In [0]:
```
import warnings
warnings.filterwarnings("ignore")

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

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

import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

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

```
from pandas_ml import ConfusionMatrix
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, precision_score, recall_score, make_scorer, f1_score, classification_report

from sklearn.model_selection import train_test_split
from sklearn.cross_validation import cross_val_score
from collections import Counter
from sklearn import cross_validation
from prettytable import PrettyTable
from sklearn.model_selection import GridSearchCV,TimeSeriesSplit
from sklearn.model_selection import RandomizedSearchCV
from sklearn.utils import resample
from sklearn.cluster import KMeans
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import DBSCAN
from sklearn.metrics import silhouette_score
from wordcloud import WordCloud
from sklearn.neighbors import NearestNeighbors
import plotly.plotly as py
from plotly.graph_objs import *
import plotly
plotly.tools.set_credentials_file(username='MyUsername', api_key='MyAPIkey')
import plotly.graph_objs as go
```

Utility Functions

In [0]:
```
#We create a few utility functions whose use is described below


def plot_df_wordcloud(data,labels_list):
    #This function plots the wordcloud for DBScan. These will be called in the following functions.
    print('\n')
    print("Creating a dataframe with Reviews, Cleaned text and Clusters...")
    lst_reviews = list(data['Text'].values) #Saving all reviews into a list
    lst_cleanedtext = list(data['CleanedText'].values) #Saving all Cleaned text into a list
    amz = {'Reviews': lst_reviews,'CleanedText':lst_cleanedtext ,'Clusters': labels_list} #Creating a dictionary with Reviews, Cleaned text and clusters
    df = pd.DataFrame(amz, index=[labels_list], columns=['Reviews','CleanedText', 'Clusters']) #Creating a dataframe of the above dictionary
    print("Dataframe is created!")
    print('\n')

    print("The number of reviews in each cluster is:")
    print(df['Clusters'].value_counts()) #Outputs the number of reviews in each cluster
    print('\n')
    print('*'*70)

    for i in range(min(labels_list),max(labels_list)+1): #Iterates through K and prints wordcloud and reviews for all the clusters
        print('\n')
        print('\n')
        print("*"*40,'Cluster ',i,"*"*40)
        words=[] #Create an empty list to store words
        for sent in df['CleanedText'][df['Clusters']==i].values: #Splits sentences into words and stores it in a list
            words.append([a for a in sent.split()]) #Appends the split words to above created list

        word_list = [item for sublist in words for item in sublist] #Creates a new list of words by flattening the above list since it is a nested list
        print('\n')
        print('Plot of Word Cloud')
        wordcloud1 = WordCloud().generate(" ".join(word_list)) #Initiate wordcloud for the list of words
        plt.figure() #Plots the wordcloud
        plt.imshow(wordcloud1)
        plt.axis("off")
        plt.show()

        count = df['Reviews'][df['Clusters']==i].count()
        if count//2>=9:#Prints random 5 reviews if the number of reviwes in that cluster is more than 100
            print('\n')
            print('\n')
            print('Printing 4 random reviews from cluster',i)
            pd.set_option('display.max_colwidth', -1) #Displays full reviews
            print(df['Reviews'][df['Clusters']==i][count//2:count//2+5])
            print('\n')
            print('\n')
        else: #Prints random reviews if the number of reviwes in that cluster is between 10 and 100
            print('\n')
            print('\n')
            print('Printing random reviews from cluster',i)
            pd.set_option('display.max_colwidth', -1) #Displays reviews in full length
            print(df['Reviews'][df['Clusters']==i][count//2:])
            print('\n')
            print('\n')

def plot_plotly(data,std_data,dist,ind): #This function plots the graph for indices vs distances using plotly. Plotly is chosen so that we can pin point the exact X and Y axis co-ord
    trace = go.Scatter(x = ind, y = dist) #Initialize trace with X as indices and Y as distances
    layout = go.Layout(title='Indicies vs Distance plot',xaxis=dict(title='Indicies'),yaxis=dict(title='Distances')) #Setting the title, x label and y label
    fig = go.Figure(data=[trace], layout=layout)
    return py.iplot(fig, filename='basic-line') #Plots the graph


def dbscan(data,std_data,eps,min_pts): #Performs Dbscan on the Eps chosen form the above graph, min_pts is the number of min points chosen which is 2*d in this case where d is dimens
    clf = DBSCAN(eps=eps, min_samples=min_pts) #Initialize DBSCAN
    clf.fit(std_data)
    labels_list_check = clf.labels_.tolist() #Gets the labels and converts it into a list
    print('\n')
    df = pd.DataFrame({'labels':labels_list_check}) #Create a dataframe with labels list to see how many points are in each cluster
    print('For Eps =',eps,'the number of points in each cluster are:')
    print('\n')
    print(df['labels'].value_counts()) #Prints the number of points in each cluster for the given Eps value
    print('\n')
    print('\n')
    print('='*20,'Plotting wordcloud and reviews for Eps =',eps)
    print('\n')
    plot_df_wordcloud(data,labels_list_check) #Plot wordcloud and print a few random reviews
```

Avg WordVec

In [0]:
```
import pandas as pd
final = pd.read_csv("drive/grouped_data_200.csv")
p = final.groupby('Score')
pos = p.get_group('Positive') #Gets the groups with Positive score
neg = p.get_group('Negative') #Gets the groups with Negative score
pos_2000 = pos.sample(30000) #Gets 30000 reviews of positive and 20000 negative scores
neg_2000 = neg.sample(20000)
grouped_data = pd.concat([pos_2000, neg_2000], ignore_index = True) #This data now contains positive and negative data in order.
grouped_data.dropna(inplace = True) #Drops rows with Nan
grouped_data.reset_index(inplace=True) #Replaces missing indexes
grouped_data.drop(['Unnamed: 0', 'Unnamed: 0.1'], axis=1, inplace=True)
grouped_data = grouped_data.sort_values('Time', axis=0, ascending=True, kind='quicksort')
scores=grouped_data['Score']
print("The shape of grouped data is {}".format(grouped_data.shape))
```

```
The shape of grouped data is (49997, 12)
```

Observations: A csv file is imported which consists of 200000 data points. These data points are already sorted on the basis of time. We create a new dataframe with over 50000 data points.

In [0]:
```
list_of_sent=[]
for sent in grouped_data['CleanedText'].values: #Splits sentences into words and stores it in a list
    list_of_sent.append(sent.split())
print(grouped_data['CleanedText'].values[9])
print("**************************************************************")
print(list_of_sent[9])
```

```
tri mani jasmin blend found dragon pearl blend basic regular jasmin novelti leav roll tight littl dri unfurl infus far tell noth enchanc flavor brew nice jasmin blend noth would re
commend jasmin lover
**************************************************************
['tri', 'mani', 'jasmin', 'blend', 'found', 'dragon', 'pearl', 'blend', 'basic', 'regular', 'jasmin', 'novelti', 'leav', 'roll', 'tight', 'littl', 'dri', 'unfurl', 'infus', 'far',
'tell', 'noth', 'enchanc', 'flavor', 'brew', 'nice', 'jasmin', 'blend', 'noth', 'would', 'recommend', 'jasmin', 'lover']
```

Observations: Get a list of all the words in data.

In [0]:
```
w2v_train=Word2Vec(list_of_sent,min_count=5,size=200, workers=4) #Initialises the Word2Vec model with words occuring more than 5 times.

