can't explain why
476617                                The best drink mix
22621                                  Dog Lover Delites
22620                                      made in china
284375                 Great recipe book for my babycook
```

```
157850                                  Excellent treats
157849                                   Sophie's Treats
157833                        THE BEST healthy dog treat!
157832                    My Alaskan Malamute Loves Them!!
157837                                  Best treat ever!
157831      my 12 year old maltese has always loved these
157830                   Dogs, Cats, Ferrets all love this
157829                                          5 snouts!
157828                                Best dog treat ever
157827                            Great for puppy training
157848                                             Great!
157834                                  Terrific Treats

                                                    Text
138683  I can remember seeing the show when it aired o...
138676  These days, when a person says, "chicken soup"...
138682  This was a favorite book of mine when I was a ...
138681  This book has been a favorite of mine since I ...
476617  This product by Archer Farms is the best drink...
22621   Our dogs just love them.  I saw them in a pet ...
22620   My dogs loves this chicken but its a product f...
284375  This book is easy to read and the ingredients ...
157850  I have been feeding my greyhounds these treats...
157849  This is one product that my welsh terrier can ...
157833  This is the ONLY dog treat that my Lhasa Apso ...
157832  These liver treas are phenomenal. When i recei...
157837  This was the only treat my dog liked during ob...
157831  No waste , even if she is having a  day when s...
157830  I wanted a treat that was accepted and well li...
157829  My Westie loves these things!  She loves anyth...
157828  This is the only dog treat that my terrier wil...
157827  New puppy loves this, only treat he will pay a...
157848  My dog loves these treats!  We started using t...
157834  This is a great treat which all three of my do...
```

OBSERVATION :- Here books with ProductId - 0006641040 and 2841233731 are also there so we have to remove all these rows with these ProductIds from the data

```
In [6]: final = final[final['ProductId'] != '2841233731']
        final = final[final['ProductId'] != '0006641040']
        final.shape

Out[6]: (364136, 10)
```

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

In [7]:
```
#set of stopwords in English
from nltk.corpus import stopwords
stop = set(stopwords.words('english'))
words_to_keep = set(('not'))
stop -= words_to_keep
#initialising the snowball stemmer
sno = nltk.stem.SnowballStemmer('english')

 #function to clean the word of any html-tags
def cleanhtml(sentence):
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', sentence)
    return cleantext

#function to clean the word of any punctuation or special characters
def cleanpunc(sentence):
    cleaned = re.sub(r'[?|!|\'|"|#]',r'',sentence)
    cleaned = re.sub(r'[.|,|)|(|\|/]',r' ',cleaned)
    return  cleaned
```

In [8]:
```
#Code for removing HTML tags , punctuations . Code for removing stopwords . Code for c
# also greater than 2 . Code for stemmimg and also to convert them to lowercase letter
i=0
str1=' '
final_string=[]
all_positive_words=[] # store words from +ve reviews here
all_negative_words=[] # store words from -ve reviews here.
s=''
for sent in final['Text'].values:
    filtered_sentence=[]
    #print(sent);
    sent=cleanhtml(sent) # remove HTMl tags
    for w in sent.split():
        for cleaned_words in cleanpunc(w).split():
            if((cleaned_words.isalpha()) & (len(cleaned_words)>2)):
                if(cleaned_words.lower() not in stop):
                    s=(sno.stem(cleaned_words.lower())).encode('utf8')
                    filtered_sentence.append(s)
                    if (final['Score'].values)[i] == 'positive':
                        all_positive_words.append(s) #list of all words used to descri
                    if(final['Score'].values)[i] == 'negative':
                        all_negative_words.append(s) #list of all words used to descri
                else:
                    continue
            else:
```

```
                    continue

            str1 = b" ".join(filtered_sentence) #final string of cleaned words


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

In [9]:
```
#adding a column of CleanedText which displays the data after pre-processing of the re
final['CleanedText']=final_string
final['CleanedText']=final['CleanedText'].str.decode("utf-8")
#below the processed review can be seen in the CleanedText Column
print('Shape of final',final.shape)
final.head()
```

```
Shape of final (364136, 11)
```

Out[9]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text	CleanedText
476617	515426	141278509X	AB1A5EGHHVA9M	CHelmic	1	1	positive	1332547200	The best drink mix	This product by Archer Farms is the best drink...	product archer farm best drink mix ever mix fl...
22621	24751	2734888454	A1C298ITT645B6	Hugh G. Pritchard	0	0	positive	1195948800	Dog Lover Delites	Our dogs just love them. I saw them in a pet ...	dog love saw pet store tag attach regard made ...
22620	24750	2734888454	A13ISQV0U9GZIC	Sandikaye	1	1	negative	1192060800	made in china	My dogs loves this chicken but its a product f...	dog love chicken product china wont buy anymor...
157850	171161	7310172001	AFXMWPNS1BLU4	H. Sandler	0	0	positive	1229385600	Excellent treats	I have been feeding my greyhounds these treats...	feed greyhound treat year hound littl finicki ...
157849	171160	7310172001	A74C7IARQEM1R	stucker	0	0	positive	1230076800	Sophie's Treats	This is one product that my welsh terrier can ...	one product welsh terrier eat sophi food alerg...

RANDOMLY SAMPLING 5K POINTS OUT OF WHOLE DATASET

In [10]:
```
##Sorting data according to Time in ascending order for Time Based Splitting
time_sorted_data = final.sort_values('Time', axis=0, ascending=True, inplace=False, ki
```

```
        # We will collect different 40K rows without repetition from time_sorted_data datafra
        my_final = time_sorted_data.take(np.random.permutation(len(final))[:5000])

        x = my_final['CleanedText'].values
```

5 (1). Bag of Words (BoW)

```
In [11]: #BoW
         count_vect = CountVectorizer(min_df = 100)
         data = count_vect.fit_transform(x)
         print("the type of count vectorizer :",type(data))
         print("the shape of out text BOW vectorizer : ",data.get_shape())
         print("the number of unique words :", data.get_shape()[1])
```

```
the type of count vectorizer : <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text BOW vectorizer :  (5000, 347)
the number of unique words : 347
```

