AMAZON FOOD REVIEWS CLUSTERING USING K-MEANS, K-MEANS++ AND AGGLOMERATIVE CLUSTERING ALGORITHMS

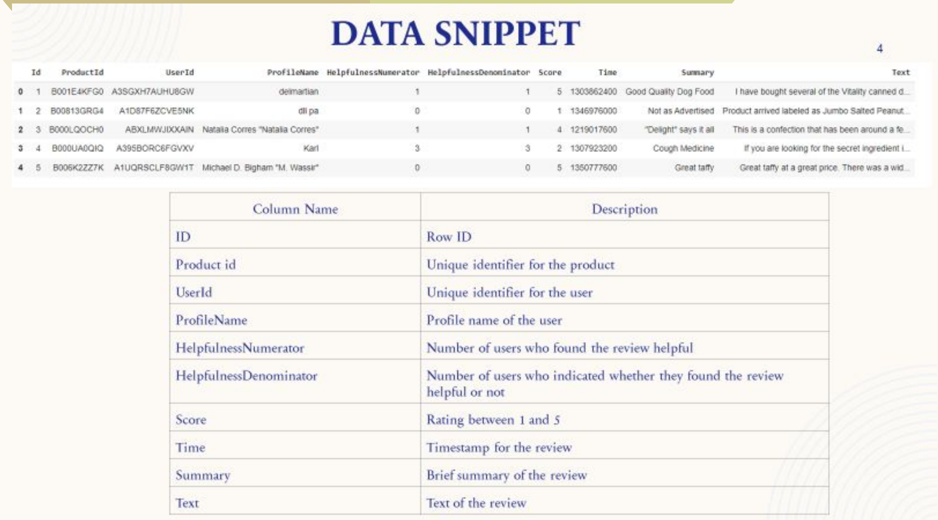
Rajashree Pethkar
Upasana Chaudhari

GOALS

- Cluster data using k-means, k-means++ and agglomerative
- Visualize the clusters
- Compare the results

HOW DID WE DO IT?

- Preprocessed the data
- Extracted features using TF-IDF
- Implemented PCA to reduce the dimensions
- Performed K-means , K-means ++, and Agglomerative clustering
- Generated the output by plotting the clusters
- Compare the results
- Library used: `sklearn`



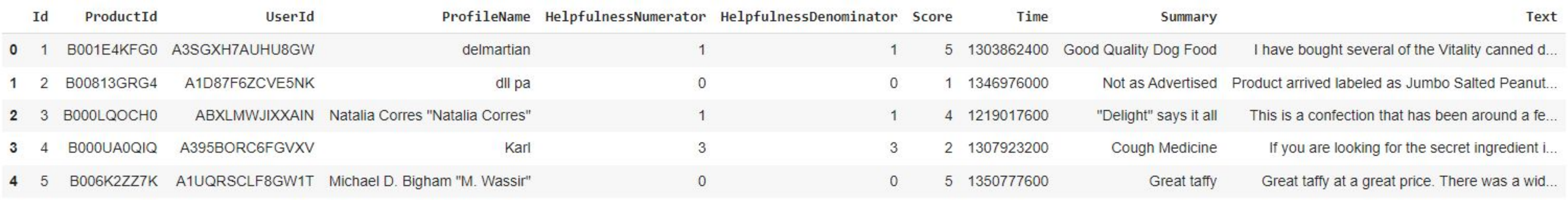
DATA SNIPPET

4

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Column Name	Description
ID	Row ID
Product id	Unique identifier for the product
UserId	Unique identifier for the user
ProfileName	Profile name of the user
HelpfulnessNumerator	Number of users who found the review helpful
HelpfulnessDenominator	Number of users who indicated whether they found the review helpful or not
Score	Rating between 1 and 5
Time	Timestamp for the review
Summary	Brief summary of the review
Text	Text of the review

DATA SNIPPET



	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Column Name	Description
ID	Row ID
Product id	Unique identifier for the product
UserId	Unique identifier for the user
ProfileName	Profile name of the user
HelpfulnessNumerator	Number of users who found the review helpful
HelpfulnessDenominator	Number of users who indicated whether they found the review helpful or not
Score	Rating between 1 and 5
Time	Timestamp for the review
Summary	Brief summary of the review
Text	Text of the review

PREPROCESSING THE DATA

- Converted the text to lowercase.
- Removed punctuation i.e any non-alphanumeric characters in the text is removed.
- Replaced occurrences of characters like /(){}\[\]\|@,;.with a space.
- Removed occurrences of bad symbol. Eg- characters like #+_“
- Removed digits.
- Split the text into individual words, removes stop words using the STOPWORDS set, and joins the remaining words back into a single string.
- Dropped unwanted columns.