w2v_train_words = list(w2v_train.wv.vocab) #This gives a dictionary of words which tells about the uniqueness of a word among other things.
print("number of words that occured minimum 5 times ",len(w2v_train_words))
print("sample words ", w2v_train_words[298:315])
```

```
number of words that occured minimum 5 times  9521
sample words  ['franci', 'chicken', 'liver', 'buy', 'pleas', 'amaz', 'flavor', 'chees', 'food', 'triangl', 'throughout', 'childhood', 'bold', 'volum', 'histori', 'age', 'virtual']
```

Observations: Train the word2vec model on the obtained list of words.

In [0]:
```
sent_vectors = [];
sent_list = []
for sent in grouped_data['CleanedText'].values:
    sent_list.append(sent.split())
for sent in sent_list: # For a sentence in the previously created list of sentences
    sent_vec = np.zeros(200) # As word vectors are of zero length, returns an array of size 50 filled with zeros
    i = 0; # Number of words with a valid vector in the sentence/review
    for word in sent: # For each word in a review/sentence
        if word in w2v_train_words:
            vec = w2v_train.wv[word] #Gets the corresponding vector for the word
            sent_vec += vec
            i += 1
    if i != 0:
        sent_vec /= i
    sent_vectors.append(sent_vec)
print(len(sent_vectors))
print(len(sent_vectors[0]))
```

```
49997
200
```

Observations: Gets the sentence vectors for data.

In [0]:
```
data = StandardScaler(with_mean=False).fit_transform(sent_vectors)
```

Observations: Standardize the data.

In [0]:
```
min_pts = len(data[0])*2 #Min_pts is taken as 2*d where d is dimensionality
neib = NearestNeighbors(n_neighbors=min_pts).fit(data) #Perform nearest neighbors on the standardized data
dist, ind = neib.kneighbors() #Get distance and indices matrix
dist[:,-1].sort() #Sort the last column of the distance matrix
ind[:,0].sort() #Sort the first column of the indices matrix
```

Observations: Get the min point value and nearest neighbours distances by taking min points as n_neighbors. Obtain distance and indices matrix.

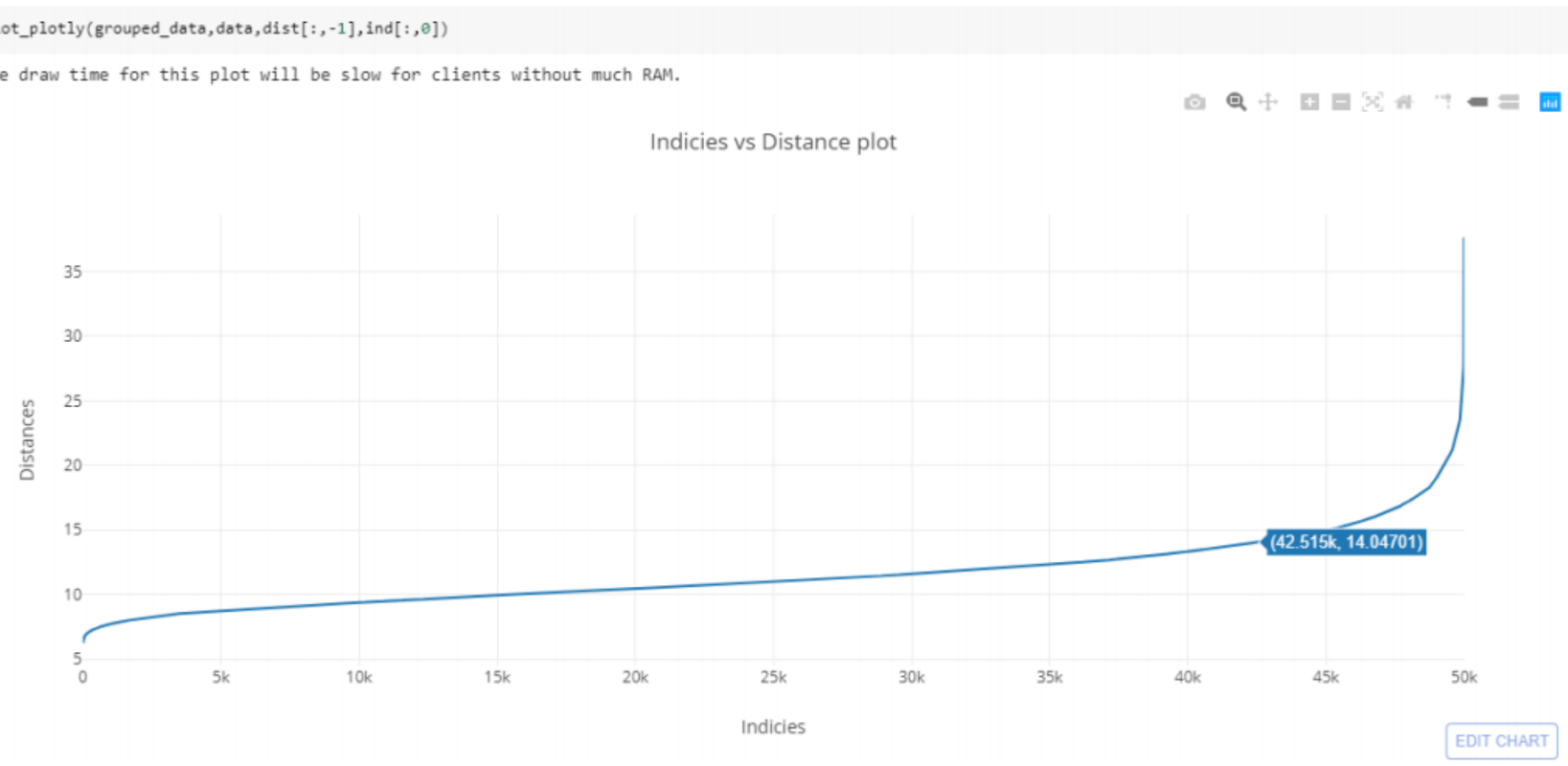
In [0]:
```
plot_plotly(grouped_data,data,dist[:,-1],ind[:,0])
```

```
The draw time for this plot will be slow for clients without much RAM.
```

Out[40]:

[ ]
```
plot_plotly(grouped_data,data,dist[:,-1],ind[:,0])
```