6 Hierarchical Clustering with 2 clusters

```
In [12]: from sklearn.cluster import AgglomerativeClustering

         model = AgglomerativeClustering(n_clusters=2).fit(data.toarray())

         reviews = my_final['Text'].values
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             else :
                 cluster2.append(reviews[i])


         # Number of reviews in different clusters
         print("No. of reviews in Cluster-1 : ",len(cluster1))
         print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
```

```
No. of reviews in Cluster-1 :  4824

No. of reviews in Cluster-2 :  176
```

READING REVIEWS MANUALLY:

```
In [13]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-3 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick
```

```
In [14]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  In

Review-2 :
 When I read comments that this tea was similar to Earl Gray I decided to try it.  It is nothin

Review-3 :
 Twinings English Afternoon Tea is a superb hot tea, delicious with milk and sugar, with a full
```

7 Hierarchical Clustering with 5 clusters

```
In [15]: model = AgglomerativeClustering(n_clusters=5).fit(data.toarray())

         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []

         for i in range(model.labels_.shape[0]):
```

```
            if model.labels_[i] == 0:
                cluster1.append(reviews[i])
            elif model.labels_[i] == 1:
                cluster2.append(reviews[i])
            elif model.labels_[i] == 2:
                cluster3.append(reviews[i])
            elif model.labels_[i] == 3:
                cluster4.append(reviews[i])
            else :
                cluster5.append(reviews[i])

        # Number of reviews in different clusters
        print("No. of reviews in Cluster-1 : ",len(cluster1))
        print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
        print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
        print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
        print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
```

```
No. of reviews in Cluster-1 :  1537

No. of reviews in Cluster-2 :  176

No. of reviews in Cluster-3 :  302

No. of reviews in Cluster-4 :  2984

No. of reviews in Cluster-5 :  1
```

READING REVIEWS MANUALLY:

```
In [16]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-2 :
 I should have listened to the other one star reviewer. My antler was NOT like the picture. Uni

Review-3 :
 I ordered this product to make white, strawberry-flavored icing for some cupcakes for a weddi
```

In [17]:
```
# Three Reviews of cluster 2
count=1
for i in range(3):
    if i < len(cluster2):
        print('Review-%d : \n %s\n'%(count,cluster2[i]))
        count +=1
```

```
Review-1 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  In

Review-2 :
 When I read comments that this tea was similar to Earl Gray I decided to try it.  It is nothin

Review-3 :
 Twinings English Afternoon Tea is a superb hot tea, delicious with milk and sugar, with a full
```

In [18]:
```
# Three Reviews of cluster 3
count=1
for i in range(3):
    if i < len(cluster3):
        print('Review-%d : \n %s\n'%(count,cluster3[i]))
        count +=1
```

```
Review-1 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder c

Review-2 :
 If you're looking for a delicious, smooth, medium roast coffee blended with a good dose of cho

Review-3 :
 Great coffee flavor in a decaf blend, and I like my coffee. And the convenience of the K cup.
```

In [19]:
```
# Three Reviews of cluster 4
count=1
for i in range(3):
    if i < len(cluster4):
        print('Review-%d : \n %s\n'%(count,cluster4[i]))
        count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th
```

```
Review-3 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast
```

```
In [20]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 Fuzzy Wuzzy's Summary:<br />**** Recommended with warm fuzzies.<br /><br />I just received my
```

8 Hierarchical Clustering with 10 clusters

```
In [21]: model = AgglomerativeClustering(n_clusters=10).fit(data.toarray())
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []
         cluster6 = []
         cluster7 = []
         cluster8 = []
         cluster9 = []
         cluster10 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             elif model.labels_[i] == 1:
                 cluster2.append(reviews[i])
             elif model.labels_[i] == 2:
                 cluster3.append(reviews[i])
             elif model.labels_[i] == 3:
                 cluster4.append(reviews[i])
             elif model.labels_[i] == 4:
                 cluster5.append(reviews[i])
             elif model.labels_[i] == 5:
                 cluster6.append(reviews[i])
             elif model.labels_[i] == 6:
                 cluster7.append(reviews[i])
             elif model.labels_[i] == 7:
```

```
            cluster8.append(reviews[i])
        elif model.labels_[i] == 8:
            cluster9.append(reviews[i])
        else :
            cluster10.append(reviews[i])
```

In [22]:
```
# Number of reviews in different clusters
print("No. of reviews in Cluster-1 : ",len(cluster1))
print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
print("\nNo. of reviews in Cluster-6 : ",len(cluster6))
print("\nNo. of reviews in Cluster-7 : ",len(cluster7))
print("\nNo. of reviews in Cluster-8 : ",len(cluster8))
print("\nNo. of reviews in Cluster-9 : ",len(cluster9))
print("\nNo. of reviews in Cluster-10 : ",len(cluster10))
```

```
No. of reviews in Cluster-1 :  1174

No. of reviews in Cluster-2 :  2984

No. of reviews in Cluster-3 :  308

No. of reviews in Cluster-4 :  48

No. of reviews in Cluster-5 :  21

No. of reviews in Cluster-6 :  18

No. of reviews in Cluster-7 :  34

No. of reviews in Cluster-8 :  254

No. of reviews in Cluster-9 :  158

No. of reviews in Cluster-10 :  1
```

READING REVIEWS MANUALLY:

In [23]:
```
# Three Reviews of cluster 1
count=1
for i in range(3):
    if i < len(cluster1):
        print('Review-%d : \n %s\n'%(count,cluster1[i]))
        count +=1
```

```
Review-1 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick
```

```
Review-2 :
 I ordered this product to make white, strawberry-flavored icing for some cupcakes for a weddi

Review-3 :
 Worst Thing I have ever put in my mouth. Ever. I opened them, immediately noticed a 'funky' sm
```

```
In [24]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-3 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast
```

```
In [25]: # Three Reviews of cluster 3
         count=1
         for i in range(3):
             if i < len(cluster3):
                 print('Review-%d : \n %s\n'%(count,cluster3[i]))
                 count +=1
```

```
Review-1 :
 I should have listened to the other one star reviewer. My antler was NOT like the picture. Uni

Review-2 :
 Given that our dogs don't get rawhide treats often, we thought they might like to try these. I

Review-3 :
 This is the best cat food for my feral cats! The 40 lb size is a great buy, and the cat food i
```

```
In [26]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
```

```
                print('Review-%d : \n %s\n'%(count,cluster4[i]))
                count +=1
```

```
Review-1 :
 If you're looking for a delicious, smooth, medium roast coffee blended with a good dose of ch

Review-2 :
 I have tried many different coffee's over the past severeal years and I am a hard customer to

Review-3 :
 I do not typically drink coffee, just because I never got into the habit. However, if and whe
```

```
In [27]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 What a waste of money. All four cats (three indoor and one feral outdoor cat) won't eat the w

Review-2 :
 After our father died, my sister inherited Duke, our dad's 4 month old orange tabby kitten.  

Review-3 :
 I don't know how ANYONE could rate this so-called "cat" food above a 1 star - it is CRUD.<br
```

```
In [28]: # Three Reviews of cluster 6
         count=1
         for i in range(3):
             if i < len(cluster6):
                 print('Review-%d : \n %s\n'%(count,cluster6[i]))
                 count +=1
```

```
Review-1 :
 I've been a tea fan for decades; I've written about tea and published stories about tea. I ca

Review-2 :
 *****<br />This Bedtime Tea from Yogi Tea is an herbal tea that is relaxing and a natural sle

Review-3 :
 I shared this product with a person in my office who drinks tea daily.  (I'm a big coffee dri
```

```
In [29]: # Three Reviews of cluster 7
         count=1
         for i in range(3):
             if i < len(cluster7):
                 print('Review-%d : \n %s\n'%(count,cluster7[i]))
                 count +=1
```

```
Review-1 :
 First of all, I have no ties with Truvia. In fact, I decided to replace my Truvia recently wi

Review-2 :
 we love sodastream soda maker but not its syrups.  my bf is a soda addict, drink soda like wa

Review-3 :
 I also reviewed the Jones Soda Berry Pomegranate vitamin water and that one fell short becaus
```

```
In [30]: # Three Reviews of cluster 8
         count=1
         for i in range(3):
             if i < len(cluster8):
                 print('Review-%d : \n %s\n'%(count,cluster8[i]))
                 count +=1
```

```
Review-1 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder

Review-2 :
 Great coffee flavor in a decaf blend, and I like my coffee. And the convenience of the K cup.

Review-3 :
 The coffee arrived quickly. It has a good flavor. I'm not sure what the "(Misc.)" means in th
```

```
In [31]: # Three Reviews of cluster 9
         count=1
         for i in range(3):
             if i < len(cluster9):
                 print('Review-%d : \n %s\n'%(count,cluster9[i]))
                 count +=1
```

```
Review-1 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  I

Review-2 :
 When I read comments that this tea was similar to Earl Gray I decided to try it.  It is nothi
```

```
Review-3 :
 Twinings English Afternoon Tea is a superb hot tea, delicious with milk and sugar, with a full
```

```
In [32]: # Three Reviews of cluster 10
         count=1
         for i in range(3):
             if i < len(cluster10):
                 print('Review-%d : \n %s\n'%(count,cluster10[i]))
                 count +=1
```

```
Review-1 :
 Fuzzy Wuzzy's Summary:<br />**** Recommended with warm fuzzies.<br /><br />I just received my
```

9 (2) TFIDF

```
In [33]: tf_idf_vect = TfidfVectorizer(min_df=100)
         data = tf_idf_vect.fit_transform(x)
         print("the type of count vectorizer :",type(data))
         print("the shape of out text TFIDF vectorizer : ",data.get_shape())
         print("the number of unique words :", data.get_shape()[1])
```

```
the type of count vectorizer : <class 'scipy.sparse.csr.csr_matrix'>
the shape of out text TFIDF vectorizer :  (5000, 347)
the number of unique words : 347
```

10 Hierarchical Clustering with 2 clusters

```
In [34]: model = AgglomerativeClustering(n_clusters=2).fit(data.toarray())

         reviews = my_final['Text'].values
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             else :
                 cluster2.append(reviews[i])


         # Number of reviews in different clusters
```

```
        print("No. of reviews in Cluster-1 : ",len(cluster1))
        print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
```

```
No. of reviews in Cluster-1 :  4743

No. of reviews in Cluster-2 :  257
```

READING REVIEWS MANUALLY:

```
In [35]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             print('Review-%d : \n %s\n'%(count,cluster1[i]))
             count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-3 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick
```

```
In [36]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 I have tried literally dozens of teas since being introduced to the Russian custom of preparin

Review-2 :
 No idea why some people are saying this is bad... I guess it doesnt compare to the teas when y

Review-3 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  In
```

11 Hierarchical Clustering with 5 clusters

```
In [37]: model = AgglomerativeClustering(n_clusters=5).fit(data.toarray())
```

```
        # Getting all the reviews in different clusters
        cluster1 = []
        cluster2 = []
        cluster3 = []
        cluster4 = []
        cluster5 = []

        for i in range(model.labels_.shape[0]):
            if model.labels_[i] == 0:
                cluster1.append(reviews[i])
            elif model.labels_[i] == 1:
                cluster2.append(reviews[i])
            elif model.labels_[i] == 2:
                cluster3.append(reviews[i])
            elif model.labels_[i] == 3:
                cluster4.append(reviews[i])
            else :
                cluster5.append(reviews[i])

        # Number of reviews in different clusters
        print("No. of reviews in Cluster-1 : ",len(cluster1))
        print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
        print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
        print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
        print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
```

```
No. of reviews in Cluster-1 :  4146

No. of reviews in Cluster-2 :  212

No. of reviews in Cluster-3 :  289

No. of reviews in Cluster-4 :  257

No. of reviews in Cluster-5 :  96
```