FEATURE EXTRACTION USING TF-IDF

- TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic used to evaluate the importance of a term (word) within a document.

- TF gives us information on how often a term appears in a document.

- IDF gives us information about the relative rarity of a term in the collection of documents. By multiplying these values together we can get our final TF-IDF value.

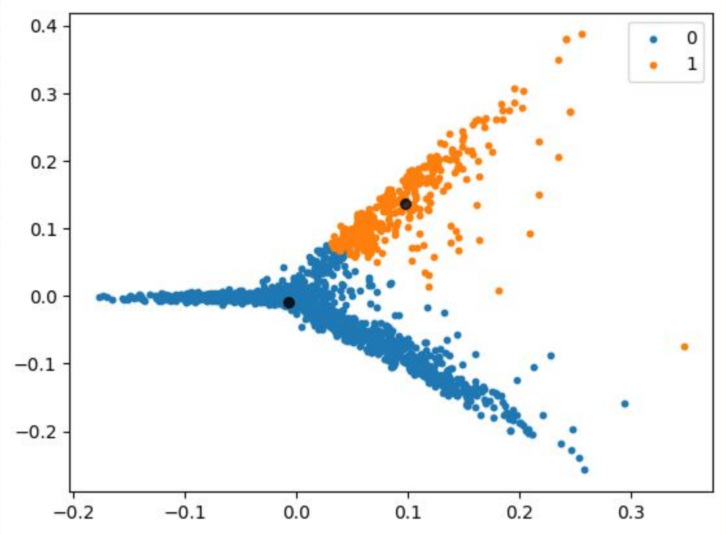
- The higher the TF-IDF score the more important or relevant the term is; as a term gets less relevant, its TF-IDF score will approach 0

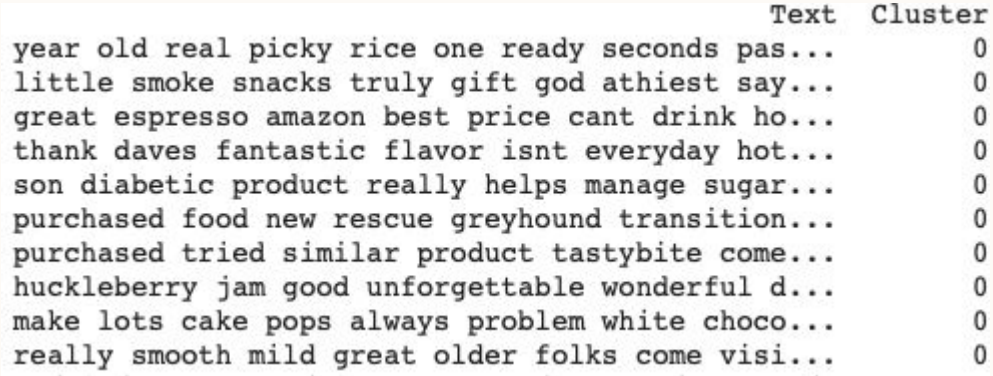
- # of features extracted: (5000, 163035)

