Assignment 4 (ML-II)

Clustering Text (Example 4)

Wali Ullah (09745)

In [80]:

```
import warnings
warnings.filterwarnings('ignore')
warnings.simplefilter('ignore')
```

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.patches import Rectangle
import numpy as np
from pprint import pprint as pp
import csv
from pathlib import Path
import seaborn as sns
from itertools import product
import string
from sklearn.cluster import KMeans
from sklearn.cluster import OPTICS
import scipy.cluster.hierarchy as sch
from matplotlib import pyplot
import nltk
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import BorderlineSMOTE
from imblearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import r2_score, classification_report, confusion_matrix, accuracy
from sklearn.metrics import homogeneity_score, silhouette_score
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import MiniBatchKMeans, DBSCAN
import gensim
from gensim import corpora
```

In [2]:

```
# Load Data
def load_data(file_name):
    def readcsv(file_name):
        return pd.read_csv(file_name)
    def readexcel(file_name):
        return pd.read_excel(file_name)
    func_map = {
        "csv": readcsv,
        "xlsx": readexcel,
    }

    # default reader = readcsv
    reader = func_map.get("csv")
    for k,v in func_map.items():
        if file_name.endswith(k):
            reader = v
            break
    return reader(file_name)
```

Data Discription

The dataset consists of 13 year's data which consists of 10 attributes for 568000 reviews. Due to the computational complexity, I am useing a random sample of 10,000 reviews for our analysis.

The dataset contains the following columns :

1.Id->Review for each ID

2.Product Id->Unique identifier for the product

3.User Id->Unique identifier for the user

4.Profile Name->A user who has given the review

5.Helpful Numerator->No. of users who found the review helpful

6.Helpful Denominator->No. of users who found the review helpful or not

7.Score->Five being is the highest rating and 1 being the lowest rating

8.Time->Date and time when the review was given

9.Summary->Summary of the review

10.Text->Review text

In [3]:

```
FILE_NAME = "reviews1.csv"
#FILE_NAME = "banksim_adj.csv"
#LABEL_COL = "fraud"
sample = load_data(FILE_NAME)
display(sample.head())
print(sample.shape)
print(sample.dtypes)
```

```
C:\Users\waliullah\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3418: Dt
ypeWarning: Columns (1,2,3,8,9) have mixed types.Specify dtype option on import or set l
ow_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)
```

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator
0	1.0	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1.0	1.0

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator
1	2.0	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0.0	0.0
2	3.0	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1.0	1.0
3	4.0	B000UA0QIQ	A395BORC6FGVXV	Karl	3.0	3.0
4	5.0	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0.0	0.0

```
(556249, 10)
Id                         float64
ProductId                   object
UserId                      object
ProfileName                 object
HelpfulnessNumerator       float64
HelpfulnessDenominator     float64
Score                      float64
Time                       float64
Summary                     object
Text                        object
dtype: object
```

In [4]:
```
#check the loaded data
print(sample.shape)
```

```
(556249, 10)
```

In [5]:
```
#look of the dataset
sample.head()
```

Out[5]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator
0	1.0	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1.0	1.0

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator
1	2.0	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0.0	0.0
2	3.0	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1.0	1.0
3	4.0	B000UA0QIQ	A395BORC6FGVXV	Karl	3.0	3.0
4	5.0	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0.0	0.0

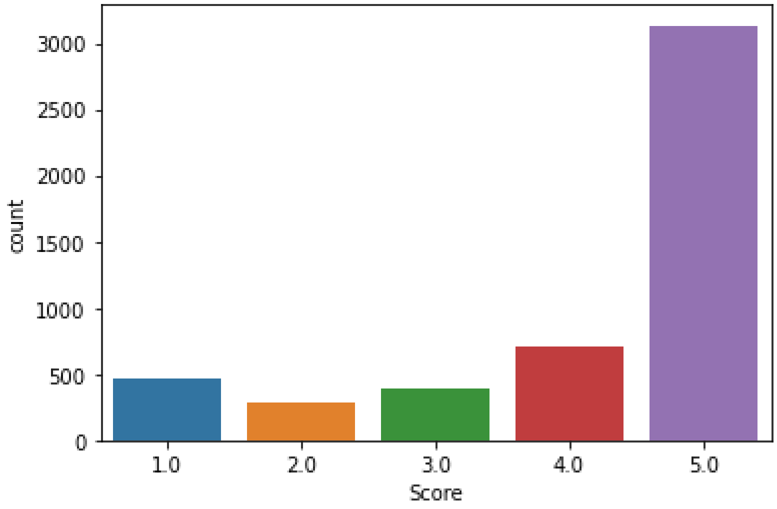
In [6]:

```
# Understand how customer ratings are distributed
import seaborn as sns
sns.countplot(sample.Score)
```

```
C:\Users\waliullah\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
ional argument will be `data`, and passing other arguments without an explicit keyword w
ill result in an error or misinterpretation.
  warnings.warn(
```

Out[6]:

```
<AxesSubplot:xlabel='Score', ylabel='count'>
```



Data Cleaning

In [7]:

```
#converting the Numerical reviws to categorical reviews on codition above 3 are
#positive and below 3 are negative as reviews rating with 3 are not much useful
#for analysis

#function
def partition(x):
    if x < 3:
        return 'negative'
    return 'positive'

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

In [8]:

```
sample.head()
```

Out[8]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator
0	1.0	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1.0	1.0
1	2.0	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0.0	0.0
2	3.0	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1.0	1.0
3	4.0	B000UA0QIQ	A395BORC6FGVXV	Karl	3.0	3.0
4	5.0	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0.0	0.0

In [9]:

```
# no of positive and negative reviews
sample["Score"].value_counts()
#here we can say it is a unbalanced data set
```

Out[9]:
```
positive    555490
negative       759
Name: Score, dtype: int64
```

In [10]:
```
#dropping  the duplicates column if any using drop duplicates from pandas
sorted_data=sample.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='
final.shape
```

Out[10]:
```
(4986, 10)
```

In [11]:
```
# no duplicate columns found
(final['Id'].size*1.0)/(sample['Id'].size*1.0)*100
```

Out[11]:
```
0.8963611619975945
```

In [12]:
```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
# Help..Num is always less than Denom.. as Denom is people who upvote and donwvote
#Before understanding text 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()

# after removing duplicate rows we found, 8346 positive and 1457 negative
```

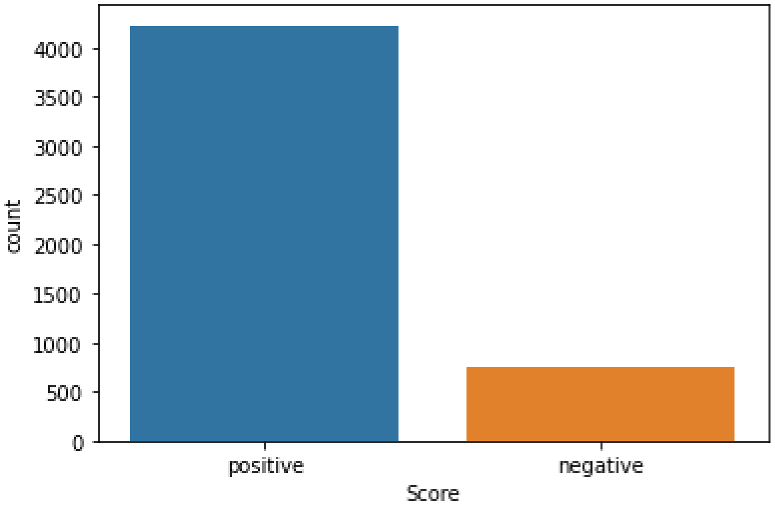
```
(4985, 10)
```

Out[12]:
```
positive    4231
negative     754
Name: Score, dtype: int64
```

In [13]:
```
# After Removing Duplicate rows
import seaborn as sns
sns.countplot(final.Score)
```

```
C:\Users\waliullah\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid posit
ional argument will be `data`, and passing other arguments without an explicit keyword w
ill result in an error or misinterpretation.
  warnings.warn(
```

Out[13]:
```
<AxesSubplot:xlabel='Score', ylabel='count'>
```



Text Processing

To make the text clean by removing HTML tag reviews, stopwords to segregate and adding timestamp

In [14]:

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

```
1
Why is this $[...] when the same product is available for $[...] here?<br />http://www.a
mazon.com/VICTOR-FLY-MAGNET-BAIT-REFILL/dp/B00004RBDY<br /><br />The Victor M380 and M50
2 traps are unreal, of course -- total fly genocide. Pretty stinky, but only right nearb
y.
```

In [15]:

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
sno = nltk.stem.SnowballStemmer('english') #initialising the snowball stemmer which is
stop=set(stopwords.words('english'))



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

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

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\waliullah\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [16]:

```
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 describ
                    if(final['Score'].values)[i] == 'negative':
                        all_negative_words.append(s) #list of all words used to describ
                else:
                    continue
            else:
                continue
    #print(filtered_sentence)
    str1 = b" ".join(filtered_sentence) #final string of cleaned words
    #print("***********************************************************************")

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

In [17]:

```
final['CleanedText']=final_string #adding a column of CleanedText which displays the da
final['CleanedText']=final['CleanedText'].str.decode("utf-8")
```

In [18]:

```
final.shape # cleaned text column added
```

Out[18]:

```
(4985, 11)
```

In [19]:

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

Out[19]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
2774	2775.0	B00002NCJC	A13RRPGE79XFFH	reader48	0.0	(
2773	2774.0	B00002NCJC	A196AJHU9EASJN	Alex Chaffee	0.0	(

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	1(

In [20]:

```
# however, this is not required for clustering, just segregaring positive,negative and
data_pos = final[final["Score"] == "positive"]
data_neg = final[final["Score"] == "negative"]
final = pd.concat([data_pos, data_neg])
score =final["Score"]
final.head()
```

Out[20]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenomina
2774	2775.0	B00002NCJC	A13RRPGE79XFFH	reader48	0.0	
2773	2774.0	B00002NCJC	A196AJHU9EASJN	Alex Chaffee	0.0	
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	1
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	
3202	3203.0	B000084DVR	A3DKGXWUEP1AI2	Glenna E. Bauer "Puppy Mum"	3.0	

In [21]:

```
#Converting the time frame and sorting in increasing order for easyness
final["Time"] = pd.to_datetime(final["Time"], unit = "s")
final= final.sort_values(by = "Time")
final.head()
```

Out[21]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenomi
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	
3782	3783.0	B00016UX0K	AF1PV3DIC0XM7	Robert Ashton	1.0	
1205	1206.0	B005O072PC	A3BD5B8Y8MY25X	J. L. K. "special_k"	13.0	
1275	1276.0	B000WNJ73Q	A394MHK3CSDGUV	kaleinor	2.0	

Clustering

Find Clustering models for both Bag of words, term frequcny/ inverse document frequcny and avg word to vector

K means using bag of words

In [22]:

```
# Generating bag of words features.
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
bow = count_vect.fit_transform(final['CleanedText'].values)
bow.shape
```

```
(4985, 8565)
```

Out[22]:

In [23]:

```
bow
```

Out[23]:

```
<4985x8565 sparse matrix of type '<class 'numpy.int64'>'
	with 150241 stored elements in Compressed Sparse Row format>
```

In [24]:

```
# to understand what kind of words generated as columns by BOW
terms = count_vect.get_feature_names()
```

In [25]:

```
#first 10 columns generated by BOW
terms[1:10]
```

Out[25]:

```
['aback',
 'abandon',
 'abat',
 'abbi',
 'abbott',
 'abdomin',
 'abid',
 'abil',
 'abl']
```

In [26]:

```
#using all processes jobs=-1 and k means++ for starting initilization advantage
from sklearn.cluster import KMeans
model = KMeans(n_clusters = 10,init='k-means++', n_jobs = -1,random_state=99)
model.fit(bow)
```

```
C:\Users\waliullah\Anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:792: FutureWar
ning: 'n_jobs' was deprecated in version 0.23 and will be removed in 1.0 (renaming of 0.
25).
  warnings.warn("'n_jobs' was deprecated in version 0.23 and will be"
```

Out[26]:

```
KMeans(n_clusters=10, n_jobs=-1, random_state=99)
```

In [27]:

```
labels = model.labels_
cluster_center=model.cluster_centers_
```

In [28]:

```
cluster_center
```

Out[28]:

```
array([[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 [29]:

```
from sklearn import metrics
silhouette_score = metrics.silhouette_score(bow, labels, metric='euclidean')
```

In [30]:

```
# which tells us that clusters are far away from each other
silhouette_score
```

Out[30]:

```
0.0738151157266508
```

In [31]:

```
# Giving Labels/assigning a cluster to each point/text
df = final
df['Bow Clus Label'] = model.labels_ # the last column you can see the label numebers
df.head(2)
```

Out[31]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	7
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	10

In [32]:

```python
# How many points belong to each cluster -> using group by in pandas
df.groupby(['Bow Clus Label'])['Text'].count()
```

Out[32]:

```
Bow Clus Label
0     120
1     182
2       1
3      25
4    3438
5      31
6     353
7     673
8      29
9     133
Name: Text, dtype: int64
```

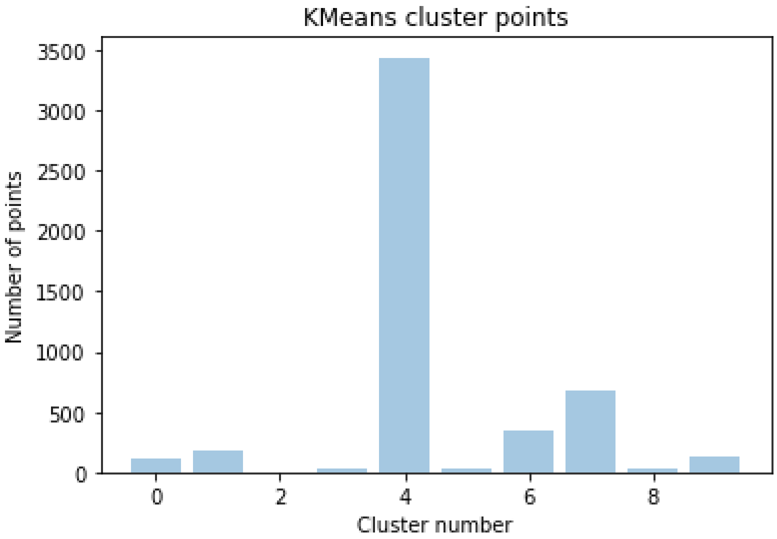
In [33]:

```python
#Refrence credit - to find the top 10 features of cluster centriod
#https://stackoverflow.com/questions/47452119/kmean-clustering-top-terms-in-cluster
print("Top terms per cluster:")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = count_vect.get_feature_names()
for i in range(10):
    print("Cluster %d:" % i, end='')
    for ind in order_centroids[i, :10]:
        print(' %s' % terms[ind], end='')
        print()
```

```
Top terms per cluster:
Cluster 0: coffe
 tast
 flavor
 like
 cup
 decaf
 good
 tri
 drink
 use
Cluster 1: food
 dog
 eat
 love
 like
 newman
 one
 cat
 year
 bag
Cluster 2: egg
 allergi
 calcium
 formula
 babi
 phosphorus
 yolk
 dha
 food
 protein
Cluster 3: one
 oreo
 product
 like
 use
 eat
 make
 would
 get
 cake
Cluster 4: great
 tast
 love
 good
 like
 product
 flavor
 use
 one
 tri
Cluster 5: mix
 pancak
 make
 use
 recip
 product
 like
 tast
 waffl
 good
Cluster 6: chip
 flavor
 bag
 like
 tast
 salt
 good
 great
 potato
 love
Cluster 7: like
 tast
 use
 product
 one
 tri
 good
 flavor
 make
 water
Cluster 8: chip
 bag
 flavor
 potato
 kettl
 like
 tast
 brand
 salt
 good
Cluster 9: tea
 green
 like
 flavor
 tast
 drink
 use
 water
 tri
 bag
```

In [34]:

```python
# visually how points or reviews are distributed across 10 clusters
import matplotlib.pyplot as plt
plt.bar([x for x in range(10)], df.groupby(['Bow Clus Label'])['Text'].count(), alpha =
plt.title('KMeans cluster points')
plt.xlabel("Cluster number")
plt.ylabel("Number of points")
plt.show()
```



In [35]:

```python
# Reading a review which belong to each group.
for i in range(10):
    print("A review of assigned to cluster ", i)
    print("-" * 70)
    print(df.iloc[df.groupby(['Bow Clus Label']).groups[i][0]]['Text'])
    print('\n')
    print("_" * 70)
```

```
A review of assigned to cluster  0
----------------------------------------------------------------------
I LOVE THESE CHIPS, I HAVE THEM ON AUTO ORDER EVERY 2 MONTHS, THEY TASTE GREAT, I CAN NO
T BELIVE THE WHOLE BAG HAS 100 CALORIES, I HAVE A BAG EVERY DAY, IT SURE HAS HELPED MY W
EIGHT LOSS BY HAVEING THEM IN LITTLE BAGS, SO I DO NOT EAT A HUGE AMOUNT
```

```
______________________________________________________________________
A review of assigned to cluster  1
----------------------------------------------------------------------
Best Bit-O-Honey I have ever eaten.<br />I have always bought this product at a local gr
ocery store and it was always hard, I figured it was suppose to be that way. But after e
ating the box I bought here I now know better; the candy was soft and delicious.<br />I 
plan to continue buying Bit-O-Honey from here.


______________________________________________________________________
A review of assigned to cluster  2
----------------------------------------------------------------------
These are the best alternative chips I have ever tried. I have shared them with friends 
and relatives and they all agree. You get enough in the single serve packs to be truly s
atisfied and if you are a weight watchers customer, they are GREAT, only 2 points for ea
ch bag!!!


______________________________________________________________________
A review of assigned to cluster  3
----------------------------------------------------------------------
My dogs have been eating this brand for a few years now and having found it available th
rough Amazon I am able to save money per pound and trips to the store. It arrives prompt
ly and is the same quality as always. Harmony Farms is much easier to digest and healthi
er than most commercial foods. LOVE IT!


______________________________________________________________________
A review of assigned to cluster  4
----------------------------------------------------------------------
I have tried many bread machine mixes and most have been okay but this one rates an "exc
ellent" because not only is it a very good basic loaf, it also works very well with addi
tives, either prepared entirely in the bread machine, or (after the second rise) removed
from the machine, shaped and baked on a stone in the oven.<br />Stretched and flattened 
with a filling of either savory or sweet ingredients, then rolled and set to rise and th
en bake in a loaf pan, it produced excellent herb and onion bread, cinnamon raisin bread
and a brown butter and seed bread.<br />I buy it via the subscribe-and-save plan as I us
e it often.


______________________________________________________________________
A review of assigned to cluster  5
----------------------------------------------------------------------
This  coffee is the smoothest dark roast coffee I have ever tasted,and it was a pleasure
to sip this full bodied coffee before breakfast and after dinner.


______________________________________________________________________
A review of assigned to cluster  6
----------------------------------------------------------------------
I was sorely disappointed in these cookies.. They are pretty tasteless and damn hard to 
o.. I wouldn't buy them again and I've certainly tasted much better low-cal cookies.. I 
definitely recommend saving your money, folks..


______________________________________________________________________
A review of assigned to cluster  7
----------------------------------------------------------------------
I love this stuff.  I put it over chicken and even steaks.  It's very sweet, and just a 
bit spicy.  Just a little goes a long way.


______________________________________________________________________
```

```
A review of assigned to cluster  8
----------------------------------------------------------------------
this is a great product. Perfect for the celiac searching for biscuit mix or pancake mix
that actually tastes good.  Also great topping for chicken pot pie.


______________________________________________________________________
A review of assigned to cluster  9
----------------------------------------------------------------------
The number one ingredient is chicken... not organic chicken just the hormone pumped junk
you wouldn't eat yourself. I will stick to Natural Balance which is sold here on Amazon
as well.


______________________________________________________________________
```