READING REVIEWS MANUALLY:

```
In [38]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like
```

```
Review-2 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-3 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick
```

```
In [39]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 Given that our dogs don't get rawhide treats often, we thought they might like to try these. I

Review-2 :
 I must say these are the best puffed lamb ears we've tried and my dog who normally has a very

Review-3 :
 I have two golden retrievers with hearty appetites.  Feeding them regular dog food makes them
```

```
In [40]: # Three Reviews of cluster 3
         count=1
         for i in range(3):
             if i < len(cluster3):
                 print('Review-%d : \n %s\n'%(count,cluster3[i]))
                 count +=1
```

```
Review-1 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder c

Review-2 :
 I ordered a second case.  Makes great Dark and Stormies or very refreshing and medicinal cold

Review-3 :
 Usually do not get the breakfast blends but it was on sale and decided at price would try it c
```

```
In [41]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
                 print('Review-%d : \n %s\n'%(count,cluster4[i]))
                 count +=1
```

```
Review-1 :
 I have tried literally dozens of teas since being introduced to the Russian custom of preparir

Review-2 :
 No idea why some people are saying this is bad... I guess it doesnt compare to the teas when y

Review-3 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  In
```

```
In [42]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 This is the best cat food for my feral cats! The 40 lb size is a great buy, and the cat food i

Review-2 :
 What a waste of money. All four cats (three indoor and one feral outdoor cat) won't eat the wo

Review-3 :
 These are like candy for our cat, who goes crazy when she hears the bag rustle (or anything th
```

12 Hierarchical Clustering with 10 clusters

```
In [43]: model = AgglomerativeClustering(n_clusters=10).fit(data.toarray())
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []
         cluster6 = []
         cluster7 = []
         cluster8 = []
         cluster9 = []
         cluster10 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             elif model.labels_[i] == 1:
```

```
            cluster2.append(reviews[i])
        elif model.labels_[i] == 2:
            cluster3.append(reviews[i])
        elif model.labels_[i] == 3:
            cluster4.append(reviews[i])
        elif model.labels_[i] == 4:
            cluster5.append(reviews[i])
        elif model.labels_[i] == 5:
            cluster6.append(reviews[i])
        elif model.labels_[i] == 6:
            cluster7.append(reviews[i])
        elif model.labels_[i] == 7:
            cluster8.append(reviews[i])
        elif model.labels_[i] == 8:
            cluster9.append(reviews[i])
        else :
            cluster10.append(reviews[i])
```

In [44]:
```
# Number of reviews in different clusters
print("No. of reviews in Cluster-1 : ",len(cluster1))
print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
print("\nNo. of reviews in Cluster-6 : ",len(cluster6))
print("\nNo. of reviews in Cluster-7 : ",len(cluster7))
print("\nNo. of reviews in Cluster-8 : ",len(cluster8))
print("\nNo. of reviews in Cluster-9 : ",len(cluster9))
print("\nNo. of reviews in Cluster-10 : ",len(cluster10))
```

```
No. of reviews in Cluster-1 :  3745

No. of reviews in Cluster-2 :  212

No. of reviews in Cluster-3 :  147

No. of reviews in Cluster-4 :  257

No. of reviews in Cluster-5 :  71

No. of reviews in Cluster-6 :  289

No. of reviews in Cluster-7 :  64

No. of reviews in Cluster-8 :  62

No. of reviews in Cluster-9 :  57
```

```
No. of reviews in Cluster-10 :  96
```

READING REVIEWS MANUALLY:

```
In [45]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1

Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-3 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast
```

```
In [46]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1

Review-1 :
 Given that our dogs don't get rawhide treats often, we thought they might like to try these. I

Review-2 :
 I must say these are the best puffed lamb ears we've tried and my dog who normally has a very

Review-3 :
 I have two golden retrievers with hearty appetites.  Feeding them regular dog food makes them
```

```
In [47]: # Three Reviews of cluster 3
         count=1
         for i in range(3):
             if i < len(cluster3):
                 print('Review-%d : \n %s\n'%(count,cluster3[i]))
                 count +=1

Review-1 :
 25calories! And chocolate! I am on weight watchers and this is only one point! I love this for
```

```
Review-2 :
 Both the oatmeal and double chocolate chunk taste and look like quality products. They are bot

Review-3 :
 I received a free sample of this bar from Influenster.com.  I really enjoyed this bar.  It's a
```

```
In [48]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
                 print('Review-%d : \n %s\n'%(count,cluster4[i]))
                 count +=1
```

```
Review-1 :
 I have tried literally dozens of teas since being introduced to the Russian custom of preparin

Review-2 :
 No idea why some people are saying this is bad... I guess it doesnt compare to the teas when y

Review-3 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  In
```

```
In [49]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 This oil is not liquid at room temperature. Also, it has a relatively low smoke point. Those i

Review-2 :
 I had been wanting to try rice bran oil for several months, but couldn't find it anywhere loca

Review-3 :
 I'm new to truffle oils and the first few purchases I couldn't get what I was looking for. Thi
```

```
In [50]: # Three Reviews of cluster 6
         count=1
         for i in range(3):
             if i < len(cluster6):
```

```
                print('Review-%d : \n %s\n'%(count,cluster6[i]))
                count +=1
```

```
Review-1 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder
```

```
Review-2 :
 I ordered a second case.  Makes great Dark and Stormies or very refreshing and medicinal cold
```

```
Review-3 :
 Usually do not get the breakfast blends but it was on sale and decided at price would try it
```

```
In [51]: # Three Reviews of cluster 7
         count=1
         for i in range(3):
             if i < len(cluster7):
                 print('Review-%d : \n %s\n'%(count,cluster7[i]))
                 count +=1
```

```
Review-1 :
 Completely opposite of other reviewer.  I love these cookies.  However, if you are looking fo
```

```
Review-2 :
 These cookies are tasty.  At times they're a normal part of my daily diet.  I've eaten a bag
```

```
Review-3 :
 These are delicious! They taste like little shortbread cookies and are coated with a powdery
```

```
In [52]: # Three Reviews of cluster 8
         count=1
         for i in range(3):
             if i < len(cluster8):
                 print('Review-%d : \n %s\n'%(count,cluster8[i]))
                 count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact t
```

```
Review-2 :
 Best hot sauce, I've ever tried!!! DO NOT BE SKEPTIC about this! Read all reviews on the I-ne
```

```
Review-3 :
 I decided to try these noodles with pesto sauce. They were great! They are whole grain buckwh
```