IMPLEMENTED PCA

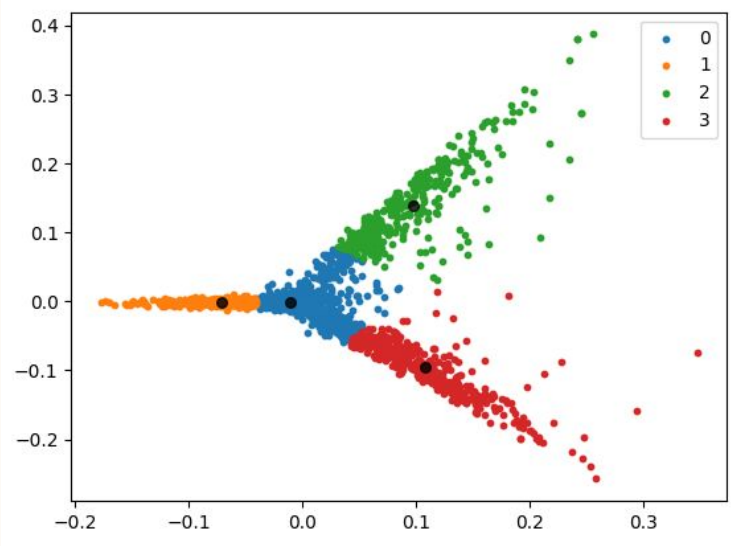
- Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of data while retaining the most important information.
- Implemented PCA using sklearn library.
- Reduced dimension: (5000, 2)

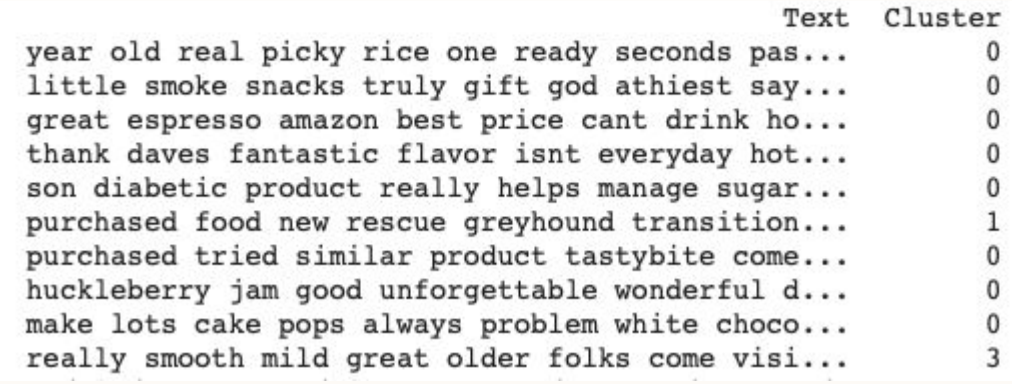
KMEANS AND KMEANS++





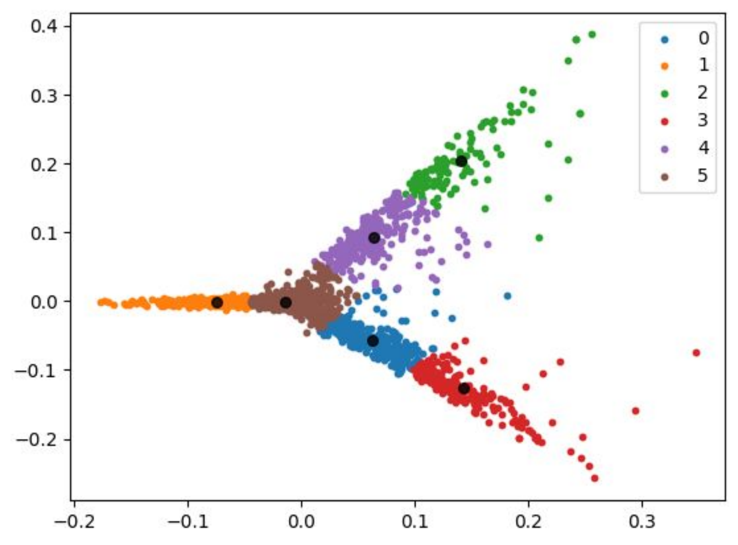
```
                                                 Text  Cluster
year old real picky rice one ready seconds pas...        0
little smoke snacks truly gift god athiest say...        0
great espresso amazon best price cant drink ho...        0
thank daves fantastic flavor isnt everyday hot...        0
son diabetic product really helps manage sugar...        0
purchased food new rescue greyhound transition...        0
purchased tried similar product tastybite come...        0
huckleberry jam good unforgettable wonderful d...        0
make lots cake pops always problem white choco...        0
really smooth mild great older folks come visi...        0
```