In [36]:

```
#considers sample of 3 random reviews for cluster 0

print(df.iloc[df.groupby(['Bow Clus Label']).groups[0][3]]['Text'])
print("_" * 70)
print(df.iloc[df.groupby(['Bow Clus Label']).groups[0][15]]['Text'])
print("_" * 70)
print(df.iloc[df.groupby(['Bow Clus Label']).groups[0][25]]['Text'])
```

```
I have nothing but good service ever since I started ordering from amazon.com.  Keep up
the good work.
______________________________________________________________________
I am from England and I was raised on this custard. If you like vanilla custard/pudding
you will love this. It's rich and creamy and has lots of vanilla taste.  I love that it
comes in a big canister and I can make as much as I want to. I would never make a trifle
with anything else.
______________________________________________________________________
It was shipped in a very nice package and I have no complain with the seller. My plant g
rowing bigger and now I am looking for a bigger pot to transplant. It's good....if you a
re looking for a a workspace plant that needs minimal maintenance this is the plant you
might need.
```

In [79]:

```
#consider sample of 3 random reviews for cluster 4

print(df.iloc[df.groupby(['Bow Clus Label']).groups[3][3]]['Text'])
print("_" * 70)
print(df.iloc[df.groupby(['Bow Clus Label']).groups[3][15]]['Text'])
print("_" * 70)
print(df.iloc[df.groupby(['Bow Clus Label']).groups[3][20]]['Text'])
```

```
fast and great service, my Cavashon Loves this low fat dog food. Thanks. Arrived in grea
t condition, thank you thank you thank you
______________________________________________________________________
Tasty, convenient bars for people with celiac disease. They seem to have gotten smaller
over time, but the taste and convenience outweigh the reduction in size (and the price).
______________________________________________________________________
I love these chips. They are so delicious, it is so hard to eat just one bag. All the fl
avors allow you to taste them all.  My favorite flavor is the cheese ones. Go ahead and
try them.  They are so delicious.
```

In [38]:

```
#consider sample of 3 random reviews for cluster 4

print(df.iloc[df.groupby(['Bow Clus Label']).groups[5][3]]['Text'])
print("_" * 70)
print(df.iloc[df.groupby(['Bow Clus Label']).groups[5][15]]['Text'])
print("_" * 70)
print(df.iloc[df.groupby(['Bow Clus Label']).groups[5][25]]['Text'])
```

```
I was pleasantly surprised by the stronger than I expected ginger flavor in this produc
```

```
t.  It is excellent, if you like ginger, try this.  Best on hot biscuits!  Update.  I've
just reordered, wish they sold it by the gallon, its different, something you can't find
locally and excellent.
______________________________________________________________________________________
The potato bread was easy to mix, rises well in my breadmaker, cooks as it should and be
st of all was very tasty. The bread has a near white bread texture, and was very good. I
purchased all of the Hodgson mixes and this was one of my favorites.
______________________________________________________________________________________
My daughter loves these snacks and can't get them in China where she is working now so I
ordered them for her. Very satisfied and she was very appreciative.
```

Analysis of K means for BOW:

Of all the clusters, 0, 4 and 6 accounts to more % of reviews, undertsanding differences between these 3 clusters is key. Also, the clusters 2 and 9 have only 1 review

If we observe the top terms per cluster, The cluster 4 which consists of LIKE AND LOVE, which are top centroid features and can say this cluster consists of all positive reviews, let us obersve few reviews of each cluster and try to understand the differences

By reading the cluster 2 and 9 which contains only one review, which is clearlt negative reviews and we can conluded customers didnt liked the product at all and not word is used extensively

By reading random reviews of cluster 0, we can easily say that these reviews are extremly positive of the product usage and customers are very happy with the product

By reading random reviews of cluster 4, we can say that the key word **BUT** is repeating acorss the review which indicates some kind of peopel agree with most of the things related to the products but their is something which is slightly disagree with product quality or delivery or some thing less than their expectation

K means using TFIDF

In [39]:

```
#tfidf vector initilization
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vect = TfidfVectorizer()
tfidf = tfidf_vect.fit_transform(final['CleanedText'].values)
tfidf.shape
```

Out[39]:

```
(4985, 8565)
```

In [40]:

```
from sklearn.cluster import KMeans
model_tf = KMeans(n_clusters = 10, n_jobs = -1,random_state=99)
model_tf.fit(tfidf)
```

```
C:\Users\waliullah\Anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:792: FutureWar
ning: 'n_jobs' was deprecated in version 0.23 and will be removed in 1.0 (renaming of 0.
25).
  warnings.warn("'n_jobs' was deprecated in version 0.23 and will be"
```

Out[40]:

```
KMeans(n_clusters=10, n_jobs=-1, random_state=99)
```

In [41]:

```
labels_tf = model_tf.labels_
cluster_center_tf=model_tf.cluster_centers_
```

In [42]:

```
cluster_center_tf
```

Out[42]:
```
array([[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.        , 0.00061507,
        0.        ]])
```

In [43]:
```
# to understand what kind of words generated as columns by BOW
terms1 = tfidf_vect.get_feature_names()
```

In [44]:
```
terms1[1:10]
```

Out[44]:
```
['aback',
 'abandon',
 'abat',
 'abbi',
 'abbott',
 'abdomin',
 'abid',
 'abil',
 'abl']
```

In [45]:
```
from sklearn import metrics
silhouette_score_tf = metrics.silhouette_score(tfidf, labels_tf, metric='euclidean')
```

In [46]:
```
silhouette_score_tf
```

Out[46]:
```
0.016420551824604532
```

In [47]:
```
# Giving Labels/assigning a cluster to each point/text
df1 = df
df1['Tfidf Clus Label'] = model_tf.labels_
df1.head(5)
```

Out[47]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenomi
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenomi
3782	3783.0	B00016UX0K	AF1PV3DIC0XM7	Robert Ashton	1.0	
1205	1206.0	B005O072PC	A3BD5B8Y8MY25X	J. L. K. "special_k"	13.0	
1275	1276.0	B000WNJ73Q	A394MHK3CSDGUV	kaleinor	2.0	

In [48]:

```
# How many points belong to each cluster ->

df1.groupby(['Tfidf Clus Label'])['Text'].count()
```

Out[48]:

```
Tfidf Clus Label
0     279
1     308
2     326
3     283
4     502
5     289
6    2086
7     455
8      68
9     389
Name: Text, dtype: int64
```

In [49]:

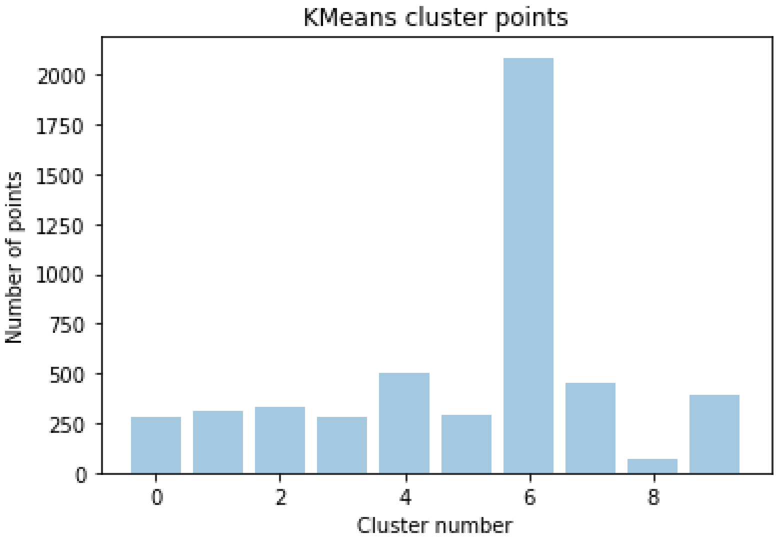
```
#Refrence credit - to find the top 10 features of cluster centriod
#https://stackoverflow.com/questions/47452119/kmean-clustering-top-terms-in-cluster
print("Top terms per cluster:")
order_centroids = model_tf.cluster_centers_.argsort()[:, ::-1]
for i in range(10):
    print("Cluster %d:" % i, end='')
    for ind in order_centroids[i, :10]:
        print(' %s' % terms1[ind], end='')
        print()
```

```
Top terms per cluster:
Cluster 0: tea
 green
 drink
 ice
 tast
 like
 flavor
 use
 water
 love
Cluster 1: pancak
 mix
 waffl
 gluten
 bisquick
 make
 free
 use
 product
 biscuit
Cluster 2: coffe
 tast
 decaf
 cup
 flavor
 like
 bitter
 strong
 smooth
 good
Cluster 3: chocol
 hot
 cocoa
 cup
 tast
 keurig
 tri
 dark
 good
 grove
Cluster 4: love
 great
 product
 snack
 flavor
 tast
 eat
 good
 one
 get
Cluster 5: dog
 food
 newman
 love
 eat
 organ
 cat
 feed
 year
 treat
Cluster 6: like
 tast
 good
 product
 flavor
 use
 one
 tri
 order
 would
Cluster 7: chip
```

```
 flavor
 bag
 salt
 potato
 kettl
 like
 vinegar
 great
 love
Cluster 8: popcorn
 pop
 kernel
 popper
 white
 hull
 small
 corn
 amish
 tender
Cluster 9: store
 amazon
 price
 find
 local
 product
 groceri
 buy
 good
 order
```

In [50]:

```
# visually how points or reviews are distributed across 10 clusters

plt.bar([x for x in range(10)], df1.groupby(['Tfidf Clus Label'])['Text'].count(), alph
plt.title('KMeans cluster points')
plt.xlabel("Cluster number")
plt.ylabel("Number of points")
plt.show()
```



In [51]:

```
# Reading a review which belong to each group.
for i in range(10):
    print("4 review of assigned to cluster ", i)
    print("-" * 70)
    print(df1.iloc[df1.groupby(['Tfidf Clus Label']).groups[i][5]]['Text'])
    print('\n')
    print(df1.iloc[df1.groupby(['Tfidf Clus Label']).groups[i][10]]['Text'])
    print('\n')
    print(df1.iloc[df1.groupby(['Tfidf Clus Label']).groups[i][20]]['Text'])
    print('\n')
    print("_" * 70)
```

```
4 review of assigned to cluster  0
----------------------------------------------------------------------
I'm trying several of the Wu Yi teas. I like this one particularly because of the subtle
citrus taste. It also has a natural sweetness.<br /><br />The ingredients include Organi
c Wu-Li Cliff Oolong Tea (700 mg), Organic Black Tea (600 mg), Organic Green Tea Extract
(100 mg), Proprietary blend (600 mg), plus some Ginseng (panax), Orange Peel, Lemon Gras
s, and Guarana. Other Ingredients: Natual Orange and Citrus flavors.<br /><br />My goal
is to drink more tea and less coffee. I am enjoying trying a different kind of tea every
day. I'll be coming back to this one.


The Price is great, The serving is just right for a snack, and ITS A COOKIE. Buy a bunch
and they send it to your house!!


thank you for this product - we use it all the time and appreciate your promptness and t
he price was excellent.  Thanks again.


______________________________________________________________________
4 review of assigned to cluster  1
----------------------------------------------------------------------
This  coffee is the smoothest dark roast coffee I have ever tasted,and it was a pleasure
to sip this full bodied coffee before breakfast and after dinner.


This popcorn is much smaller than your average popcorn, and it is virtually hulless.  Th
at is exactly what I was looking for in a popcorn because my 3 year old LOVES popcorn.
With 3 year old children, you have to be concerned with the hulls causing choking or hur
ting their tender gums.  My son has no trouble eating this popcorn.  It is soft and ligh
t, doesn't get stuck in your teeth nearly as much, and has a great flavor.<br /><br />Th
is is now my popcorn of choice for our whole family.  I usually pop it in my 10 year old
cheap hot air popper with no problems.  I get very few unpopped kernels. Another reviewe
r stated that it flies out before popping, but I have not experienced this problem (it c
ould depend on the popper, I suppose).  I have tried microwaving it in a paper bag, but
I always seem to end up burning it.  I think it takes a little more trial and error for
microwaving to get the time just right, so I just stick with my hot air popper. I get gr
eat popcorn every time!


My daughter was diagnosed with celiac about 2 years ago, and has gone without pancakes e
ver since...until I happened to come across this at the grocery store.  Really wasn't ex
pecting good results - but it's very good!  The non-celiac members of the family even pr
efer the gf version to the regular!  Try it!


______________________________________________________________________
4 review of assigned to cluster  2
----------------------------------------------------------------------
I thought it was okay but not Harbanero BBQ sauce from Hell.I would probally not get it
again.Also it was not worth the money I thought it was kind of a ripoff it was 10$ for b
arbque sauce you could have made at home.


The first time I tried this product  was 2 years ago at the Venetian spa restaurant in L
as Vegas  having  a cup of tea with some friends. We all remarked how delicious and tast
y it made our tea.  We've been using it ever since. A little goes a long way and I use i
t anytime I would ordinarily use a sweetener.  Since the artificial sweeteners are so un
safe and unhealthy and I have a sweet tooth, I felt this was a safe alternative. Not onl
```

```
y is it safer, but there's no aftertaste and doesn't raise your blood sugar level. This
particular stevia has absolutely no bitter taste to it. I've tried others before and sin
ce and there's no comparison.


Hands down, by far the absolute worst tasting tea I have ever had.... and I have ALLOT!
I am an avid tea drinker and I just can't get it down. The benefits are supposed to be a
mazing so I'll keep it around to water down and sugar up but wow it's bad. So strong to
o, tastes like dirt (*that's been pissed on in the woods).



______________________________________________________________________
4 review of assigned to cluster  3
----------------------------------------------------------------------
Very disappointed with purchase.  The dates were so dried up where it tasted like leathe
r instead of dates.  Must have been sitting there for a long long time.


This was the 2nd time that I ordered the Smokehouse USA Chicken Stix. The 1st order had
USA printed right on the bag, this one had a USA sticker over the China sticker.  In add
ition, the product looked entirely different. Obviously I am concerned. I pay more to en
sure I am not feeding my "fur kids" products from China. If it happens again, I'll stop
ordering all Smokehouse products on-line.


I expected from the extremely positive reviews on the site for full flavor healty chips.
In reality I get an OKAY taste with a strange aftertaste on basically all of the flavor
s. Chedder and Salt&Pepper honestly being some of the worst tasting chips I've ever had.
If you're just looking for a great tasting brand of chips with health as a #2 on the lis
t there are much much better brands out there. In the end I compare these chips to diet
soda vs regular; Some won't tell the difference while some will immediately sense it and
hate it.



______________________________________________________________________
4 review of assigned to cluster  4
----------------------------------------------------------------------
These sticks definitely don't look like ones in the picture.<br />They are much thinner
and IMHO not worth the money. I should have returned it but instead gave my dog (GSD) 5
at a time to keep her busy.<br /><br />I'd stick with the Redbarn or equivalent from ano
ther website.


First off, I received a bag of this coffee via the Vine Program to review at no cost.  S
econdly, I am more of a bold, "slap-me-in-the-face" French Roast coffee drinker (or Star
Bucks Gold Coast), so I had my reservations about this decaf.<br /><br />And ... Well, f
or me, this coffee had a somewhat stale aroma upon opening the package, an OK flavor onc
e brewed (sort of a nutty strange flavor at first), and once consumed, left an after-tas
te in my mouth.<br /><br />It is definitely a different roast than I am used to, but it
is "decent" for a pre-ground Melita decaf.  More of a medium roast as opposed to being a
dark bold roast in my opinion.  Overall ... not too bad.<br /><br />I give this coffee a
so-so recommendation.


Love this product.  Very flavorable to most anything.  Works great on lunch meat, ham, p
otato salad, more.  Something similar is sold at Honey Baked Ham locations, but this pro
duct is much better.



______________________________________________________________________
4 review of assigned to cluster  5
----------------------------------------------------------------------
I got this after a mention in the NY Times and immediately became an addict.  It's a lit
tle hot but not too much, and it has an amazing depth of flavor -- a lot more than the p
lain version.  I use it in chicken salad to cut back on the amount of mayo, to add flavo
```

```
r to blah soups and as a substitute for a lot of the oil in salad dressings.  A little b
it over steamed veggies and added to yogurt over baked potatoes (instead of sour cream)
is also great.


Cats are such finicky eaters some times.  My old lady kitty does not like the canned cat
foods, but she always went nuts whenever I'd open a can of tuna for myself.  So I gave h
er some one time.  These Tuna Cups are a great way to keep the tuna fresh enough over a
few days, and good for travel.<br /><br />Same items can be purchased at grocery store,
but buying in quantity through Amazon saves some $$.  Thank you


Pkg says soy free. Ingredients (and my stomach/bladder) say otherwise!<br />Also, lots o
f fiber in this one. If you can handle fiber (and soy) this food will fill you up like c
razy. If you have IBS and have trouble with fiber,or are eating a low residue diet be ca
reful....<br />I'm going back to the bread. Eating a sandwich on it grilled is a fair su
bstitute for the dreaded wheat...



______________________________________________________________________
4 review of assigned to cluster  6
----------------------------------------------------------------------
Cutting sugar out of our diet,we went to splenda. Then finding that it was also unhealth
y and baking with it made foods flat and dry. We went to stevia no calorie powder, after
much research proved that stevia has been used for hundreds of years with no bad health
effects. The trick to using stevia is to use tiny amounts and then taste it and add a bi
t more at a time to suit your taste. I also found that using a stevia that has an 80% ra
tio is the best tasting and this is the best that I have found. Because of using VERY SM
ALL amounts this jar lasts for months. This is also the best price ANYWHERE! Swanson is
a trusted vendor as is Amazon.


I was sorely disappointed in these cookies.. They are pretty tasteless and damn hard to
o.. I wouldn't buy them again and I've certainly tasted much better low-cal cookies.. I
definitely recommend saving your money, folks..


I grew up in Ohio and lived in the woods where sassafrass trees were plentiful. I would
dig tender roots or "bark" a few trees for a delightful tea.  I purchased Breezy Morning
Sassafrass tea and although was quite good it was very expensive for only 20 average siz
e tea bags and shipping was as much as the product.  Not a good bargain.  I switched to
Pappy's extract.



______________________________________________________________________
4 review of assigned to cluster  7
----------------------------------------------------------------------
These are more like a cracker - good for a little fill in for between meals.


If you are truly looking for a decaf that will no longer make your heart race, this is n
ot the coffee for you. It is quite flavorful, so I do recommend it to those who can stil
l tolerate a bit of kick in their coffee.


I Love these potato chip they r soooo good! =) sweet and spicy but not overwhelming just
enough to work up the palate, not salty just enough to balance out the sweetness I will
be ordering these again soon. LOVe IT!



______________________________________________________________________
4 review of assigned to cluster  8
----------------------------------------------------------------------
My bottle arrived just in time for a sushi dinner and game night planned with friends.
I had three soy sauces to choose from.  After some sampling we were all using the Bluegr
```

```
ass!  Everyone enjoyed the light smoothness a subtle sweetness.  We were splashing right
on every piece, rather that dunking/dipping. Friends wanted to know where I found it.
I'm going back to order more right now, and I'll be ordering 3 extra bottles because the
y are going to make a great little gift that's affordable and unique; instead of the tra
ditional bottle of wine that so common it's boring.


This is the first time we bought this tea from Amazon, we used to get it directly from R
evolution Tea. I don't know if that made any difference, or if it was just a change in t
he tea's composition (along with the packaging), but what used to be my wife's favorite 
tea now has a metallic smell and aftertaste, with much less of a pomegranate flavor and 
now a strong, almost cheap-generic-green-tea flavor. Really disappointing, especially si
nce we're stuck with six boxes.


I really liked the taste. It is well priced too. It is a great and healthy alternative t
o table salt. I even used it on salad.


_______________________________________________________________________
4 review of assigned to cluster  9
-----------------------------------------------------------------------
I really liked these chips as did my entire family. I wish you could pick and choose the
flavors in the variety pack. I would love to try the sweet potato and chili lime and lea
ve the salt and pepper behind.


As someone who suffers occasionally from digestive difficulties, these Kavli Crispy Thin
s are among the few things that I can eat at such times.  I always keep these crackers a
round to add to soups.  However, there are also times when they are the only thing I can
properly digest.  This is a great item to have around when having digestive difficulties
associated with the flu or stress.  For me they work better than the "bananas, rice, app
lesauce or toast" of the BRAT diet used in relieving IBS. These thin, easy to chew crack
ers are a really great product.


I brought these cookies as a "dish to pass" at my family Christmas get together. they we
re a hit. The cookies were fresh, and unbroken. packaging was well done, and delivery wa
s prompt. my sisters had fun with the fortunes, by adding the phrase, "while in bed" aft
er reading each fortune.<br /><br />I will use the Amazon vendor, "House of Rice" again.


_______________________________________________________________________
```