In [53]:
```
# Three Reviews of cluster 9
count=1
for i in range(3):
    if i < len(cluster9):
        print('Review-%d : \n %s\n'%(count,cluster9[i]))
        count +=1
```

```
Review-1 :
 Really pleased with first order back in July. Not so much with the second order. When the seco

Review-2 :
 All of the cherry candys leaked out due to being crushed.  The Candy syrup stuck everything to

Review-3 :
 I've probably been consuming Starburst candies since they were invented and have tried every
```

In [54]:
```
# Three Reviews of cluster 10
count=1
for i in range(3):
    if i < len(cluster10):
        print('Review-%d : \n %s\n'%(count,cluster10[i]))
        count +=1
```

```
Review-1 :
 This is the best cat food for my feral cats! The 40 lb size is a great buy, and the cat food i

Review-2 :
 What a waste of money. All four cats (three indoor and one feral outdoor cat) won't eat the wo

Review-3 :
 These are like candy for our cat, who goes crazy when she hears the bag rustle (or anything th
```

13 Word2Vec

In [55]:
```
# List of sentence in X_train text
sent_x = []
for sent in x :
    sent_x.append(sent.split())


# Train your own Word2Vec model using your own train text corpus
# min_count = 5 considers only words that occured atleast 5 times
w2v_model=Word2Vec(sent_x,min_count=5,size=50, workers=4)
```

```
        w2v_words = list(w2v_model.wv.vocab)
        print("number of words that occured minimum 5 times ",len(w2v_words))
```

```
number of words that occured minimum 5 times  3149
```

14 (3). Avg Word2Vec

```
In [56]: # compute average word2vec for each review for sent_x .
         train_vectors = [];
         for sent in sent_x:
             sent_vec = np.zeros(50)
             cnt_words =0;
             for word in sent: #
                 if word in w2v_words:
                     vec = w2v_model.wv[word]
                     sent_vec += vec
                     cnt_words += 1
             if cnt_words != 0:
                 sent_vec /= cnt_words
             train_vectors.append(sent_vec)

         data = train_vectors
```

15 Hierarchical Clustering with 2 clusters

```
In [58]: model = AgglomerativeClustering(n_clusters=2).fit(data)

         reviews = my_final['Text'].values
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             else :
                 cluster2.append(reviews[i])


         # Number of reviews in different clusters
         print("No. of reviews in Cluster-1 : ",len(cluster1))
         print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
```

```
No. of reviews in Cluster-1 :  2948

No. of reviews in Cluster-2 :  2052
```

READING REVIEWS MANUALLY:

```
In [59]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             print('Review-%d : \n %s\n'%(count,cluster1[i]))
             count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-2 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-3 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast
```

```
In [60]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 A friend introduced this tea several years ago, and I have been searching our local grocery st

Review-3 :
 So grateful for this!! What an amazing mix. It can be used to make some of the best gluten fre
```

16 Hierarchical Clustering with 5 clusters

```
In [62]: model = AgglomerativeClustering(n_clusters=5).fit(data)

         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []
```

```
        for i in range(model.labels_.shape[0]):
            if model.labels_[i] == 0:
                cluster1.append(reviews[i])
            elif model.labels_[i] == 1:
                cluster2.append(reviews[i])
            elif model.labels_[i] == 2:
                cluster3.append(reviews[i])
            elif model.labels_[i] == 3:
                cluster4.append(reviews[i])
            else :
                cluster5.append(reviews[i])

        # Number of reviews in different clusters
        print("No. of reviews in Cluster-1 : ",len(cluster1))
        print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
        print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
        print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
        print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
```

```
No. of reviews in Cluster-1 :  1688

No. of reviews in Cluster-2 :  1712

No. of reviews in Cluster-3 :  818

No. of reviews in Cluster-4 :  364

No. of reviews in Cluster-5 :  418
```

READING REVIEWS MANUALLY:

```
In [63]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 So grateful for this!! What an amazing mix. It can be used to make some of the best gluten fr

Review-2 :
 I should have listened to the other one star reviewer. My antler was NOT like the picture. Un

Review-3 :
 25calories! And chocolate! I am on weight watchers and this is only one point! I love this fo
```

```
In [64]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast

Review-2 :
 This oil is not liquid at room temperature. Also, it has a relatively low smoke point. Those

Review-3 :
 Bought this for my husband who is a major pepper head. I think I ended up eating half of it, a
```

```
In [65]: # Three Reviews of cluster 3
         count=1
         for i in range(3):
             if i < len(cluster3):
                 print('Review-%d : \n %s\n'%(count,cluster3[i]))
                 count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-2 :
 This is a great mix.  Used it in a crock pot, and it's wonderful to come home to a great  meal

Review-3 :
 Very tasty.  I was worried that this would be too hot for some of my family but everyone liked
```

```
In [66]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
                 print('Review-%d : \n %s\n'%(count,cluster4[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 A friend introduced this tea several years ago, and I have been searching our local grocery st
```

```
Review-3 :
 My sister loves Good Earth Original Caffeine Free tea.  We used to be able to get it locally,
```

```
In [67]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-2 :
 I was misled by the name and thought that the cotechino was imported from Italy. Wrong! It was

Review-3 :
 Imagine farmers on earth lose the ability to grow peanuts. And, 10000 years down the road, an
```