```
The draw time for this plot will be slow for clients without much RAM.
```



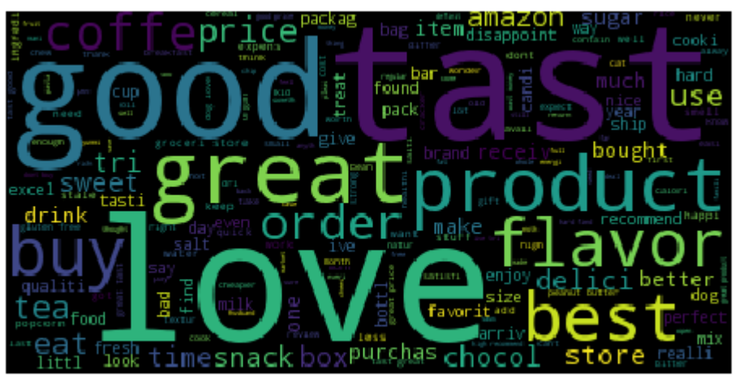
Observations:

1)Plotly plot is plotted and it is observed that Eps value around 14 will be a suitable value.

2)Since the plot is not diplayed for some reason when the ipynb file is moved, I have attached an image of the interactive plot plotted using plotly.

In [0]:
```
dbscan(grouped_data,data,13.00,min_pts)
```

```
For Eps = 13.0 the number of points in each cluster are:


 0    47463
-1    2534
Name: labels, dtype: int64



==================== Plotting wordcloud and reviews for Eps = 13.0



Creating a dataframe with Reviews, Cleaned text and Clusters...
Dataframe is created!


The number of reviews in each cluster is:
 0    47463
-1    2534
Name: Clusters, dtype: int64

**********************************************************************



************************************ Cluster  -1 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster -1
-1    As I'm always looking for a DEAL on Cranaple by OS there is no DEAL here at $9.54 a bottle. When my local store has it around $3.58 a 64.oz bottle Thanks but No Thanks
-1    Been using for about 4 months now with my 7 year old husky-shepard mix.  Seems to help her move a little better.

-1    These crackers are great.  Unfortunately, every box I've bought is about 50% broken.  And that's from several different stores.

-1    These have the perfect consistency for caramel apples. They melt easily and harden perfectly. Some caramels are too soft and almost melt off the apples but these are great!

-1    This coffee is very weak.  I am very disappointed with this coffee.  Not worth the price.  i would not recommend this coffee...  I signed up for delivery every few weeks. I w
ill cancel my subscription.
Name: Reviews, dtype: object



************************************ Cluster  0 ************************************

Plot of Word Cloud
```


```
Printing 4 random reviews from cluster 0
0    Been drinkin gormet coffee many years now, bunch of different brands and roasters. It's hard to beat the Coffee Fool for quality and freshness, but I've come back to this Jeri
mias' pick for the quality and the savings I get here at Amazon. Its a great value and superb freshness--you'll be very pleased!
0    Iopened the pkg and all but 25 waffle bowls survived. Iam left with an open house this weekend and not enough bowls!!!

0    This stuff tastes amazing though and everyone I share with loves it! Although I am sure the seller did not mean to mislead buyers, each individual limoncho container is .4 oz
and not 4 oz like written in the description.
0    This was awful coffee. I gave the other two bags away.  I asked them to tell me their opinion.  They hated it too.<br /><br />If they were giving this away for free, I would t
ake a pass.
0    I love that it is American made from renewable sources (Birch trees), and the taste is great (better than splenda or stevia)--absolutely indistinguishable from sugar. However,
a couple of days of use had my wife cramping and a bit distressed in the nether regions. Now the canister stands a bit sadly unused (hey, don't look at me--I like sugar!).
Name: Reviews, dtype: object
```

Observations:

1. When Eps = 13.00, the number of elements in cluster 0 are 47463 and points in cluster -1 (or noise points) are 2534.

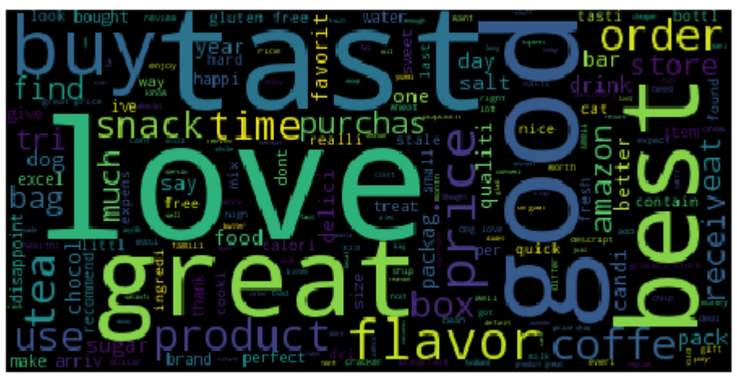
2. Cluster -1 reviews seem to be about how the customers liked the product due to a frequent occurance of the words 'great', 'love', 'good', 'best'.

3. Cluster 0 reviews talked about how the customers used the product and how good the flavour was.

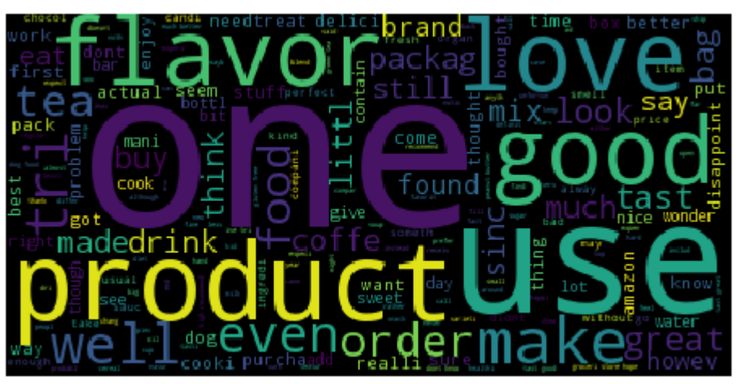
4. The number of points classified as noise increased with decrease in Eps value.

In [0]:
```
dbscan(grouped_data,data,14.06,min_pts)
```

```
For Eps = 14.06 the number of points in each cluster are:


 0    48970
-1    1027
Name: labels, dtype: int64



==================== Plotting wordcloud and reviews for Eps = 14.06



Creating a dataframe with Reviews, Cleaned text and Clusters...
Dataframe is created!


The number of reviews in each cluster is:
 0    48970
-1    1027
Name: Clusters, dtype: int64


**********************************************************************

************************************ Cluster  -1 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster -1
-1    my dogs tell me that they are very good and good for their health..

-1    I purchased this coffee on 11/4, and as 12/14 I have not received the product.  I would not recommend this seller.
-1    I have been trying to get hoodys peanut crunch and cashews for over a year now.  how do you buy them?

-1    The ingredient list for this item lists cocoa butter but the actual item contains ZERO cocoa butter, in fact the ingredient list on this listing is completely different from the package.  This is not white chocolate, avoid!!!!
-1    My cat loves these treats. I wish there was a hairball treat with salmon flav. cause my cat is alergic to chicken food/flav.

Name: Reviews, dtype: object



************************************ Cluster  0 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster 0
0    Same exact product is available at Sams Club for $8.77. Wow! My Prime Membership is really paying off on this product...