Analysis of K means for TF_IDF:

__Of all the cluster 4 accounts to more % of reviews i,e above 4000.

If we observe the top terms per cluster, The clusters based on the products and product wise like and dislikes. for example, if we oberve cluster 8, the reviews talk more about chips, potatos, and other products which are like snacks

In these, its better to understand the cluster center top features rather than individual reviews.

Average Word to Vector

In [52]:
```
# Train your own Word2Vec model using your own text corpus
i=0
list_of_sent=[]
```

```
for sent in final['CleanedText'].values:
    list_of_sent.append(sent.split())
```

In [53]:

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

```
realli good idea final product outstand use decal car window everybodi ask bought decal
made two thumb
*****************************************************************
['realli', 'good', 'idea', 'final', 'product', 'outstand', 'use', 'decal', 'car', 'windo
w', 'everybodi', 'ask', 'bought', 'decal', 'made', 'two', 'thumb']
```

In [54]:

```
# removing html tags and apostrophes if present.
import re
def cleanhtml(sentence): #function to clean the word of any html-tags
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', sentence)
    return cleantext
def cleanpunc(sentence): #function to clean the word of any punctuation or special char
    cleaned = re.sub(r'[?|!|\'|"|#]',r'',sentence)
    cleaned = re.sub(r'[.|,|)|(|\|/]',r' ',cleaned)
    return  cleaned
```

In [55]:

```
i=0
list_of_sent_train=[]
for sent in final['CleanedText'].values:
    filtered_sentence=[]
    sent=cleanhtml(sent)
    for w in sent.split():
        for cleaned_words in cleanpunc(w).split():
            if(cleaned_words.isalpha()):
                filtered_sentence.append(cleaned_words.lower())
            else:
                continue
    list_of_sent_train.append(filtered_sentence)
```

In [ ]:

```
vector_size=100
```

In [57]:

```
import gensim
# Training the wor2vec model using train dataset
w2v_model=gensim.models.Word2Vec(list_of_sent_train, vector_size=100, workers=4)
```

In [58]:

```
import numpy as np
sent_vectors = []; # the avg-w2v for each sentence/review is stored in this train
for sent in list_of_sent_train: # for each review/sentence
    sent_vec = np.zeros(100) # as word vectors are of zero length
    cnt_words =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        try:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
        except:
            pass
    sent_vec /= cnt_words
    sent_vectors.append(sent_vec)
sent_vectors = np.array(sent_vectors)
sent_vectors = np.nan_to_num(sent_vectors)
sent_vectors.shape
```

Out[58]:

```
(4985, 100)
```

K Means CLustering for Avg word to vectors

In [59]:

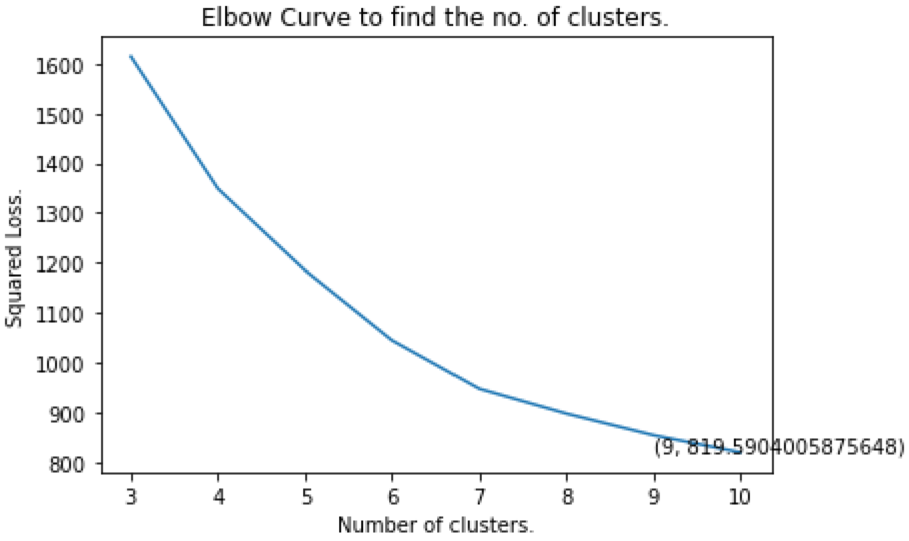
```
# Number of clusters to check.
num_clus = [x for x in range(3,11)]
num_clus
```