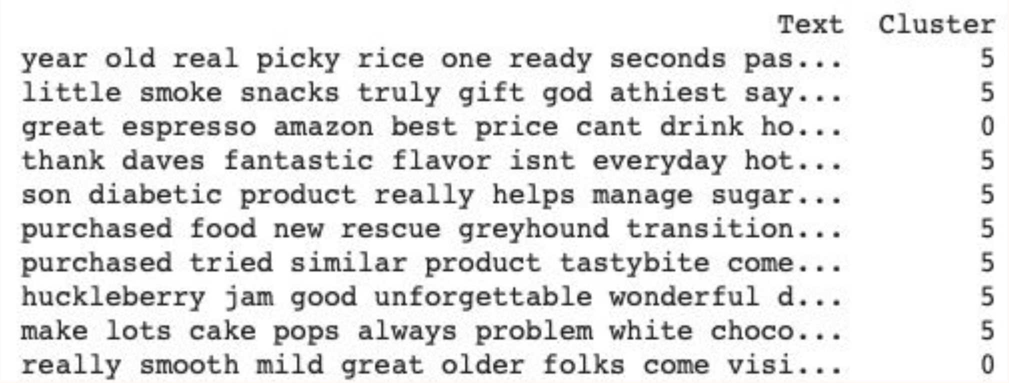




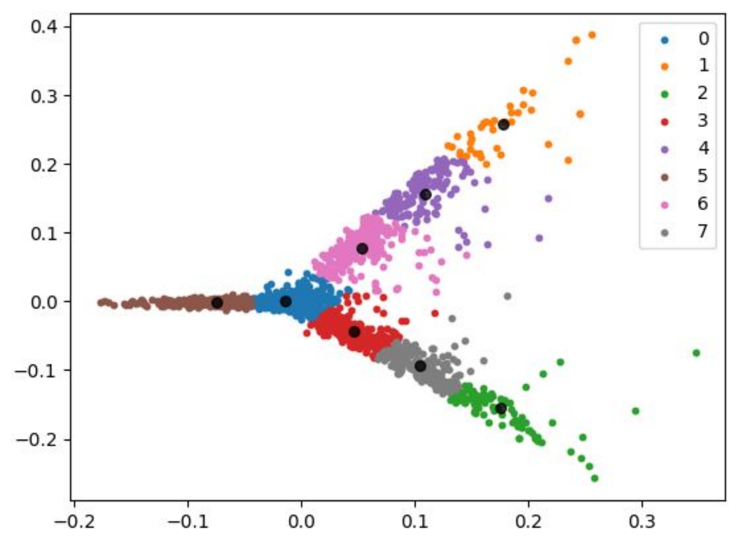
```
                                                 Text  Cluster
year old real picky rice one ready seconds pas...        0
little smoke snacks truly gift god athiest say...        0
great espresso amazon best price cant drink ho...        0
thank daves fantastic flavor isnt everyday hot...        0
son diabetic product really helps manage sugar...        0
purchased food new rescue greyhound transition...        1
purchased tried similar product tastybite come...        0
huckleberry jam good unforgettable wonderful d...        0
make lots cake pops always problem white choco...        0
really smooth mild great older folks come visi...        3
```