0    Purchasing Lundberg Jubilee, Gourmet Blend of Whole Grain Brown Rice, 16-Ounce Units (Pack of 6) was a great decision. This rice has a great nutty taste. The other benefit is that since it is whole grain it has a low glycemic index. The price is much better than I receive at the local grocers, when I can find this rice.<br />I love it and I highly recommend it. You may never go back to white rice!!<a href="http://www.amazon.com/gp/product/B000G82L5I">Lundberg Jubilee, Gourmet Blend of Whole Grain Brown Rice, 16-Ounce Units (Pack of 6)</a>
0    I'm a nut butter addict, so I was excited to try a nut butter that was finally low-cal. But, of course, I was just disappointed. It says it's "all natural," but it leaves a really distinct, gross, artificial after-taste. It doesn't really taste like peanut butter- and it just does not come close to any nut butter. I personally think it's awful. Plus it's expensive! Forget about the calories and eat regular nut butter- it definitely tastes better!

0    I'm not a big granola bar guy, but these things are crazy delicious!!! At about 200 calories it's the perfect size for a pre- or post- run/exercise snack. I also appreciate that Nature Valley uses truly whole grains instead of the fake whole wheat flour that you usually find in "whole grain" products. I'm about to dig into one as soon as I finish this review, mmm mmmmmmm!!!

0    These taste just like banana runts (my favorite)!  Great treat for tropical theme birthday parties, pinata filling, or stocking stuffers for people like me who love these banana flavored candies!


Name: Reviews, dtype: object
```

Observations:

1. When Eps = 14.06, the number of elements in cluster 0 are 48970 and points in cluster -1 (or noise points) are 1027.

2. Cluster -1 reviews seem to be about how the customers liked the product due to a frequent occurance of the words 'great', 'love', 'good', 'best'.

3. Cluster 0 reviews talked about how the customers used the product and how good the flavour was.

In [0]:
```
dbscan(grouped_data,data,15.00,min_pts)
```

```
For Eps = 15.0 the number of points in each cluster are:


 0    49556
-1    441
Name: labels, dtype: int64



==================== Plotting wordcloud and reviews for Eps = 15.0



Creating a dataframe with Reviews, Cleaned text and Clusters...
Dataframe is created!


The number of reviews in each cluster is:
 0    49556
-1    441
Name: Clusters, dtype: int64


**********************************************************************

************************************ Cluster  -1 ************************************

Plot of Word Cloud
```


```
Printing 4 random reviews from cluster -1
-1    rated as low sodium and very much true.  Very high calorie with no flavor.  I will snack on them for a low carb diet but add salt and pepper
-1    a little dry...way too much chocolate coating.  I had to slice it off the sides
-1    what to say?<br />they arrived and the dogs loved them and all is well.
-1    Quality of product is excellent, cost is excellent and delivery is superb.  Very satisfied.
-1    so so.there was great reviews on product only made it once will try it again
Name: Reviews, dtype: object



************************************ Cluster  0 ************************************

Plot of Word Cloud
```


```
Printing 4 random reviews from cluster 0
0    When you taste these you only get the taste of salt. Either the package I got came from the batch Amazon had in storage for long period of time or when they packaged these som
eone overdozed the salt. I have these only because  the Roland Kalamata Pitted (whole) olives have been out of stock for couple month. As result had to dispose to trash Rolan Kalam
ata Halves.

0    As a chocolate lover, I enthusiastically looked forward to trying the new Fiber One brownies. Unfortunately, I was quite disappointed.  The texture is not quite up to that of
a real brownie, lacking that treat's trademark chewiness and moisture. The taste was far too sweet, and lacked in chocolate flavor. I don't know why many manufacturers, and even so
me private bakeries, insist that sweet is interchangeable with chocolate. It isn't. The low calorie count and high fiber are the best things about these bars, but personally, I wou
ld rather do without than accept this as a brownie substitute. If you like sweet treats, however, this might be a good alternate snack for you.
0    These are my guilty pleasure, and a household favorite!! They are hard to find in my area, and its great to be able to get them on Amazon!<br /><br />These are just as good as
the girl scout version, and available much cheaper!!<br /><br />WOOT WOOT!! YUM YUM!!


0    A co-worker used to use this product.  The spray remains in the air and can be detected as a heavy scent several yards away (huge warehouse-type building), and caused asthmati
c reactions after use in normal-sized rooms.  He kindly shared the product information with me AND ceased using it out of concern for those affected.  Many individuals later expres
sed privately to me that the odor was repulsive, even if it had not been over-strong (it is apparently impossible to apply the spray version of this product "lightly").  So if you
are trying to be a chick magnet, DON'T buy this product.  "Your cologne is making me gag and nearly pass out" isn't the conversation you want to start with "the ladies."
0    I took this box of Cheesy Skillets to a cabin getaway and had it for dinner one night while we were there. That was not a good move. The nearest restaurant was 25 miles away.
And guess where we went for dinner that night? The skillet smelled up the whole cabin. We had to leave the doors open and let all the bugs in to get the stench out. So we had a sme
lly cabin and tons of mosquitos because we cooked something we couldn't even eat. That's right, we threw it out.<br /><br />That was my experience with this product. I also had the
Chicken and Broccoli flavor which I liked, so it's not a blanket statement for the whole product line - just this flavor.
Name: Reviews, dtype: object
```

Observations:

1. When Eps = 15.00, the number of elements in cluster 0 are 49556 and points in cluster -1 (or noise points) are 441.

2. Cluster -1 reviews seem to be about how the customers liked the product due to a frequent occurance of the words 'great', 'love', 'good', 'best'.

3. Cluster 0 reviews talked about how the customers used the product and how good the flavour was.

4. The number of points classified as noise decreased with increase in Eps value.

TFIDF WordVec

In [0]:
```
import pandas as pd
final = pd.read_csv("drive/grouped_data_200.csv")
p = final.groupby('Score')
pos = p.get_group('Positive') #Gets the groups with Positive score
neg = p.get_group('Negative') #Gets the groups with Negative score
pos_2000 = pos.sample(30000) #Gets 30000 reviews of positive and 20000 negative scores
neg_2000 = neg.sample(20000)
grouped_data = pd.concat([pos_2000, neg_2000], ignore_index = True) #This data now contains positive and negative data in order.
grouped_data.dropna(inplace = True) #Drops rows with Nan
grouped_data.reset_index(inplace=True) #Replaces missing indexes
grouped_data.drop(['Unnamed: 0', 'Unnamed: 0.1','Score'], axis=1, inplace=True)
print("The shape of grouped data is {}".format(grouped_data.shape))
```

```
The shape of grouped data is (49999, 11)
```

In [0]:
```
list_of_sent=[]
for sent in grouped_data['CleanedText'].values: #Splits sentences into words and stores it in a list
    list_of_sent.append(sent.split())
print(grouped_data['CleanedText'].values[9])
print("*****************************************************************")
print(list_of_sent[9])
```