17 Hierarchical Clustering with 10 clusters

```
In [69]: model = AgglomerativeClustering(n_clusters=10).fit(data)
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []
         cluster6 = []
         cluster7 = []
         cluster8 = []
         cluster9 = []
         cluster10 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             elif model.labels_[i] == 1:
                 cluster2.append(reviews[i])
             elif model.labels_[i] == 2:
                 cluster3.append(reviews[i])
             elif model.labels_[i] == 3:
                 cluster4.append(reviews[i])
             elif model.labels_[i] == 4:
```

```
            cluster5.append(reviews[i])
        elif model.labels_[i] == 5:
            cluster6.append(reviews[i])
        elif model.labels_[i] == 6:
            cluster7.append(reviews[i])
        elif model.labels_[i] == 7:
            cluster8.append(reviews[i])
        elif model.labels_[i] == 8:
            cluster9.append(reviews[i])
        else :
            cluster10.append(reviews[i])
```

```
In [70]: # Number of reviews in different clusters
         print("No. of reviews in Cluster-1 : ",len(cluster1))
         print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
         print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
         print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
         print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
         print("\nNo. of reviews in Cluster-6 : ",len(cluster6))
         print("\nNo. of reviews in Cluster-7 : ",len(cluster7))
         print("\nNo. of reviews in Cluster-8 : ",len(cluster8))
         print("\nNo. of reviews in Cluster-9 : ",len(cluster9))
         print("\nNo. of reviews in Cluster-10 : ",len(cluster10))
```

```
No. of reviews in Cluster-1 :  601

No. of reviews in Cluster-2 :  1260

No. of reviews in Cluster-3 :  452

No. of reviews in Cluster-4 :  317

No. of reviews in Cluster-5 :  418

No. of reviews in Cluster-6 :  413

No. of reviews in Cluster-7 :  501

No. of reviews in Cluster-8 :  298

No. of reviews in Cluster-9 :  674

No. of reviews in Cluster-10 :  66
```

READING REVIEWS MANUALLY:

```
In [71]: # Three Reviews of cluster 1
         count=1
```

```
        for i in range(3):
            if i < len(cluster1):
                print('Review-%d : \n %s\n'%(count,cluster1[i]))
                count +=1
```

```
Review-1 :
 I got a popcorn maker for Christmas and after doing some research, I found this product. The

Review-2 :
 Tried for a few years to obtain gooseberries but Amazon was the go-to place!!  Now I can bake

Review-3 :
 The product itself was inedible and mostly crumbs. It even smelled bad. The star is there beca
```

In [72]:
```
        # Three Reviews of cluster 2
        count=1
        for i in range(3):
            if i < len(cluster2):
                print('Review-%d : \n %s\n'%(count,cluster2[i]))
                count +=1
```

```
Review-1 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast

Review-2 :
 This oil is not liquid at room temperature. Also, it has a relatively low smoke point. Those

Review-3 :
 Bought this for my husband who is a major pepper head. I think I ended up eating half of it, a
```

In [73]:
```
        # Three Reviews of cluster 3
        count=1
        for i in range(3):
            if i < len(cluster3):
                print('Review-%d : \n %s\n'%(count,cluster3[i]))
                count +=1
```

```
Review-1 :
 This tea has a smooth and flavorful taste.  No bitterness. Cheaper buying it this way,than at

Review-2 :
 I have tried literally dozens of teas since being introduced to the Russian custom of preparin

Review-3 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder
```

```
In [74]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
                 print('Review-%d : \n %s\n'%(count,cluster4[i]))
                 count +=1
```

```
Review-1 :
 Completely opposite of other reviewer.  I love these cookies.  However, if you are looking fo

Review-2 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  I

Review-3 :
 U can taste the lemon in this blend.  I have tried several and the blend had more black peppe
```

```
In [75]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-2 :
 I was misled by the name and thought that the cotechino was imported from Italy. Wrong! It wa

Review-3 :
 Imagine farmers on earth lose the ability to grow peanuts. And, 10000 years down the road, an
```

```
In [76]: # Three Reviews of cluster 6
         count=1
         for i in range(3):
             if i < len(cluster6):
                 print('Review-%d : \n %s\n'%(count,cluster6[i]))
                 count +=1
```

```
Review-1 :
 I should have listened to the other one star reviewer. My antler was NOT like the picture. Un
```

```
Review-2 :
 25calories! And chocolate! I am on weight watchers and this is only one point! I love this fo

Review-3 :
 Excellent cocktails I am preparing and having fun with my girlfriend and friends.<br />Great
```

```
In [77]: # Three Reviews of cluster 7
         count=1
         for i in range(3):
             if i < len(cluster7):
                 print('Review-%d : \n %s\n'%(count,cluster7[i]))
                 count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact t

Review-2 :
 This is a great mix.  Used it in a crock pot, and it's wonderful to come home to a great  mea

Review-3 :
 Very tasty.  I was worried that this would be too hot for some of my family but everyone like
```

```
In [78]: # Three Reviews of cluster 8
         count=1
         for i in range(3):
             if i < len(cluster8):
                 print('Review-%d : \n %s\n'%(count,cluster8[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 A friend introduced this tea several years ago, and I have been searching our local grocery s

Review-3 :
 My sister loves Good Earth Original Caffeine Free tea.  We used to be able to get it locally,
```

```
In [79]: # Three Reviews of cluster 9
         count=1
         for i in range(3):
             if i < len(cluster9):
                 print('Review-%d : \n %s\n'%(count,cluster9[i]))
                 count +=1
```

```
Review-1 :
 So grateful for this!! What an amazing mix. It can be used to make some of the best gluten fr

Review-2 :
 I love licorice these were dry and the flavor was nasty! I think they were many years old no

Review-3 :
 I have to say this is the best canned soup I have ever eaten. The fact that it is organic wit
```

```
In [80]: # Three Reviews of cluster 10
         count=1
         for i in range(3):
             if i < len(cluster10):
                 print('Review-%d : \n %s\n'%(count,cluster10[i]))
                 count +=1
```

```
Review-1 :
 This is not a good deal.  I can go to target or other grocery stores and buy four boxes for a

Review-2 :
 The item arrived on-time and in the advertised condition. Would order this product from Amazo

Review-3 :
 The price was considerably better a while back at $19.44. How can they figure the original pr
```