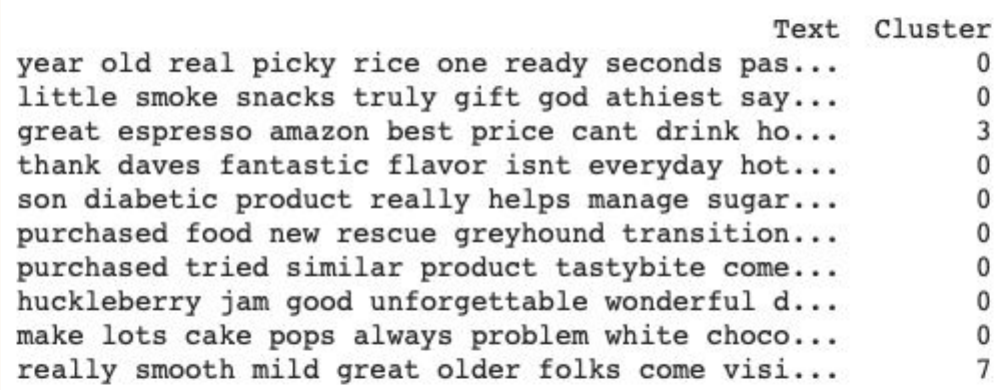
KMEANS AND KMEANS++





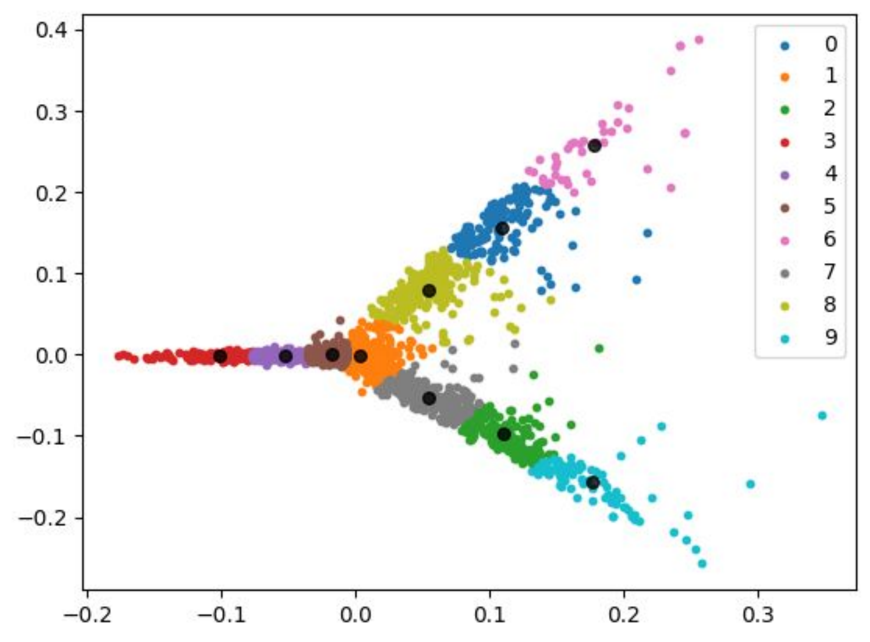
```
                                                Text  Cluster
year old real picky rice one ready seconds pas...        5
little smoke snacks truly gift god athiest say...        5
great espresso amazon best price cant drink ho...        0
thank daves fantastic flavor isnt everyday hot...        5
son diabetic product really helps manage sugar...        5
purchased food new rescue greyhound transition...        5
purchased tried similar product tastybite come...        5
huckleberry jam good unforgettable wonderful d...        5
make lots cake pops always problem white choco...        5
really smooth mild great older folks come visi...        0
```

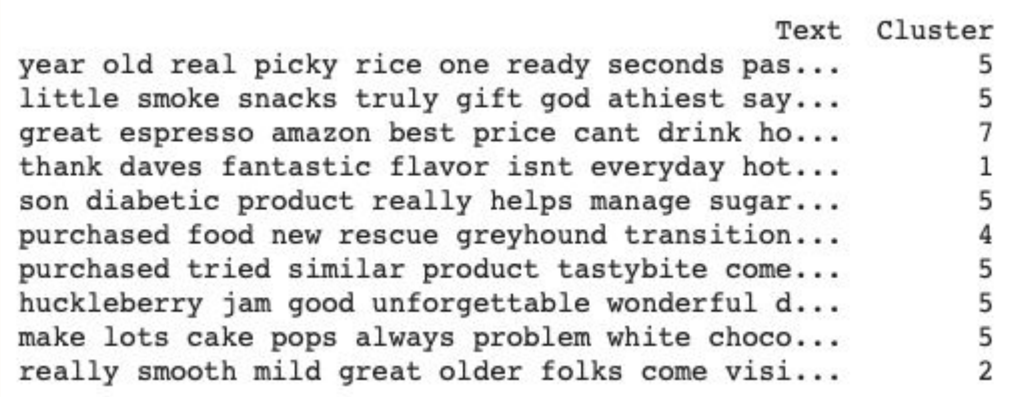




```
                                                Text  Cluster
year old real picky rice one ready seconds pas...        0
little smoke snacks truly gift god athiest say...        0
great espresso amazon best price cant drink ho...        3
thank daves fantastic flavor isnt everyday hot...        0
son diabetic product really helps manage sugar...        0
purchased food new rescue greyhound transition...        0
purchased tried similar product tastybite come...        0
huckleberry jam good unforgettable wonderful d...        0
make lots cake pops always problem white choco...        0
really smooth mild great older folks come visi...        7
```