```
sauc hot sake hot get heat flavor one sauc sauc kick sill lot wonder flavor great kind meat use beef chicken pork like scrambl egg despit flavor lower sugar sodium mani sauc less g
uilti guilti pleasur watch diet cant wrong sauc
*****************************************************************
['sauc', 'hot', 'sake', 'hot', 'get', 'heat', 'flavor', 'one', 'sauc', 'sauc', 'kick', 'sill', 'lot', 'wonder', 'flavor', 'great', 'kind', 'meat', 'use', 'beef', 'chicken', 'pork',
'like', 'scrambl', 'egg', 'despit', 'flavor', 'lower', 'sugar', 'sodium', 'mani', 'sauc', 'less', 'guilti', 'guilti', 'pleasur', 'watch', 'diet', 'cant', 'wrong', 'sauc']
```

Observations: Get a list of all the words in data.

In [0]:
```
w2v_train=Word2Vec(list_of_sent,min_count=5,size=200, workers=4) #Initialises the Word2Vec model with words occuring more than 5 times.

w2v_train_words = list(w2v_train.wv.vocab) #This gives a dictionary of words which tells about the uniqueness of a word among other things.
print("number of words that occured minimum 5 times ",len(w2v_train_words))
print("sample words ", w2v_train_words[298:315])
```

```
number of words that occured minimum 5 times  9549
sample words  ['krave', 'crunchi', 'outsid', 'middl', 'pleas', 'also', 'white', 'togeth', 'hope', 'stay', 'around', 'yumm', 'recov', 'addict', 'suggest', 'replac', 'real']
```

Observations: Word2Vec model is built. We can see the number of times a word occured minimum 5 times.

In [0]:
```
tf_idf_vect = TfidfVectorizer(ngram_range=(1,2))
vocab_tf_idf = tf_idf_vect.fit_transform(grouped_data['CleanedText'].values)
tfidf_feat = tf_idf_vect.get_feature_names()
dictionary = dict(zip(tfidf_feat, list(tf_idf_vect.idf_)))
```

Observations: Train the TFIDF vectorizations.

In [0]:
```
sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
sent_list = []
for sent in grouped_data['CleanedText'].values:
    sent_list.append(sent.split())
for sent in sent_list: # for each review/sentence
    sent_vec = np.zeros(200) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_train_words:
            try:
                vec = w2v_train.wv[word] # obtain the tf_idfidf of a word in a sentence/review
                tf_idf  = dictionary[word]*sent.count(word)
                sent_vec += (vec * tf_idf)
                weight_sum += tf_idf
            except:
                pass
    if weight_sum != 0:
        sent_vec /= weight_sum
    sent_vectors.append(sent_vec)
    row += 1
print(len(sent_vectors))
print(len(sent_vectors[0]))
```

```
49999
200
```

Observations: Gets the sentence vectors for data.

In [0]:
```
data = StandardScaler(with_mean=False).fit_transform(sent_vectors)
```

Observations: Standardize the data.

In [0]:
```
min_pts = len(data[0])*2 #Min_pts is taken as 2*d where d is dimensionality
neib = NearestNeighbors(n_neighbors=min_pts).fit(data) #Perform nearest neighbors on the standardized data
dist, ind = neib.kneighbors() #Get distance and indices matrix
dist[:,-1].sort() #Sort the last column of the distance matrix
ind[:,0].sort() #Sort the first column of the indices matrix
```

Observations: Get the min point value and nearest neighbours distances by taking min points as n_neighbors. Obtain distance and indices matrix.

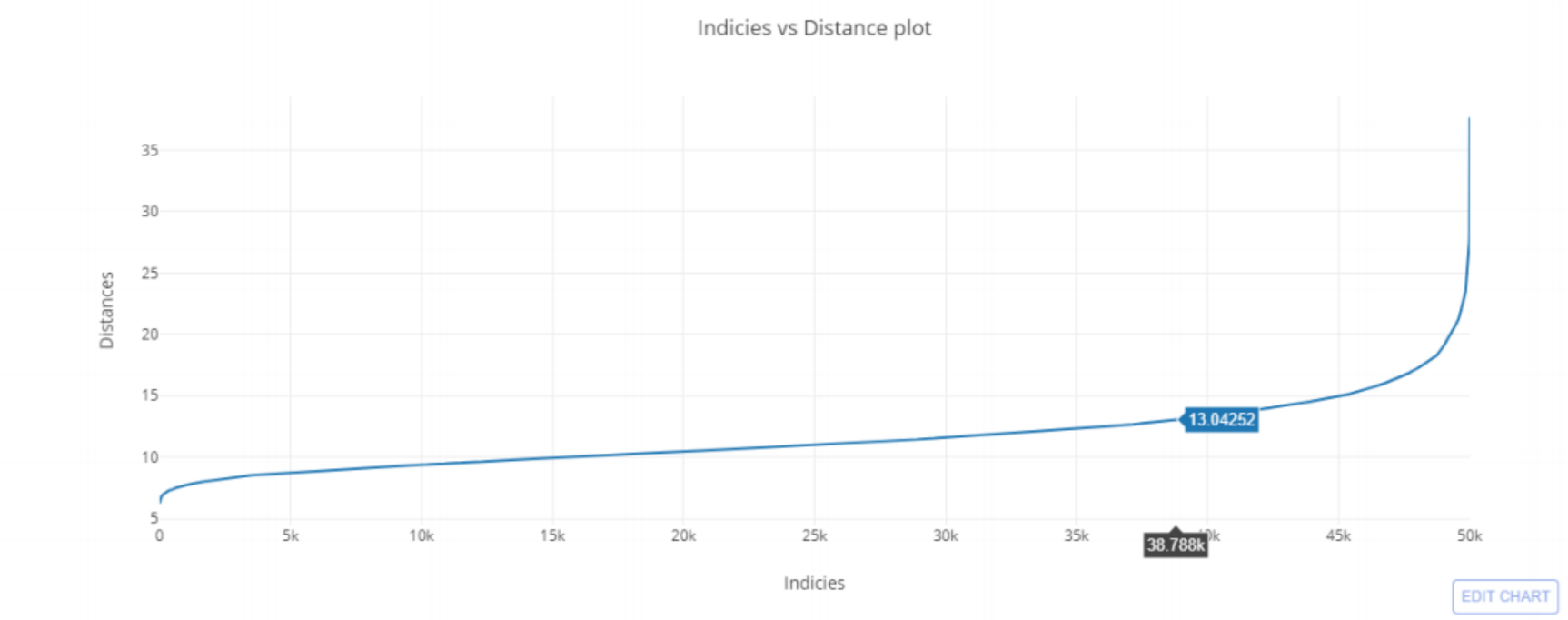
In [0]:
```
plot_plotly(grouped_data,data,dist[:,-1],ind[:,0])
```