Out[59]:

```
[3, 4, 5, 6, 7, 8, 9, 10]
```

In [60]:

```
# Choosing the best cluster using Elbow Method.
# source credit,few parts of min squred loss info is taken from different parts of the
# this is used to understand to find the optimal clusters in differen way rather than u
squared_errors = []
for cluster in num_clus:
    kmeans = KMeans(n_clusters = cluster).fit(sent_vectors) # Train Cluster
    squared_errors.append(kmeans.inertia_) # Appending the squared loss obtained in the

optimal_clusters = np.argmin(squared_errors) + 2 # As argmin return the index of minimu
plt.plot(num_clus, squared_errors)
plt.title("Elbow Curve to find the no. of clusters.")
plt.xlabel("Number of clusters.")
plt.ylabel("Squared Loss.")
xy = (optimal_clusters, min(squared_errors))
plt.annotate('(%s, %s)' % xy, xy = xy, textcoords='data')
plt.show()

print ("The optimal number of clusters obtained is - ", optimal_clusters)
print ("The loss for optimal cluster is - ", min(squared_errors))
```



```
The optimal number of clusters obtained is -  9
The loss for optimal cluster is -  819.5904005875648
```

In [61]:

```
# Training the best model --
from sklearn.cluster import KMeans
model2 = KMeans(n_clusters = optimal_clusters)
model2.fit(sent_vectors)
```

```
KMeans(n_clusters=9)
```

Out[61]:

In [62]:
```
word_cluster_pred=model2.predict(sent_vectors)
word_cluster_pred_2=model2.labels_
word_cluster_center=model2.cluster_centers_
```

In [63]:
```
word_cluster_center[1:2]
```

Out[63]:
```
array([[ 2.97770039e-01,  4.33285802e-01,  1.04807388e-03,
         1.41850606e-01,  4.85069192e-02, -4.39646394e-01,
         3.45700605e-01,  3.93962006e-01, -2.87123992e-01,
         9.05883804e-04,  1.95094611e-01, -2.04462879e-01,
         6.42368877e-02, -1.03781047e-01,  5.39754948e-02,
        -2.22620047e-01,  1.54635033e-01, -4.82459264e-01,
        -1.52842011e-01, -8.27133050e-01,  1.04009122e-01,
         8.03299383e-02, -9.65052480e-02, -7.62683143e-03,
        -4.54422366e-01, -4.99152779e-03, -2.23128314e-01,
        -2.52930300e-01, -3.00263015e-02, -1.01688197e-01,
         2.90673683e-01,  9.25762006e-02, -5.70009608e-02,
        -5.53611791e-02, -1.39321250e-01,  4.17029860e-01,
        -1.72431287e-01, -1.43766007e-01, -1.35696035e-01,
        -7.84194974e-01,  5.75856137e-02, -3.76448343e-01,
        -1.25271509e-01,  1.27192182e-01,  2.27435965e-01,
        -1.10459405e-02, -3.71644328e-01,  2.07651520e-02,
         3.23662472e-01,  1.66144826e-01,  2.81013745e-02,
        -3.99828662e-01,  1.30941917e-01, -2.29755008e-01,
        -2.76465574e-01,  1.84475974e-01,  2.20360954e-01,
         1.09782180e-01, -5.05657162e-01,  3.27083643e-02,
        -3.53201182e-02,  1.93665378e-01, -7.27729599e-02,
        -1.68167273e-01, -4.96079563e-01,  4.59261651e-01,
         1.09780297e-01,  6.06671145e-02, -3.57360144e-01,
         2.80058705e-01, -4.57922284e-01,  1.16778751e-01,
         2.32815491e-01, -1.60592215e-01,  4.28951897e-01,
         3.39953114e-01,  9.98965812e-02,  2.45811086e-02,
        -3.39693043e-01, -1.26303522e-01,  8.47159986e-03,
        -6.24809295e-02, -1.18867293e-01,  4.04400039e-01,
        -3.31583142e-02, -2.09467503e-01, -1.13615192e-01,
         2.34949132e-01,  6.00831357e-01,  2.03904592e-01,
         3.01512162e-01,  2.44110647e-01, -6.60909198e-03,
         1.90956685e-01,  5.74312300e-01,  2.05615032e-01,
         1.55608763e-02, -3.71224923e-01,  2.74746237e-02,
         3.46828746e-04]])
```

In [64]:
```
# Giving Labels/assigning a cluster to each point/text
dfa = df1
dfa['AVG-W2V Clus Label'] = model2.labels_
dfa.head(2)
```

Out[64]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	7

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	10

In [65]:

```
# How many points belong to each cluster ->
dfa.groupby(['AVG-W2V Clus Label'])['Text'].count()
```

Out[65]:

```
AVG-W2V Clus Label
0     863
1     295
2    1002
3     223
4     357
5     294
6     398
7    1202
8     351
Name: Text, dtype: int64
```

In [66]:

```
# Reading a review which belong to each group.
for i in range(optimal_clusters):
    print("A review of assigned to cluster ", i)
    print("-" * 70)
    print(dfa.iloc[dfa.groupby(['AVG-W2V Clus Label']).groups[i][0]]['Text'])
    print('\n')
    print(dfa.iloc[dfa.groupby(['AVG-W2V Clus Label']).groups[i][1]]['Text'])
    print('\n')
    print("_" * 70)
```

```
A review of assigned to cluster  0
----------------------------------------------------------------------
I was given a pack of this coffee as a gift and it had been sitting around for quite a w
hile (over a year) when I decided to try it.  It was without a doubt one of the best cof
fees that I have tasted.  Very smooth and flavorful.  I would highly recommend this.  Af
ter that I ordered a supply from Amazon. They are vacuum packed (I believe much better t
han store bought coffee).  I may try the decaf version of this to see how that tastes.


I love this stuff.  I put it over chicken and even steaks.  It's very sweet, and just a
bit spicy.  Just a little goes a long way.


______________________________________________________________________
A review of assigned to cluster  1
----------------------------------------------------------------------
They are OK, but my husband only eats them when we are in car. We keep them there for a
celiac snack. (Hard to find when roaming around.) He won't eat them otherwise. They are
definitely raspberry, but are kind of dry.<br />  We have never tried any other snack ba
rs, so in all honesty they may be the norm. I don't know.<br />  We DO like Enjoy Life's
other products.<br />  They are GF, have no MSG or related products or aspartates.
```

```
It appears a little watery, but the taste is not bad at all.  If you have kids that are
as fascinated by the Keurig as you are, they'll probably enjoy this.


______________________________________________________________________
A review of assigned to cluster  2
----------------------------------------------------------------------
I have tried many bread machine mixes and most have been okay but this one rates an "exc
ellent" because not only is it a very good basic loaf, it also works very well with addi
tives, either prepared entirely in the bread machine, or (after the second rise) removed
from the machine, shaped and baked on a stone in the oven.<br />Stretched and flattened
with a filling of either savory or sweet ingredients, then rolled and set to rise and th
en bake in a loaf pan, it produced excellent herb and onion bread, cinnamon raisin bread
and a brown butter and seed bread.<br />I buy it via the subscribe-and-save plan as I us
e it often.


I have a 2 year old Portuguese Water Dog who always seemed to have a sensitive stomach,
and a 15 year old Shepherd X dog who was beginning to lose weight, she slept almost all
the time, and she was getting very fussy about what she'd eat. (I believe that with a do
gs sense of smell they KNOW exactly what is in their food; however they have no choice b
ut to eat what we feed them...just Google what is in most commercial dog foods, and yo
u'll see why your dog may not be thrilled to eat it up...if a dog can detect cancer in a
person, they can detect all sorts of other things that aren't supposed to be eaten.) Any
way...I used to feed Purina lamb and rice to my dogs, but I was finding my young Porty w
as having trouble with it, and my 15 yr old would eat a few bits and then leave her dis
h. So, I started making their food, and came up with a wonderfully nutritious and tasty
recipe. The dogs loved my homemade dog food, and my 15 yr old began putting on weight ag
ain, but I didn't always have time to make their food. And, homemade dog food is kind of
hard on the wallet. So, I started to research all the natural dog foods out there, compa
re costs etc. I arrived on Harmony Farms, and guess what...my dogs LOVE it just as much
as they love my home made dog food. My young Porty no longer has stomach troubles and sh
e looks great; my 15 yr old finishes her meals and looks for more. Both have shiny coats
and energy. No one can believe my old girl is 15 years old. I have told my friends about
Harmony Farms dog food, and when they've switched over, they've reported similar positiv
e results. My brother-in-law's dog had always been a very fussy eater, and had skin prob
lems with a thinning coat. When they started to feed her Harmony Farms, she became an ea
ger eater, more happy and outgoing, and her skin and coat condition has completely clear
ed up (...I'm thinking she had an allergy to whatever was in her other food, which I thi
nk was Iams). Other friends have commented on how happy their dogs are at meal time no
w...and they had all been feeding their dogs premium, top of the line dogs foods AND pay
ing top of the line prices. Harmony Farms products are very reasonably priced. Not the c
heapest dog food on the shelf, but, in my opinion, you are buying the best quality food
on the market, so it's an incredibly GOOD DEAL! It may also mean you make fewer trips to
the veterinarian's office. I and my friends are so happy we have found this food for our
dogs. :)


______________________________________________________________________
A review of assigned to cluster  3
----------------------------------------------------------------------
Awesome little snack treat for your favorite pup/dog.  My two pekeingese just love thes
e.  You can use them for training as well as just a treat.  They come in a variety of fl
avors.  My pups like the liver the best.  For senior dogs, these are perfect if they do
n't have the "jaw" power or teeth use anymore.  They are easy to carry so you can treat
your dog whenever you want to for example on a walk around the neighborhood.  They are j
ust 3 calories each so you don't have to worry about their weight.  Absolutely recommend
these to every dog owner!


For those who love salt and vinegar potato chips, this is the one to choose. The flavor
is zippy and tart, with no unpleasant chemical aftertaste like the less "natural" versio
ns of this snack. The 2-ounce bags are just right to share at lunch. The chips are a lit
tle greasy, and that's why I've given them 4 stars instead of 5.
```

```
______________________________________________________________________
A review of assigned to cluster  4
----------------------------------------------------------------------
These things are AWESOME. Perfect size, truly wafer thin, consistent thickness of dark c
hocolate over minty interior. They are like what would happen if you could run a steamro
ller over a junior mint. I first tried them when I was living in Germany 16 years ago, a
nd to this day no other mint comes close. And the copycat ones? Disgusting. Go for the o
riginal. You won't regret it. Each mint is in its own sleeve, so you can pass them out w
ithout ever touching one. Elegant and delicious. Love them!


When I got my Keurig brewer last month, I bought two different types of hot chocolate.
The Grove Square Hot Chocolate is far superior.  I have so far brewed it on 6oz and 8oz
with good flavor at both sizes.  The flavor is rich and it leaves no residue in the bott
om of the kcup (the other brand does, maybe I got a bad batch?).  Overall, I will happil
y buy this one again.



______________________________________________________________________
A review of assigned to cluster  5
----------------------------------------------------------------------
I've tried a few different 'Dirty Martini' mix's, etc. but I prefer the actual juice fro
m the olive jar. Well, as any experienced dirty martini drinker knows; you soon have way
too many olives and not enough juice. Boscoli Family Dirty Martini Olive Juice is the re
al thing. I usually buy 4 of the 25oz. bottles and that'll last me several months. Shipp
ing is pricey but the bottles come not only packed in peanuts but also bubble wrapped. T
hey pack 'em in a sturdy cardboard box and tape it up real well. I live in a rural area
and usually get my order in about 4 days. Fantastic experience all the way around.


The noodles in the box were all broken.  The sauce was over salted and did not have a go
od flavor.  I threw out most of the skillet.  I would recommend not purchasing this prod
uct.



______________________________________________________________________
A review of assigned to cluster  6
----------------------------------------------------------------------
Cutting sugar out of our diet,we went to splenda. Then finding that it was also unhealth
y and baking with it made foods flat and dry. We went to stevia no calorie powder, after
much research proved that stevia has been used for hundreds of years with no bad health
effects. The trick to using stevia is to use tiny amounts and then taste it and add a bi
t more at a time to suit your taste. I also found that using a stevia that has an 80% ra
tio is the best tasting and this is the best that I have found. Because of using VERY SM
ALL amounts this jar lasts for months. This is also the best price ANYWHERE! Swanson is
a trusted vendor as is Amazon.


High quality coca products do not have a paper wraps since they can harbor dirt and bact
eria and the printing ink can change or contaminate the delicate flavor of the tea bags.
<br />Organic coca tea bags do not contain preservatives or additives; therefore they mu
st be properly stored. All coca tea  Air Tight Bags are re-sealable with zip lock closur
es and once opened the bags must be kept closed at all time to avoid contamination in a
dry, cool, dark place away from strong-flavored foods.  An extra airtight container is r
ecommended when the tea is stored over one month.<br />Do not store coca tea  products i
nside a refrigerator if you have produce inside. Many fruits and vegetables, especially
if they have been damaged, give off ethylene gas as they ripen. Coca tea bags are very s
ensitive to the presence of even very low levels of ethylene gas. The refrigerator acts
as a trap for the ethylene gas given off by the generating varieties, allowing it to bui
ld up to damaging levels. Although not hazardous to humans, the ethylene gas leads to th
e early aging and rotting of the tea.
```

```
______________________________________________________________________
A review of assigned to cluster  7
----------------------------------------------------------------------
This is not jerky, this is processed, hard like a rock, very greasy and stale smelling s
tripe of something that you can't break into anything smaller than 2 inches long and tha
t certainly is not the size of a training treat! The dogs- 45lb dogs that will eat anyth
ing- were not impressed, it was hard to chew, and it sounded like they were crunching ro
cks, most of them spat it out after a few chews, left it there, this would be the first 
time they would not eat something in their entire lives, these dogs will work for lettuc
e. Where is a zero star button?


This stuff isn't bad at all. But you know how you can just eat spoon after spoon of real
ly good caviar without toast points or crackers or anything? This roe is more for sprink
ling over fish dishes, morning eggs or a caesar salad. All in all, for the price, it's p
retty good,



______________________________________________________________________
A review of assigned to cluster  8
----------------------------------------------------------------------
the pop chips are really incredible. They are very flavorful and crispy. My Favorite is 
the BBQ, but that is just me, they are all good. I would highly recommend these to anyon
e who is watching their weight of just for overall better health.


4-year old Elkhound loves this food, and it keeps her in great health.  Amazon has the b
est price I've found for this food, and it's even better with subscribe and save.  I hig
hly recommend it.


______________________________________________________________________
```