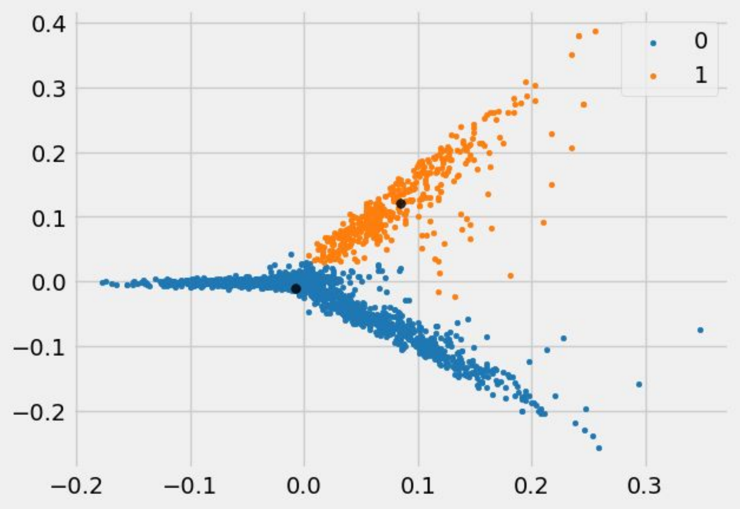
KMEANS AND KMEANS++

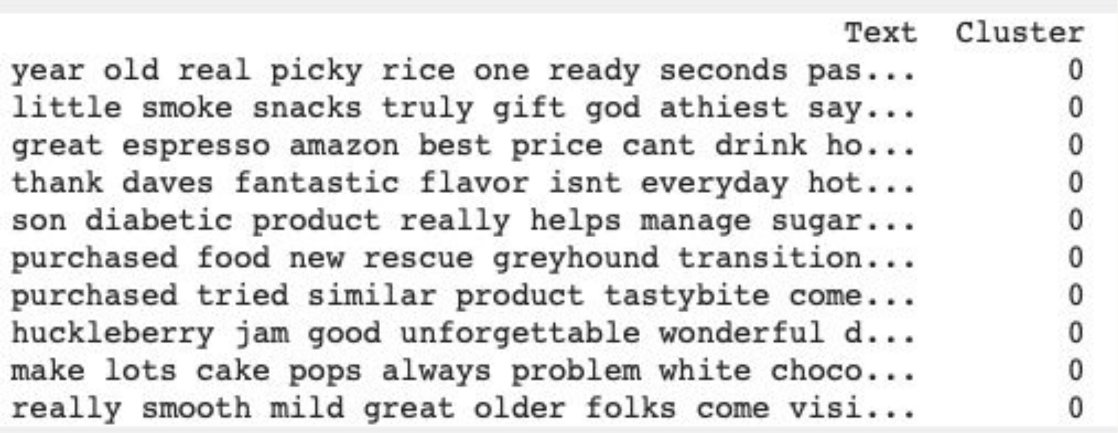




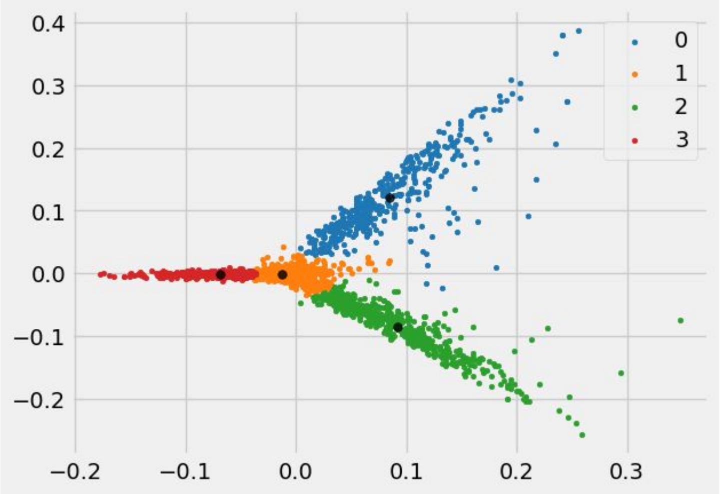
```
                                                 Text  Cluster
year old real picky rice one ready seconds pas...        5
little smoke snacks truly gift god athiest say...        5
great espresso amazon best price cant drink ho...        7
thank daves fantastic flavor isnt everyday hot...        1
son diabetic product really helps manage sugar...        5
purchased food new rescue greyhound transition...        4
purchased tried similar product tastybite come...        5
huckleberry jam good unforgettable wonderful d...        5
make lots cake pops always problem white choco...        5
really smooth mild great older folks come visi...        2
```