```
The draw time for this plot will be slow for clients without much RAM.
```

Out[17]:

[ ] plot_plotly(grouped_data,data,dist[:,-1],ind[:,0])

The draw time for this plot will be slow for clients without much RAM.



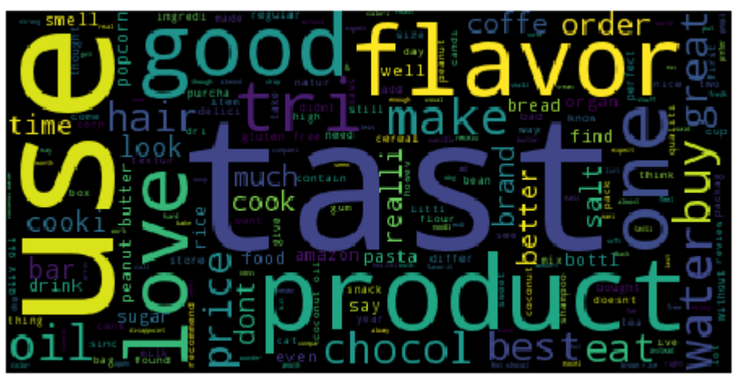
Observations:

1) Plotly plot is plotted and it is observed that Eps value around 13 will be a suitable value.

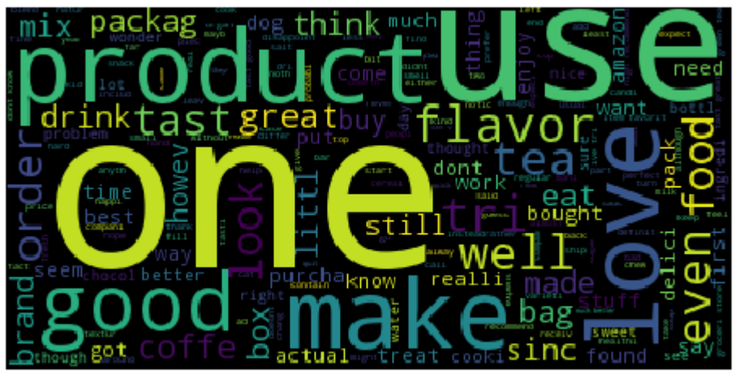
2) Since the plot is not displayed when the ipynb file is moved for some reason, I have attached an image of the interactive graph plotted using plotly.

In [0]:
```
dbscan(grouped_data,data,12.05,min_pts)
```

```
For Eps = 12.05 the number of points in each cluster are:


 0    48156
-1    1843
Name: labels, dtype: int64



==================== Plotting wordcloud and reviews for Eps = 12.05



Creating a dataframe with Reviews, Cleaned text and Clusters...
Dataframe is created!


The number of reviews in each cluster is:
 0    48156
-1    1843
Name: Clusters, dtype: int64


**********************************************************************

************************************ Cluster  -1 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster -1
-1    Works great. So much easier to use than the one you buy for the machine. Less mess - just throw it away!

-1    In the category it's in, I rate this 5 stars.  Sure, there are higher quality cocoas and chocolates out there, but for the price, and the ease of use, this rates high in my 'recipe' book.  This does have a bit more sweetness than what I would normally drink, but adding semi-sweet chocolate chips to the cup cuts down on that aspect.<br /><br />I use the Back to Basic Hot Cocoa/Latte maker and it does a wonderful job of mixing all lumps out (of any brand) and heating my hot chocolate or hot cocoa fast and thoroughly.  The difference is that hot chocolate is made by melting chocolate bars or morsels, it has a higher cocoa butter content than your cocoas.  I sometimes use a mix of both cocoa powder and a handful of chocolate, makes for a decadent hot drink.  I also use a splash of heavy cream with whole milk to make my hot cocoa, it makes it extra rich and creamy.  Don't forget to add a few drops of real vanilla extract to add a sublime nuance to your cocoa and chocolate drinks.  I also enjoy mixing a few brands of cocoa powders and chocolate bars to make outrageous hot cocoa/hot chocolate drinks.  I haven't tried the frozen hot chocolate by a famous restaurant (S3) yet, but this brand, along with others I purchased on Amazon will all be thrown into the mix :)<br /><br />This mixes smoothly, is slightly sweeter than others I use, has a perfect color, and the powder is very finely milled making it excellent for sprinkling on top of cappuccino or whipped cream.  Not bitter, just a good chocolate taste that comes through.  Can't beat it for the connoisseur hot cocoa drinker to add to their arsenal.
-1    My daughter is sensitive to gluten, and fortunately we love most of the gluten free pastas. This corn pasta has a nice springy texture. My daughter and her friends love it, and it works well for mac and cheese.

-1    I looked at all grocery stores in my area for some GF soy sauce. Glad to have found it online.

-1    I've been drinking this chai for years and  no other chai product compares. Well worth the money.



Name: Reviews, dtype: object



************************************ Cluster  0 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster 0
0    H. Mitchell, I suspect you're joking about the granola providing you only 2 servings, but the back of the package suggests 3/4 cup per serving for a total of 6 servings.  I lo
ve the taste and texture of the ingredients, the delicate flavor, the gratification I get from knowing it will keep me satisfied until lunchtime.  Good job, whoever designed the gr
anola.

0    For those hazelnut lovers out there, you MUST TRY THIS!!!  I had no problems with the product as far as separation.  A good thorough mixing remedies this and I put some that w
as a little more solidified in the microwave for a few seconds to soften and then it mixed up just fine.<br /><br />I used the recipe on the label to make the filling for a brownie
torte I made over the holidays.  People RAVED over it.  Someone even said it would win contests if I entered it into a baking contest.  And I HATE TO BAKE!  I plan on purchasing th
e 2 pack next time and can't wait to make the full cake recipe from the label, ice cream and other deserts with it!
0    This is actually very Good Honey, Will order again in the Future When Stock runs Lower Recommend to everyone who ever Likes the sweet Taste

0    Being a diabetic for 29 years, I have tasted virtually every sugar free candy on the market. These are wonderful -- hard to tell from the real thing. No more watching my frien
ds indulge without being able to join in! Use with caution, however -- excess consumption (like eating half the box during a really good movie -- not hard to do) will make you rath
er ill. On the whole, though, these are great.

0    arrived quickly, undamaged and as described, probably not the best priced.  cats love it.  contains low magnesium - so as not to form urine crystals.



Name: Reviews, dtype: object
```

Observations:

1. When Eps = 13.05, the number of elements in cluster 0 are 49206 and points in cluster -1 (or noise points) are 793.

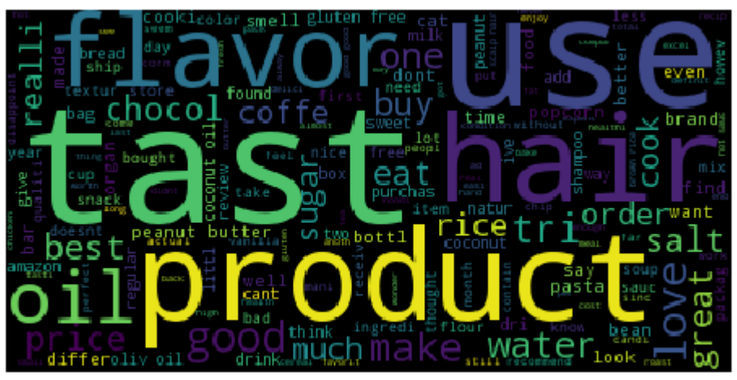
2. Cluster -1 reviews seem to be about the taste and flavour of the products. Many reviews also mention products about chocolate and hair.

3. Cluster 0 reviews talked about how the customers used the product and how good the flavour was. They also talked about how they made use of them.

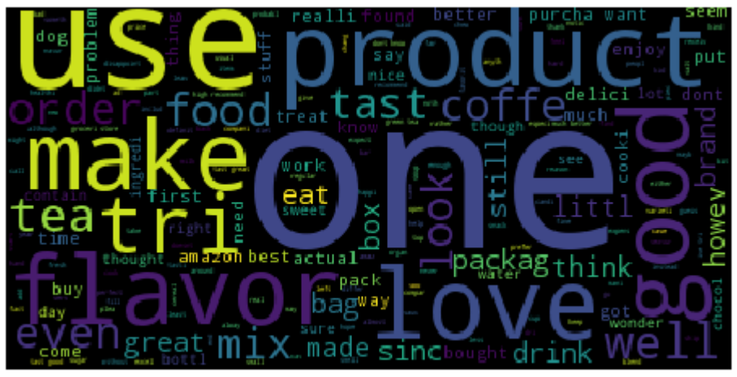
4. The number of points classified as noise increased with decrease in Eps value.

In [0]:
```
dbscan(grouped_data,data,13.05,min_pts)
```