18 (4). TFIDF-Word2Vec

```
In [81]: # TF-IDF weighted Word2Vec
         tf_idf_vect = TfidfVectorizer()

         # final_tf_idf1 is the sparse matrix with row= sentence, col=word and cell_val = tfid
         final_tf_idf1 = tf_idf_vect.fit_transform(x)

         # tfidf words/col-names
         tfidf_feat = tf_idf_vect.get_feature_names()

         # compute TFIDF Weighted Word2Vec for each review for sent_x .
         tfidf_vectors = [];
         row=0;
         for sent in sent_x:
             sent_vec = np.zeros(50)
             weight_sum =0;
             for word in sent:
                 if word in w2v_words:
```

```
                    vec = w2v_model.wv[word]
                    # obtain the tf_idfidf of a word in a sentence/review
                    tf_idf = final_tf_idf1[row, tfidf_feat.index(word)]
                    sent_vec += (vec * tf_idf)
                    weight_sum += tf_idf
            if weight_sum != 0:
                sent_vec /= weight_sum
            tfidf_vectors.append(sent_vec)
            row += 1

        data = tfidf_vectors
```

19 Hierarchical Clustering with 2 clusters

```
In [82]: model = AgglomerativeClustering(n_clusters=2).fit(data)

         reviews = my_final['Text'].values
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             else :
                 cluster2.append(reviews[i])


         # Number of reviews in different clusters
         print("No. of reviews in Cluster-1 : ",len(cluster1))
         print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
```

```
No. of reviews in Cluster-1 :  3254

No. of reviews in Cluster-2 :  1746
```

READING REVIEWS MANUALLY:

```
In [84]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact tl
```

```
Review-2 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-3 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast
```

```
In [85]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 A friend introduced this tea several years ago, and I have been searching our local grocery st

Review-3 :
 So grateful for this!! What an amazing mix. It can be used to make some of the best gluten fre
```

20 Hierarchical Clustering with 5 clusters

```
In [86]: model = AgglomerativeClustering(n_clusters=5).fit(data)

         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             elif model.labels_[i] == 1:
                 cluster2.append(reviews[i])
             elif model.labels_[i] == 2:
                 cluster3.append(reviews[i])
             elif model.labels_[i] == 3:
                 cluster4.append(reviews[i])
             else :
```

```
                cluster5.append(reviews[i])

        # Number of reviews in different clusters
        print("No. of reviews in Cluster-1 : ",len(cluster1))
        print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
        print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
        print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
        print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
```

```
No. of reviews in Cluster-1 :  1425

No. of reviews in Cluster-2 :  1452

No. of reviews in Cluster-3 :  967

No. of reviews in Cluster-4 :  835

No. of reviews in Cluster-5 :  321
```

READING REVIEWS MANUALLY:

```
In [87]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like

Review-2 :
 So grateful for this!! What an amazing mix. It can be used to make some of the best gluten fr

Review-3 :
 This tea has a smooth and flavorful taste.  No bitterness. Cheaper buying it this way,than at
```

```
In [88]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast
```

```
Review-2 :
 This oil is not liquid at room temperature. Also, it has a relatively low smoke point. Those

Review-3 :
 Bought this for my husband who is a major pepper head. I think I ended up eating half of it,
```

```
In [89]: # Three Reviews of cluster 3
         count=1
         for i in range(3):
             if i < len(cluster3):
                 print('Review-%d : \n %s\n'%(count,cluster3[i]))
                 count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact t

Review-2 :
 I have tried literally dozens of teas since being introduced to the Russian custom of prepari

Review-3 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder
```

```
In [90]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
                 print('Review-%d : \n %s\n'%(count,cluster4[i]))
                 count +=1
```

```
Review-1 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-2 :
 I was misled by the name and thought that the cotechino was imported from Italy. Wrong! It wa

Review-3 :
 I should have listened to the other one star reviewer. My antler was NOT like the picture. Un
```

```
In [91]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
```

```
                print('Review-%d : \n %s\n'%(count,cluster5[i]))
                count +=1
```

```
Review-1 :
 A friend introduced this tea several years ago, and I have been searching our local grocery st

Review-2 :
 My sister loves Good Earth Original Caffeine Free tea.  We used to be able to get it locally,

Review-3 :
 This is not a good deal.  I can go to target or other grocery stores and buy four boxes for at
```

21 Hierarchical Clustering with 10 clusters

```
In [93]: model = AgglomerativeClustering(n_clusters=10).fit(data)
         # Getting all the reviews in different clusters
         cluster1 = []
         cluster2 = []
         cluster3 = []
         cluster4 = []
         cluster5 = []
         cluster6 = []
         cluster7 = []
         cluster8 = []
         cluster9 = []
         cluster10 = []

         for i in range(model.labels_.shape[0]):
             if model.labels_[i] == 0:
                 cluster1.append(reviews[i])
             elif model.labels_[i] == 1:
                 cluster2.append(reviews[i])
             elif model.labels_[i] == 2:
                 cluster3.append(reviews[i])
             elif model.labels_[i] == 3:
                 cluster4.append(reviews[i])
             elif model.labels_[i] == 4:
                 cluster5.append(reviews[i])
             elif model.labels_[i] == 5:
                 cluster6.append(reviews[i])
             elif model.labels_[i] == 6:
                 cluster7.append(reviews[i])
             elif model.labels_[i] == 7:
                 cluster8.append(reviews[i])
             elif model.labels_[i] == 8:
```

```
                cluster9.append(reviews[i])
            else :
                cluster10.append(reviews[i])
```

```
In [94]: # Number of reviews in different clusters
         print("No. of reviews in Cluster-1 : ",len(cluster1))
         print("\nNo. of reviews in Cluster-2 : ",len(cluster2))
         print("\nNo. of reviews in Cluster-3 : ",len(cluster3))
         print("\nNo. of reviews in Cluster-4 : ",len(cluster4))
         print("\nNo. of reviews in Cluster-5 : ",len(cluster5))
         print("\nNo. of reviews in Cluster-6 : ",len(cluster6))
         print("\nNo. of reviews in Cluster-7 : ",len(cluster7))
         print("\nNo. of reviews in Cluster-8 : ",len(cluster8))
         print("\nNo. of reviews in Cluster-9 : ",len(cluster9))
         print("\nNo. of reviews in Cluster-10 : ",len(cluster10))
```

```
No. of reviews in Cluster-1 :  1078

No. of reviews in Cluster-2 :  835

No. of reviews in Cluster-3 :  264

No. of reviews in Cluster-4 :  731

No. of reviews in Cluster-5 :  321

No. of reviews in Cluster-6 :  366

No. of reviews in Cluster-7 :  288

No. of reviews in Cluster-8 :  721

No. of reviews in Cluster-9 :  337

No. of reviews in Cluster-10 :  59
```