Clustering DBSCAN

In [67]:

```python
from sklearn.cluster import DBSCAN
```

In [68]:

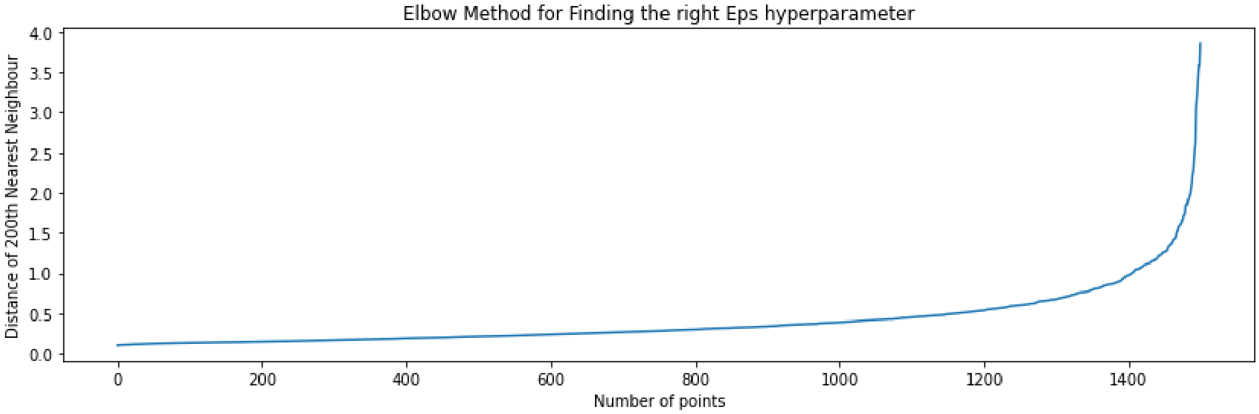
```python
# Computing 200th Nearest neighbour distance
minPts = 2 * 100
# Lower bound function copied from -> https://gist.github.com/m00nlight/0f9306b4d4e61ba
def lower_bound(nums, target): # This function return the number in the array just grea
    l, r = 0, len(nums) - 1
    while l <= r: # Binary searching.
        mid = int(l + (r - l) / 2)
        if nums[mid] >= target:
            r = mid - 1
        else:
            l = mid + 1
    return l

def compute200thnearestneighbour(x, data): # Returns the distance of 200th nearest neig
    dists = []
    for val in data:
        dist = np.sum((x - val) **2 ) # computing distances.
        if(len(dists) == 200 and dists[199] > dist): # If distance is larger than curre
            l = int(lower_bound(dists, dist)) # Using the Lower bound function to get t
            if l < 200 and l >= 0 and dists[l] > dist:
                dists[l] = dist
        else:
            dists.append(dist)
            dists.sort()
    return dists[199] # Dist 199 contains the distance of 200th nearest neighbour.
```

In [69]:

```
# Computing the 200th nearest neighbour distance of some point the dataset:
twohundrethneigh = []
for val in sent_vectors[:1500]:
    twohundrethneigh.append( compute200thnearestneighbour(val, sent_vectors[:1500]) )
twohundrethneigh.sort()
```

In [70]:

```
# Plotting for the Elbow Method :
plt.figure(figsize=(14,4))
plt.title("Elbow Method for Finding the right Eps hyperparameter")
plt.plot([x for x in range(len(twohundrethneigh))], twohundrethneigh)
plt.xlabel("Number of points")
plt.ylabel("Distance of 200th Nearest Neighbour")
plt.show()
```



Conclusions for Elbow Method

The Knee point seems to be 5. So Eps = 5

In [71]:

```
# Training DBSCAN :
model = DBSCAN(eps = 5, min_samples = minPts, n_jobs=-1)
model.fit(sent_vectors)
```

Out[71]:

```
DBSCAN(eps=5, min_samples=200, n_jobs=-1)
```

In [72]:

```
dfdb = dfa
dfdb['AVG-W2V Clus Label'] = model.labels_
dfdb.head(2)
```