```
For Eps = 13.05 the number of points in each cluster are:


 0    49206
-1      793
Name: labels, dtype: int64



==================== Plotting wordcloud and reviews for Eps = 13.05



Creating a dataframe with Reviews, Cleaned text and Clusters...
Dataframe is created!


The number of reviews in each cluster is:
 0    49206
-1      793
Name: Clusters, dtype: int64


**********************************************************************

************************************ Cluster  -1 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster -1
-1    I love this gluten.  I've used other, more expensive gluten & this is just as good, if not better.<br /><br />I use it all the time in my baking.  I love it.  Connie

-1    This spice is wonderful! It is very expensive if a person can find it at all. This was a fantastic buy. LOW Low Low price and fast shipping. Thanks so very muc
h!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

-1    Most of these cookies are great - some lack a little flavor (but worth the calorie savings!)

-1    I am writing this review because I feel that the only other review here is not justified. It says on the packaging that it is whole milk powder and it should be assumed that
it will be just that. I have been interested in purchasing this product for some time and recently found it at a grocery store. I have not had good luck finding a quality powdered
milk and this product has satisfied my desire for a quality powdered milk. It has a good flavor and does not taste stale or like cardboard. Of the brands of powdered milk I have tr
ied, many taste so bad I cannot even consume it when I backpack. This is bad because, when backpacking, I need to eat all I can and often will eat all I can and have. Bringing powd
ered milk I won't drink is a waste of weight I carry and calories I'll never get. I have seriously eaten grape nuts with water because my powdered milk was so stale. This product h
as a nice rich and creamy flavor that the non-fat milk powders don't have. I prepared this product and drank the result immediately, without any refrigeration and it was already ex
cellent. I think the only other milk powder that I have drank that was good was the carnation non-fat milk. Also, I have used vanilla carnation breakfast drink in place of powdered
milk because it works as a milk substitute on the trail. I enjoy this Nido whole milk powder because it is loaded with calories and also because it can be very useful in preparing
a number of different normal every day dishes or possibly ethnic at home, such as ice cream, kulfi, or when you want to thicken dairy based foods. I also drink whole milk on a norm
al basis and prefer to have the fat anyway. However, if I didn't prefer the fat, I would choose this over a bad tasting stale milk powder any day.
-1    This is one smooth, rick dark chocolate treat! Gevalia's Dark Chocolate Truffle coffee lives up to its name you can taste the rich chocolate goodness in every sip. The mouthw
atering aroma will bring chocolate lovers running. (My kids were quite disappointed to find that there was coffee in the pot with the yummy smelling chocolate.) The Husband (not a
dedicated dark chocolate consumer) thoroughly enjoyed his cuppa. He was pleased that the chocolate flavoring was rich and sweet rather than strong and bitter (which he normally ass
ociates with the dark chocolate blends). I can't wait to get out the holiday mugs and serve this to my dark chocolate loving friends and relatives.<br /><br />A very nice treat for
the coffee lover with a bit of a sweet tooth.



Name: Reviews, dtype: object



************************************ Cluster  0 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster 0
0    I do not have a Senseo coffee pod maker.<br /><br />I use a 4cup small MR COFFEE pot and use<br />2 pods for 4 cups of coffee.  Works perfectly.<br /><br />I like the SUMATRA
coffee a lot...bold, dark<br />and rich flavor.<br /><br />I add a little bit of BAILEYS IRISH CREAM to<br />each cup.<br /><br />I have tried the MELITTA coffee pods --  they<br />
>are weak and expensive... but,they do have<br />some very very interesting flavors.<br /><br />bette b prater<br />okc ok  62 retired
0    First of all, I really don't like the flavor of black licorice, but the taste in this tea was not overpowering. I was able to drink it just fine.<br /><br />I was worried abou
t making enough milk to feed my baby and to store for when I returned to work, so I started drinking this tea. I would brew a whole box, stick it in the fridge and sip glasses thro
ughout the day. The box recommends that you drink 3-5 cups per day. So one box (16 bags) will last 3-5 days.<br /><br />I didn't notice a huge change in my milk supply with this pr
oduct. I drank the tea for about 10 days before my midwife recommended I start fenugreek supplements. I tried fenugreek capsules as well as More Milk Special blend AND I upgraded t
o hospital grade pump. I noticed a significant increase in the number of dirty diapers my baby made when I started the supplements. However, I did not pump a lot of extra milk unti
l I upgraded my pump.<br /><br />My point is that you really need to evaluate your overall lactation profile (I just made up that term, I don't know if anyone else calls it that.)
I recommend the book "The Breastfeeding Mother's Guide to Making More Milk" by Diana West. It covers all the reasons you may not be making enough milk and many strategies on how to
tackle your issues. I found it very comprehensive.<br /><br />The tea is not a bad option. I gave it 4 stars because the company doesn't make any promises about milk supply, it tas
tes alright and the price is right. If you just want a little boost or if you are sensitive to any sort of pharmaceutical (herbs are drugs too!) starting with this tea might be a g
ood option for you. I really think you should also speak to a lactation consultant if you can. Or find a breastfeeding support group in your area.<br /><br />Happy nursing!!
0    Jelly Belly jelly beans have become the favorite snack for my piano students, but have noticed that their parents cannot resist combining several flavors to make your "recipe
s"!  So I thought I would try the Belly Flops, but we were all quite disappointed to see the lack of variety in two packages.  I hope you do a better job this time!  By the way, my
students love the misshapen Flops!

0    these were very good on salads and by the handfuls. I would recommend you trying these first then if you want to try the more bland plain ones ... I tried them both and like t
he spicy ones the best. they are nor real spice just enough to give it some umph :)

0    This soup is very delicious if you like curry and lentils, a perfect blend of flavors and so healthy, too. It is so good it is addictive. I never get tired of it. I have tried
many of the other Pacific Soups but none are as good as this one. For those of us saddled with restrictive diets (not vegan), a boring diet is hard to overcome. This soup meets all
the requirements that I need to eat both protein and and gluten-free fiber at every meal. When it started disappearing from the stores (in Denver), I was devastated. Kudos to Whole
Foods for turning me onto Amazon.com and kudos to amazon.com for selling discontinued foods.

