[1] Amazon Fine Food Reviews Analysis

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

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.

Number of reviews: 568,454
Number of users: 256,059
Number of products: 74,258
Timespan: Oct 1999 - Oct 2012
Number of Attributes/Columns in data: 10

Attribute Information:

1. Id
2. ProductId - unique identifier for the product
3. UserId - unqiue identifier for the user
4. ProfileName
5. HelpfulnessNumerator - number of users who found the review helpful
6. HelpfulnessDenominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

Objective:

Given a review, determine whether the review is positive (Rating of 4 or 5) or negative (rating of 1 or 2).

[Q] How to determine if a review is positive or negative?

[Ans] We could use the Score/Rating. A rating of 4 or 5 could be cosnidered a positive review. A review of 1 or 2 could be considered negative. A review of 3 is nuetral and ignored. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

Loading the data

The dataset is available in two forms

1. .csv file
2. SQLite Database

In order to load the data, We have used the SQLITE dataset as it easier to query the data and visualise the data efficiently.

Here as we only want to get the global sentiment of the recommendations (positive or negative), we will purposefully ignore all Scores equal to 3. If the score id above 3, then the recommendation wil be set to "positive". Otherwise, it will be set to "negative".

Task to be performed:

- Find the best ‘k’ using the elbow-knee method (plot k vs inertia_)
- Once after you find the k clusters, plot the word cloud per each cluster so that at a single go we can analyze the words in a cluster.
- Also apply the k-medoids algorithm as well.

In [1]:
```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cross_validation import train_test_split
```

```
from sklearn.cross_validation import cross_val_score
from sklearn.metrics import accuracy_score,precision_score,recall_score
,classification_report,confusion_matrix
from sklearn.metrics import roc_auc_score, roc_curve
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import GridSearchCV
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
import sqlite3
```

```
C:\Users\manish dogra\Documents\anaconda\lib\site-packages\sklearn\cros
s_validation.py:41: DeprecationWarning: This module was deprecated in v
ersion 0.18 in favor of the model_selection module into which all the r
efactored classes and functions are moved. Also note that the interface
of the new CV iterators are different from that of this module. This mo
dule will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

In [2]:
```
con = sqlite3.connect("./amazon-fine-food-reviews/database.sqlite")
data = pd.read_sql_query('''
SELECT *
FROM REVIEWS
WHERE SCORE != 3''', con)
data.shape
```

Out[2]: (525814, 10)

Data Cleaning

In [3]:
```
data = data[data.HelpfulnessNumerator <= data.HelpfulnessDenominator]
data.shape
```

Out[3]: (525812, 10)

Observation:- It was also seen that in two rows given below the value of HelpfulnessNumerator is greater than HelpfulnessDenominator which is not practically possible hence these two rows too are removed from calcualtions

In [4]:
```
data['Score'] = data["Score"].apply(lambda x: "positive" if x > 3 else "negative")
sorted_data = data.sort_values('ProductId',axis = 0, inplace = False, kind = 'quicksort',ascending = True)
sorted_data.head()
```

Out[4]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	He
138693	150511	0006641040	A1C9K534BCI9GO	Laura Purdie Salas	0	0
138708	150526	0006641040	A3E9QZFE9KXH8J	R. Mitchell	11	18
138707	150525	0006641040	A2QID6VCFTY51R	Rick	1	2
138706	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0	0

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	He
138705	150523	0006641040	A2P4F2UO0UMP8C	Elizabeth A. Curry "Lovely Librarian"	0	0

In [5]:
```
filtered_data = sorted_data.drop_duplicates(subset = {'UserId','Profile
Name','Time'} ,keep = 'first', inplace = False)
filtered_data.shape
```

Out[5]: (328770, 10)

In [6]:
```
final = filtered_data.copy()
import nltk
nltk.download('stopwords')
```

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

Out[6]: True

In [7]:
```
stop = set(stopwords.words("english"))
st = PorterStemmer()
```

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

Now that we have finished deduplication our data requires some preprocessing before we go on further with analysis and making the prediction model.

Hence in the Preprocessing phase we do the following in the order below:-

1. Begin by removing the html tags
2. Remove any punctuations or limited set of special characters like , or . or # etc.
3. Check if the word is made up of english letters and is not alpha-numeric
4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)
5. Convert the word to lowercase
6. Remove Stopwords
7. Finally Snowball Stemming the word (it was obsereved to be better than Porter Stemming)

After which we collect the words used to describe positive and negative reviews

In [8]:
```
def cleanhtml(sent):
    cleanr = re.compile('<.*?>')
    cleaned = re.sub(cleanr,' ',sent)
    return cleaned
def cleanpunc(sent):
    clean = re.sub(r'[?|!|$|#|\'|"|:]',r'',sent)
    clean = re.sub(r'[,|(|)|.|\|/]',r' ',clean)
    return clean
```

In [9]:
```
i=0
all_positive_reviews =[]
all_negative_reviews = []
final_string = []
stem_data = " "
for p in final['Text'].values:
    filtered_sens = []#filtered word
    p = cleanhtml(p)
    for w in p.split():
       # print(w)
        punc = cleanpunc(w)
        for s in punc.split():
            #print(w)
            if (s.isalpha()) & (len(s)>2):
                if s.lower() not in stop:
                    stem_data = (st.stem(s.lower())).encode('utf8')
```