READING REVIEWS MANUALLY:

```
In [95]: # Three Reviews of cluster 1
         count=1
         for i in range(3):
             if i < len(cluster1):
                 print('Review-%d : \n %s\n'%(count,cluster1[i]))
                 count +=1
```

```
Review-1 :
 I first tried this product on Princess cruise and since bought it online from Amazon.  I like
```

```
Review-2 :
 So grateful for this!! What an amazing mix. It can be used to make some of the best gluten fr

Review-3 :
 This tea has a smooth and flavorful taste.  No bitterness. Cheaper buying it this way,than at
```

```
In [96]: # Three Reviews of cluster 2
         count=1
         for i in range(3):
             if i < len(cluster2):
                 print('Review-%d : \n %s\n'%(count,cluster2[i]))
                 count +=1
```

```
Review-1 :
 Awful Awful taste...and phosphoric sick color...<br />Awful Awful taste...and phosphoric sick

Review-2 :
 I was misled by the name and thought that the cotechino was imported from Italy. Wrong! It wa

Review-3 :
 I should have listened to the other one star reviewer. My antler was NOT like the picture. Un
```

```
In [97]: # Three Reviews of cluster 3
         count=1
         for i in range(3):
             if i < len(cluster3):
                 print('Review-%d : \n %s\n'%(count,cluster3[i]))
                 count +=1
```

```
Review-1 :
 Some of the finest tea I've had.  It is a pleasure on the palette, as well as to the nose.  I

Review-2 :
 Great tasting tea I have been looking for a good  Sassafras tea forever.Finally. I will be or

Review-3 :
 Twinings English Afternoon Tea is a superb hot tea, delicious with milk and sugar, with a ful
```

```
In [98]: # Three Reviews of cluster 4
         count=1
         for i in range(3):
             if i < len(cluster4):
                 print('Review-%d : \n %s\n'%(count,cluster4[i]))
                 count +=1
```

```
Review-1 :
 This oil is not liquid at room temperature. Also, it has a relatively low smoke point. Those i

Review-2 :
 I ordered this product to make white, strawberry-flavored icing for some cupcakes for a weddi

Review-3 :
 Worst Thing I have ever put in my mouth. Ever. I opened them, immediately noticed a 'funky' sm
```

```
In [99]: # Three Reviews of cluster 5
         count=1
         for i in range(3):
             if i < len(cluster5):
                 print('Review-%d : \n %s\n'%(count,cluster5[i]))
                 count +=1
```

```
Review-1 :
 A friend introduced this tea several years ago, and I have been searching our local grocery st

Review-2 :
 My sister loves Good Earth Original Caffeine Free tea.  We used to be able to get it locally,

Review-3 :
 This is not a good deal.  I can go to target or other grocery stores and buy four boxes for at
```

```
In [100]: # Three Reviews of cluster 6
          count=1
          for i in range(3):
              if i < len(cluster6):
                  print('Review-%d : \n %s\n'%(count,cluster6[i]))
                  count +=1
```

```
Review-1 :
 I have tried literally dozens of teas since being introduced to the Russian custom of preparin

Review-2 :
 I thought this coffee was too weak, and had a slightly sour aftertaste. I do prefer a bolder c

Review-3 :
 Usually do not get the breakfast blends but it was on sale and decided at price would try it c
```

```
In [101]: # Three Reviews of cluster 7
          count=1
```

```
        for i in range(3):
            if i < len(cluster7):
                print('Review-%d : \n %s\n'%(count,cluster7[i]))
                count +=1
```

```
Review-1 :
 Given that our dogs don't get rawhide treats often, we thought they might like to try these. I

Review-2 :
 I must say these are the best puffed lamb ears we've tried and my dog who normally has a very

Review-3 :
 These are not too small nor too big, an excellent snack sized treat.  only problem i have with
```

```
In [102]: # Three Reviews of cluster 8
          count=1
          for i in range(3):
              if i < len(cluster8):
                  print('Review-%d : \n %s\n'%(count,cluster8[i]))
                  count +=1
```

```
Review-1 :
 This is a really nice product for those who want to consume healthful things. Refreshing, tast

Review-2 :
 Bought this for my husband who is a major pepper head. I think I ended up eating half of it, a

Review-3 :
 This is a great mix.  Used it in a crock pot, and it's wonderful to come home to a great  meal
```

```
In [103]: # Three Reviews of cluster 9
          count=1
          for i in range(3):
              if i < len(cluster9):
                  print('Review-%d : \n %s\n'%(count,cluster9[i]))
                  count +=1
```

```
Review-1 :
 I made crab rangoon and used this sauce as a dipping sauce.  It was great.  I love the fact th

Review-2 :
 Completely opposite of other reviewer.  I love these cookies.  However, if you are looking for

Review-3 :
 The label says it's Watermelon and Strawberry but I did not taste either one of these flavors
```

```
In [104]: # Three Reviews of cluster 10
          count=1
          for i in range(3):
              if i < len(cluster10):
                  print('Review-%d : \n %s\n'%(count,cluster10[i]))
                  count +=1
```

```
Review-1 :
 This is the best cat food for my feral cats! The 40 lb size is a great buy, and the cat food i

Review-2 :
 I have two golden retrievers with hearty appetites.  Feeding them regular dog food makes them

Review-3 :
 What a waste of money. All four cats (three indoor and one feral outdoor cat) won't eat the wo
```

22 CONCLUSION :-

23 Procedure Followed :

STEP 1 :- Text Preprocessing

STEP 2 :- Taking all text data and ignoring class variable .

STEP 3:- Training the vectorizer on text_data and later applying same vectorizer on text_data to transform it into vectors

STEP 4:- Implementing Hierarchical Clustering using multiple values of clusters .

STEP 5:- Reading reviews manually for each cluster

Repeat from STEP 3 to STEP 5 for each of these four vectorizers : Bag Of Words(BoW), TFIDF, Avg Word2Vec and TFIDF Word2Vec