Name: Reviews, dtype: object
```

Observations:

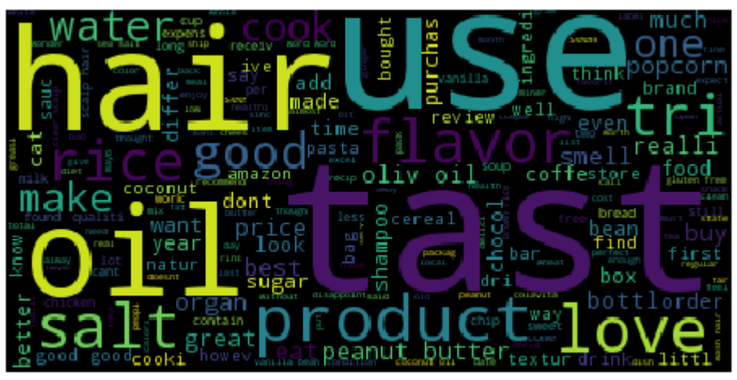
1. When Eps = 13.05, the number of elements in cluster 0 are 49206 and points in cluster -1 (or noise points) are 793.

2. Cluster -1 reviews seem to be about the taste and flavour of the products. Also a significant number of reviews mention that the product has something to do with the hair and oil.

3. Cluster 0 reviews talked about how the customers used the product and how good the flavour was. They also talked about how they made use of them.

In [0]:
```
dbscan(grouped_data,data,14.05,min_pts)
```

```
For Eps = 14.05 the number of points in each cluster are:


 0    49677
-1    322
Name: labels, dtype: int64



==================== Plotting wordcloud and reviews for Eps = 14.05



Creating a dataframe with Reviews, Cleaned text and Clusters...
Dataframe is created!


The number of reviews in each cluster is:
 0    49677
-1    322
Name: Clusters, dtype: int64


**********************************************************************

************************************ Cluster  -1 ************************************

Plot of Word Cloud
```




```
Printing 4 random reviews from cluster -1
-1    The coconut oil works great on my hair and skin. my hair responded great to the oil and it grew a little because of the oil.
-1    Bought 5LB of this cheese and I use it for mac and cheese. Very good cheese sauce I would definitely buy more!
-1    Came as expected on expected time. Was what I was hoping for when comes for pacaging and taste. Very good!
-1    this can chicken taste great in hot dishes, chicken salad, chicken helper, salads, or whatever calls for chicken well worth the money will buy again.
-1    Love this pancake mix.  This is the only brand we have used n years.
Name: Reviews, dtype: object



************************************ Cluster  0 ************************************

Plot of Word Cloud
```


```
Printing 4 random reviews from cluster 0
0    These things are ridiculous.  To open a bag is to finish it.  I wish that they could be just, like, *medium* good, or even just "quite good" to "very good".  Unfortunately the
y're like the snack version of crack.



0    We love whole bean french roast coffee by Eight O'Clock, but were having such a hard time finding it in super markets. Amazon had it and we are so pleased we signed up to rece
ive the coffee once a month. The price is right that is for sure. If you like French Roast whole bean coffee, I highly recommend this purchase.



0    I do not have a Senseo coffee pod maker.<br /><br />I use a 4cup small MR COFFEE pot and use<br />2 pods for 4 cups of coffee.  Works perfectly.<br /><br />I like the SUMATRA
coffee a lot...bold, dark<br />and rich flavor.<br /><br />I add a little bit of BAILEYS IRISH CREAM to<br />each cup.<br /><br />I have tried the MELITTA coffee pods --  they<br />
>are weak and expensive... but,they do have<br />some very very interesting flavors.<br /><br />bette b prater<br />okc ok  62 retired
```

```
0    First of all, I really don't like the flavor of black licorice, but the taste in this tea was not overpowering. I was able to drink it just fine.<br /><br />I was worried abou
t making enough milk to feed my baby and to store for when I returned to work, so I started drinking this tea. I would brew a whole box, stick it in the fridge and sip glasses thro
ughout the day. The box recommends that you drink 3-5 cups per day. So one box (16 bags) will last 3-5 days.<br /><br />I didn't notice a huge change in my milk supply with this pr
oduct. I drank the tea for about 10 days before my midwife recommended I start fenugreek supplements. I tried fenugreek capsules as well as More Milk Special blend AND I upgraded t
o hospital grade pump. I noticed a significant increase in the number of dirty diapers my baby made when I started the supplements. However, I did not pump a lot of extra milk unti
l I upgraded my pump.<br /><br />My point is that you really need to evaluate your overall lactation profile (I just made up that term, I don't know if anyone else calls it that.)
I recommend the book "The Breastfeeding Mother's Guide to Making More Milk" by Diana West. It covers all the reasons you may not be making enough milk and many strategies on how to
tackle your issues. I found it very comprehensive.<br /><br />The tea is not a bad option. I gave it 4 stars because the company doesn't make any promises about milk supply, it tas
tes alright and the price is right. If you just want a little boost or if you are sensitive to any sort of pharmaceutical (herbs are drugs too!) starting with this tea might be a g
ood option for you. I really think you should also speak to a lactation consultant if you can. Or find a breastfeeding support group in your area.<br /><br />Happy nursing!!
0    Jelly Belly jelly beans have become the favorite snack for my piano students, but have noticed that their parents cannot resist combining several flavors to make your "recipe
s"!  So I thought I would try the Belly Flops, but we were all quite disappointed to see the lack of variety in two packages.  I hope you do a better job this time!  By the way, my
students love the misshapen Flops!



Name: Reviews, dtype: object
```

Observations:

1. When Eps = 14.05, the number of elements in cluster 0 are 49677 and points in cluster -1 (or noise points) are 322.

2. Cluster -1 reviews seem to be about hair products and oil products and how most customers used them. A significant number of reviews talked about taste and flavour of the products.

3. Cluster 0 reviews talked about how the customers used the product and how good the flavour was. They also talked about how they made use of them.

4. The number of points classified as noise decreased with increase in Eps value.

Summary and Conclusions

1) Import the csv file containing pre processed data which is already arranged on the basis of time.

2) Three functions are created for plotting wordcloud, for plotting graph in plotly and performing DBSCAN.

3) The first function creates a dataframe with all the reviews, cleaned text and clusters. Then it prints the number of data points in each cluster. A wordcloud is then plotted along with a few random reviews from that cluster.

4) The second function plots graph between indices and distance array using plotly. Plotly is chosen so that we can pin point the exact coordinate for a point on the graph.

5) The third function performs DBSCAN and gets the best Eps value by plotting indices vs nearest distances. Then it gets the labels for visualisation through wordcloud.

6) These functions are applied on avg Word2Vec and TFIDF Word2Vec.

6) It can be observed that as the eps value increases, the number of points classified as a noise point decreases. Hence it is important to choose the right Eps value so that noise points are well seperated from other data points.