```
                    #can we use lemmatizer and stemming altogether??
                    filtered_sens.append(stem_data)
                    if (final['Score'].values)[i] == 'positive':
                        all_positive_reviews.append(stem_data)
                    if (final['Score'].values)[i] == 'negative':
                        all_negative_reviews.append(stem_data)
                else:
                    continue
            else:
                continue
    #print(filtered_sens)
    str1 = b" ".join(filtered_sens)
    #print(str1)
    final_string.append(str1)
    i+=1
```

In [10]:
```
final['CleanedText'] = final_string
final.head()
```

Out[10]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	He
138693	150511	0006641040	A1C9K534BCI9GO	Laura Purdie Salas	0	0
138708	150526	0006641040	A3E9QZFE9KXH8J	R. Mitchell	11	18

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	He
138707	150525	0006641040	A2QID6VCFTY51R	Rick	1	2
138706	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0	0
138705	150523	0006641040	A2P4F2UO0UMP8C	Elizabeth A. Curry "Lovely Librarian"	0	0

In [11]:
```
final = final.sort_values('Time',axis= 0,inplace = False , na_position
= 'last',ascending = True)
X = final['CleanedText'].values
X = X[:100000]
```

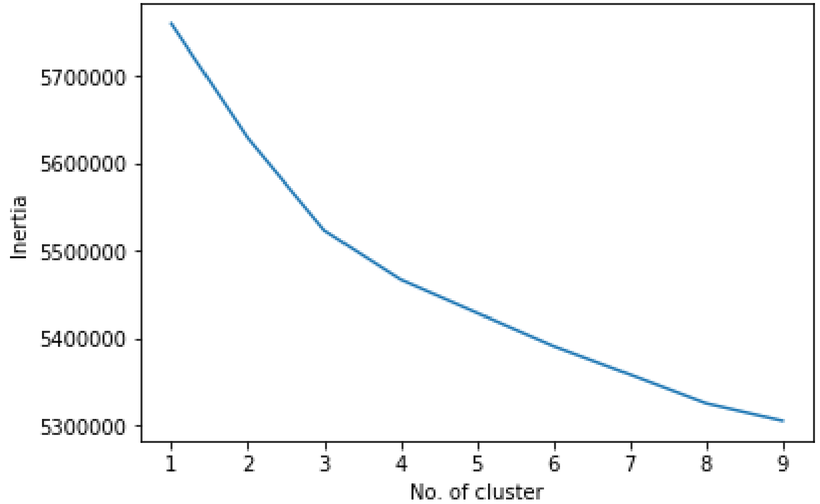
Bow

In [12]:
```
count_vect = CountVectorizer()
bow = count_vect.fit_transform(X)
```

In [13]:
```
list_of_sent = []
for i in X:
```

```
    sent = []
    for word in i.split():
        sent.append(word.decode('utf-8'))
    list_of_sent.append(sent)
```

In [14]:
```
from sklearn.cluster import KMeans
dic = {}
for i in range(1,10):
    clus = KMeans(n_clusters = i)
    clus.fit(bow)
    dic[i] = clus.inertia_
plt.plot(list(dic.keys()), list(dic.values()))
plt.xlabel("No. of cluster")
plt.ylabel("Inertia")
plt.show()
```



In [15]:
```
optimal_k = KMeans(n_clusters = 8)
p = optimal_k.fit_predict(bow)
```

In [34]:
```
index = []
for i in range(len(p)):
```