Out[72]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	7

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	10

In [73]:

```
dfdb.groupby(['AVG-W2V Clus Label'])['Id'].count()
```

Out[73]:

```
AVG-W2V Clus Label
0    4985
Name: Id, dtype: int64
```

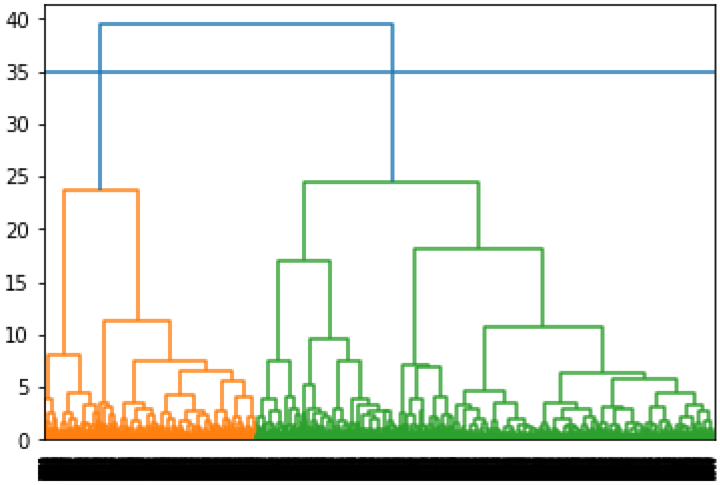
Clustering Hierarchical

In [74]:

```
import scipy
from scipy.cluster import hierarchy
dendro=hierarchy.dendrogram(hierarchy.linkage(sent_vectors,method='ward'))
plt.axhline(y=35)# cut at 30 to get 5 clusters
```

Out[74]:

```
<matplotlib.lines.Line2D at 0x1e1810989d0>
```



In [75]:

```
from sklearn.cluster import AgglomerativeClustering

cluster = AgglomerativeClustering(n_clusters=5, affinity='euclidean', linkage='ward')
Agg=cluster.fit_predict(sent_vectors)
```

In [76]:

```
# Giving Labels/assigning a cluster to each point/text
aggdfa = dfdb
aggdfa['AVG-W2V Clus Label'] = cluster.labels_
aggdfa.head(2)
```

Out[76]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominat
1244	1245.0	B00002Z754	A29Z5PI9BW2PU3	Robbie	7.0	7
1243	1244.0	B00002Z754	A3B8RCEI0FXFI6	B G Chase	10.0	10

In [77]:

```
# How many points belong to each cluster ->
aggdfa.groupby(['AVG-W2V Clus Label'])['Text'].count()
```

Out[77]:

```
AVG-W2V Clus Label
0    1065
1    1169
2    1933
3     414
4     404
Name: Text, dtype: int64
```

In [78]:

```
# Reading a review which belong to each group.
for i in range(5):
    print("2 reviews of assigned to cluster ", i)
    print("-" * 70)
    print(aggdfa.iloc[aggdfa.groupby(['AVG-W2V Clus Label']).groups[i][0]]['Text'])
    print('\n')
    print(aggdfa.iloc[aggdfa.groupby(['AVG-W2V Clus Label']).groups[i][1]]['Text'])
    print('\n')
    print("_" * 70)
```

```
2 reviews of assigned to cluster  0
----------------------------------------------------------------------
They are OK, but my husband only eats them when we are in car. We keep them there for a
celiac snack. (Hard to find when roaming around.) He won't eat them otherwise. They are
definitely raspberry, but are kind of dry.<br />  We have never tried any other snack ba
rs, so in all honesty they may be the norm. I don't know.<br />  We DO like Enjoy Life's
other products.<br />  They are GF, have no MSG or related products or aspartates.


the pop chips are really incredible. They are very flavorful and crispy. My Favorite is
the BBQ, but that is just me, they are all good. I would highly recommend these to anyon
e who is watching their weight of just for overall better health.


______________________________________________________________________
2 reviews of assigned to cluster  1
----------------------------------------------------------------------
I love this stuff.  I put it over chicken and even steaks.  It's very sweet, and just a
```

```
bit spicy.  Just a little goes a long way.


This stuff makes great pancakes and shortcake that I am actually allowed to eat!<br />My
doctor tells me I'm celiac- this was three years ago now. I just hope I can continue to
find this Bisquick!



______________________________________________________________________
2 reviews of assigned to cluster  2
----------------------------------------------------------------------
I have tried many bread machine mixes and most have been okay but this one rates an "exc
ellent" because not only is it a very good basic loaf, it also works very well with addi
tives, either prepared entirely in the bread machine, or (after the second rise) removed
from the machine, shaped and baked on a stone in the oven.<br />Stretched and flattened
with a filling of either savory or sweet ingredients, then rolled and set to rise and th
en bake in a loaf pan, it produced excellent herb and onion bread, cinnamon raisin bread
and a brown butter and seed bread.<br />I buy it via the subscribe-and-save plan as I us
e it often.


I have a 2 year old Portuguese Water Dog who always seemed to have a sensitive stomach,
and a 15 year old Shepherd X dog who was beginning to lose weight, she slept almost all
the time, and she was getting very fussy about what she'd eat. (I believe that with a do
gs sense of smell they KNOW exactly what is in their food; however they have no choice b
ut to eat what we feed them...just Google what is in most commercial dog foods, and yo
u'll see why your dog may not be thrilled to eat it up...if a dog can detect cancer in a
person, they can detect all sorts of other things that aren't supposed to be eaten.) Any
way...I used to feed Purina lamb and rice to my dogs, but I was finding my young Porty w
as having trouble with it, and my 15 yr old would eat a few bits and then leave her dis
h. So, I started making their food, and came up with a wonderfully nutritious and tasty
recipe. The dogs loved my homemade dog food, and my 15 yr old began putting on weight ag
ain, but I didn't always have time to make their food. And, homemade dog food is kind of
hard on the wallet. So, I started to research all the natural dog foods out there, compa
re costs etc. I arrived on Harmony Farms, and guess what...my dogs LOVE it just as much
as they love my home made dog food. My young Porty no longer has stomach troubles and sh
e looks great; my 15 yr old finishes her meals and looks for more. Both have shiny coats
and energy. No one can believe my old girl is 15 years old. I have told my friends about
Harmony Farms dog food, and when they've switched over, they've reported similar positiv
e results. My brother-in-law's dog had always been a very fussy eater, and had skin prob
lems with a thinning coat. When they started to feed her Harmony Farms, she became an ea
ger eater, more happy and outgoing, and her skin and coat condition has completely clear
ed up (...I'm thinking she had an allergy to whatever was in her other food, which I thi
nk was Iams). Other friends have commented on how happy their dogs are at meal time no
w...and they had all been feeding their dogs premium, top of the line dogs foods AND pay
ing top of the line prices. Harmony Farms products are very reasonably priced. Not the c
heapest dog food on the shelf, but, in my opinion, you are buying the best quality food
on the market, so it's an incredibly GOOD DEAL! It may also mean you make fewer trips to
the veterinarian's office. I and my friends are so happy we have found this food for our
dogs. :)



______________________________________________________________________
2 reviews of assigned to cluster  3
----------------------------------------------------------------------
Cutting sugar out of our diet,we went to splenda. Then finding that it was also unhealth
y and baking with it made foods flat and dry. We went to stevia no calorie powder, after
much research proved that stevia has been used for hundreds of years with no bad health
effects. The trick to using stevia is to use tiny amounts and then taste it and add a bi
t more at a time to suit your taste. I also found that using a stevia that has an 80% ra
tio is the best tasting and this is the best that I have found. Because of using VERY SM
ALL amounts this jar lasts for months. This is also the best price ANYWHERE! Swanson is
a trusted vendor as is Amazon.
```

```
these chips tatse great, and the serving size is good. we especially liked the variety p
ack with six flavors


___________________________________________________________________
2 reviews of assigned to cluster  4
-------------------------------------------------------------------
I love these treats for my two german shepherds. I think these treats are great for even
large size dogs. My dogs absolutely go crazy for them, and they are only 3 calories per
treat, perfect for training sessions. I also love the texture of them... you can put the
m in your pocket without a ton of flaking or gross smell. This product also has a simpli
fied ingredient list... opposed to so many other dog treats out there. You can pronounce
all the ingredients in the list and you know what they are getting.


Stonewall Kitchen products are a big favorite at our house.  This pancake mix makes the
best pancakes you will ever eat.  We add fresh blueberries to them, and top them with th
e Stonewall Blueberry Syrup, which you can buy on their website.  They make wonderful ja
ms and jellies too, as well as sauces, and flavored syrups.  Try this pancake mix thoug
h, you won't be disappointed!


___________________________________________________________________
```

Conclusion

Kmeans for bag of words and TFIDF

1. By using Elbow method, we generated optimal 10 clusters for both the bag of words and tfidf techniques
2. In both the cases, one cluster accounts around 6000 reviews which is large chunk from 10k reviews and rest are distributed unevenly
3. we can ignore 2 clusters or keep 2 clusters depending upon the business goal for bag of words generation as both contain only 1 review

Final Observations:

FOR TFIDF K means is best for identification than K MEANS for BOW, all the clusters are clearly refelcting they were grouped based on the categories/products. However, K means did best on the cluster centers top terms but however when we caopare reviews , few places it is not correalting.

DBSCAN is very poorly performining on the 10k columns as it is grouping all reviews in one cluster

Hierarchical, for BOW and TFIDF, we cannot identify the clusters and not divded unevenly, but for avg word to vectors all are grouped and divided evenly. It is very difficult to identify the type of reveiws based on Hirarchial formation.

```
In [ ]:
```