```
        if p[i] == 1:
            index.append(i)
```

In [35]:
```
len(index)
```

Out[35]: 66621

In [36]:
```
text = []
for i in range(len(index)):
    text.append(list_of_sent[index[i]])
```

In [37]:
```
from wordcloud import WordCloud
from matplotlib.pyplot import figure
t_b = ''
for j in range(len(text)):
    for i in range(len(text[j])):
        t_b = t_b + text[j][i] + ' '
#print(t_b)
word_cloud = WordCloud(relative_scaling = 1.0).generate(t_b)
plt.imshow(word_cloud,aspect='auto')
plt.axis('off')
plt.show()
```

conclusion

- we have applied the KMeans clustering on Bag of words and created the word cloud of cluster(1) to analyse the what kind of word are there in this cluster exist.
- Most common words in cluster(1) are written in bolder size in the image.

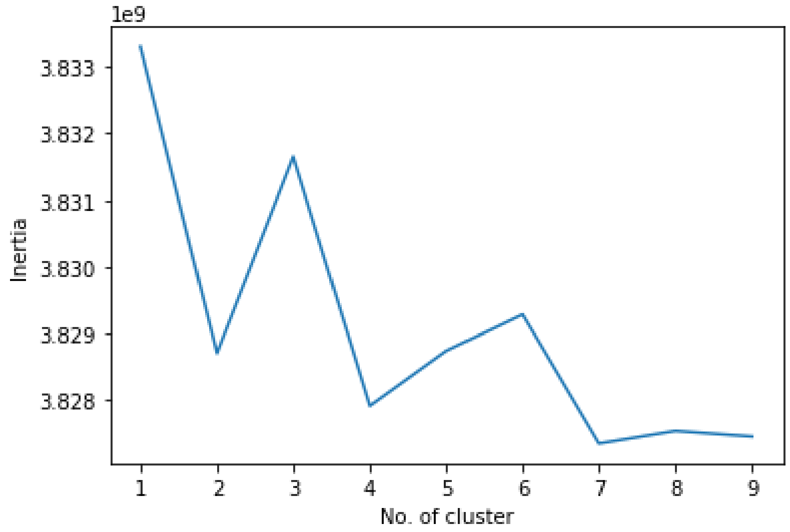
Tfidf

In [38]:

```
tfidf_vect = TfidfVectorizer()
tfidf_train = tfidf_vect.fit_transform (X)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler(with_mean = False)
X_tr = sc.fit_transform(tfidf_train)
```

In [39]:

```
from sklearn.cluster import KMeans
dit = {}
for i in range(1,10):
    clus_tf = KMeans(n_clusters = i)
    clus_tf.fit(X_tr)
    dit[i] = clus_tf.inertia_
plt.plot(list(dit.keys()), list(dit.values()))
plt.xlabel("No. of cluster")
plt.ylabel("Inertia")
plt.show()
```



In [40]:
```
opt_k = KMeans(n_clusters = 7)
p = optimal_k.fit_predict(tfidf_train)
```

In [57]:
```
index = []
for i in range(len(p)):
    if p[i] == 5:
        index.append(i)
```

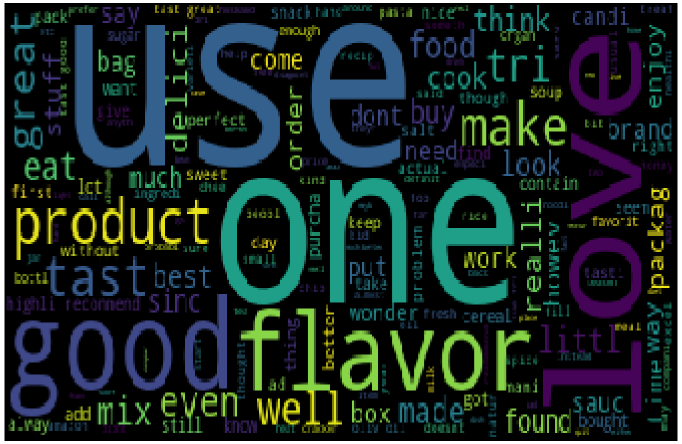
In [58]:
```
len(index)
```

Out[58]: 50241

In [59]:
```
text = []
for i in range(len(index)):
    text.append(list_of_sent[index[i]])
```

In [60]:
```
from wordcloud import WordCloud
from matplotlib.pyplot import figure
t_b = ''
for j in range(len(text)):
```

```
    for i in range(len(text[j])):
        t_b = t_b + text[j][i] + ' '
#print(t_b)
word_cloud = WordCloud(relative_scaling = 1.0).generate(t_b)
plt.imshow(word_cloud,aspect='auto')
plt.axis('off')
plt.show()
```



conclusion

- we have applied the KMeans clustering on Tfidf and created the word cloud of cluster(5) to analyse the what kind of word are there in this cluster exist.
- Most common words in cluster(5) are written in bolder size in the image.

Avg W2vec

In [61]:
```
list_of_sent = []
for i in X:
    sent = []
```

```
    for word in i.split():
        sent.append(word.decode('utf-8'))
    list_of_sent.append(sent)
```

In [62]:
```
from gensim.models import Word2Vec
w2v_model = Word2Vec(list_of_sent,min_count = 5,size = 50,workers = 4)
sent_vectors = []
for sent in list_of_sent:
    sent_vec = np.zeros(50)
    cnt_word = 0
    for word in sent:
        try:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_word += 1
        except:
            pass
    sent_vec /= cnt_word
    sent_vectors.append(sent_vec)
print(len(sent_vectors))
```

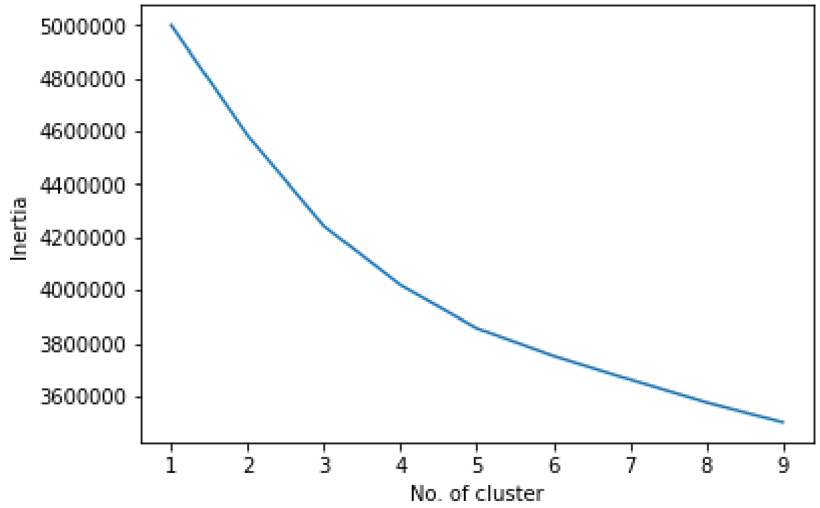
```
C:\Users\manish dogra\Documents\anaconda\lib\site-packages\gensim\util
s.py:1197: UserWarning: detected Windows; aliasing chunkize to chunkize
_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_seria
l")
```

```
100000
```

In [63]:
```
sc = StandardScaler()
w2v = sc.fit_transform(sent_vectors)
```

In [64]:
```
from sklearn.cluster import KMeans
diw = {}
for i in range(1,10):
    clus_w2 = KMeans(n_clusters = i)
    clus_w2.fit(w2v)
    diw[i] = clus_w2.inertia_
plt.plot(list(diw.keys()), list(diw.values()))
```

```
plt.xlabel("No. of cluster")
plt.ylabel("Inertia")
plt.show()
```



In [67]:
```
opt_k = KMeans(n_clusters = 8)
p = opt_k.fit_predict(w2v)
```

In [89]:
```
index = []
for i in range(len(p)):
    if p[i] == 2:
        index.append(i)
```

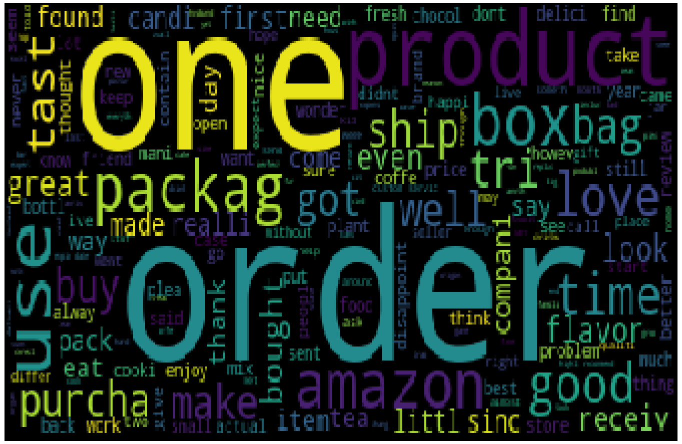
In [91]:
```
len(index)
```

Out[91]: 15215

In [92]:
```
text = []
for i in range(len(index)):
    text.append(list_of_sent[index[i]])
```

In [93]:
```
from wordcloud import WordCloud
```

```
from matplotlib.pyplot import figure
t_b = ''
for j in range(len(text)):
    for i in range(len(text[j])):
        t_b = t_b + text[j][i] + ' '
#print(t_b)
word_cloud = WordCloud(relative_scaling = 1.0).generate(t_b)
plt.imshow(word_cloud,aspect='auto')
plt.axis('off')
plt.show()
```



conclusion

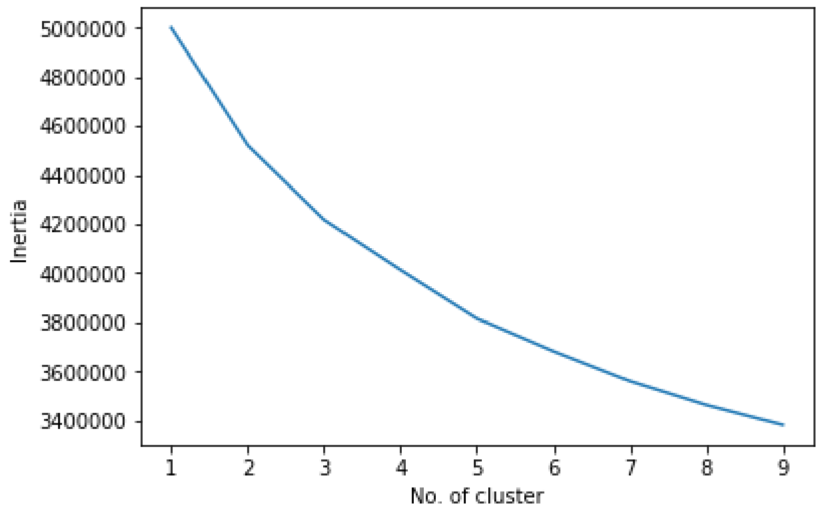
- we have applied the KMeans clustering on Avg-w2vec and created the word cloud of cluster(2) to analyse the what kind of word are there in this cluster exist.
- Most common words in cluster(2) are written in bolder size in the image.

Tfidf W2vec

In [65]:
```
tf_idf_feat = tfidf_vect.get_feature_names()
tfidf_sent_vec = []
row = 0
for sent in list_of_sent:
    sent_vec = np.zeros(50)
    weight_sum = 0
    for word in sent:
        try:
            vec = w2v_model.wv[word]
            tfidf = tfidf_train[row,tf_idf_feat.index(word)]
            sent_vec += (vec*tfidf)
            weight_sum += tfidf
        except:
            pass
    sent_vec/= weight_sum
    tfidf_sent_vec.append(sent_vec)
    row += 1
```

In [94]:
```
sc =  StandardScaler()
tfidf_w2v = sc.fit_transform(tfidf_sent_vec)
```

In [95]:
```
from sklearn.cluster import KMeans
diw = {}
for i in range(1,10):
    clus_w = KMeans(n_clusters = i)
    clus_w.fit(tfidf_w2v)
    diw[i] = clus_w.inertia_
plt.plot(list(diw.keys()), list(diw.values()))
plt.xlabel("No. of cluster")
plt.ylabel("Inertia")
plt.show()
```



In [96]:
```
opt_k = KMeans(n_clusters = 8)
p = opt_k.fit_predict(tfidf_w2v)
```

In [111]:
```
index = []
for i in range(len(p)):
    if p[i] == 5:
        index.append(i)
```

In [112]:
```
len(index)
```

Out[112]: 23244

In [113]:
```
text = []
for i in range(len(index)):
    text.append(list_of_sent[index[i]])
```

In [114]:
```
from wordcloud import WordCloud
from matplotlib.pyplot import figure
t_b = ''
for j in range(len(text)):
    for i in range(len(text[j])):
```

```
        t_b = t_b + text[j][i] + ' '
#print(t_b)
word_cloud = WordCloud(relative_scaling = 1.0).generate(t_b)
plt.imshow(word_cloud,aspect='auto')
plt.axis('off')
plt.show()
```

conclusion

- we have applied the KMeans clustering on Tfidf-w2vec and created the word cloud of cluster(5) to analyse the what kind of word are there in this cluster exist.
- Most common words in cluster(5) are written in bolder size in the image.