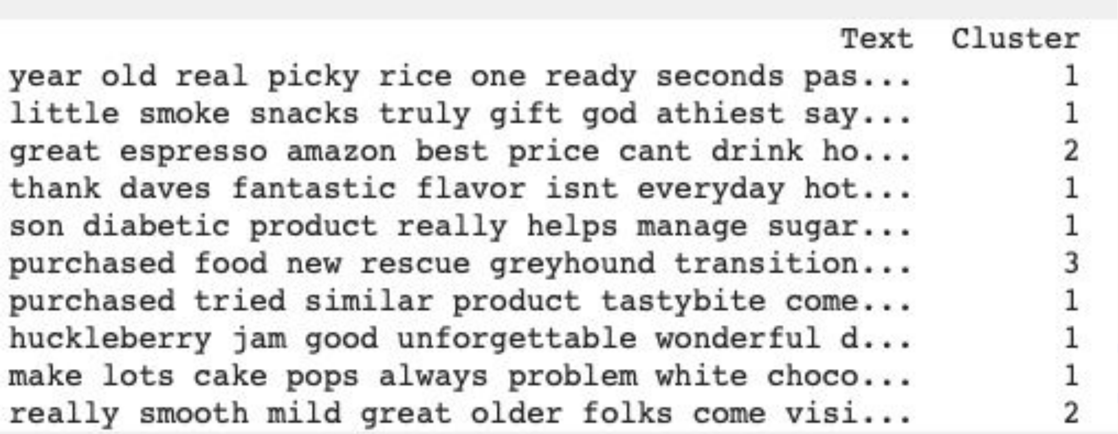
AGGLOMERATIVE CLUSTERING





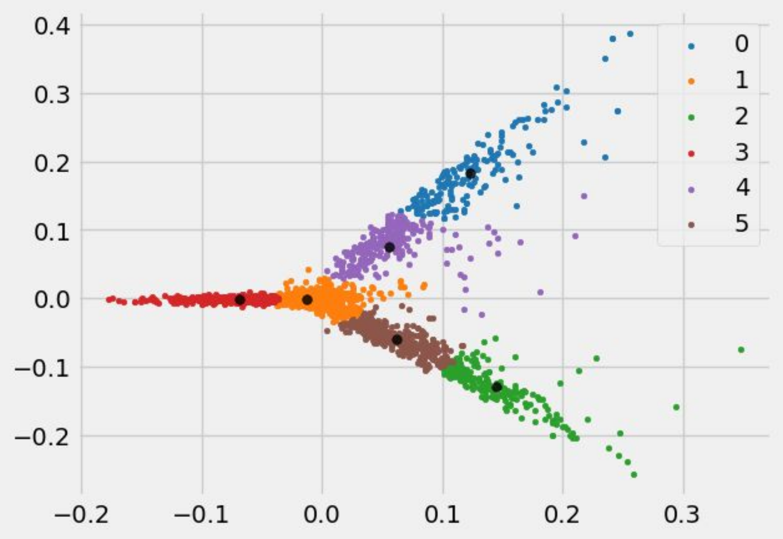
```
                                                   Text  Cluster
year old real picky rice one ready seconds pas...        0
little smoke snacks truly gift god athiest say...        0
great espresso amazon best price cant drink ho...        0
thank daves fantastic flavor isnt everyday hot...        0
son diabetic product really helps manage sugar...        0
purchased food new rescue greyhound transition...        0
purchased tried similar product tastybite come...        0
huckleberry jam good unforgettable wonderful d...        0
make lots cake pops always problem white choco...        0
really smooth mild great older folks come visi...        0
```

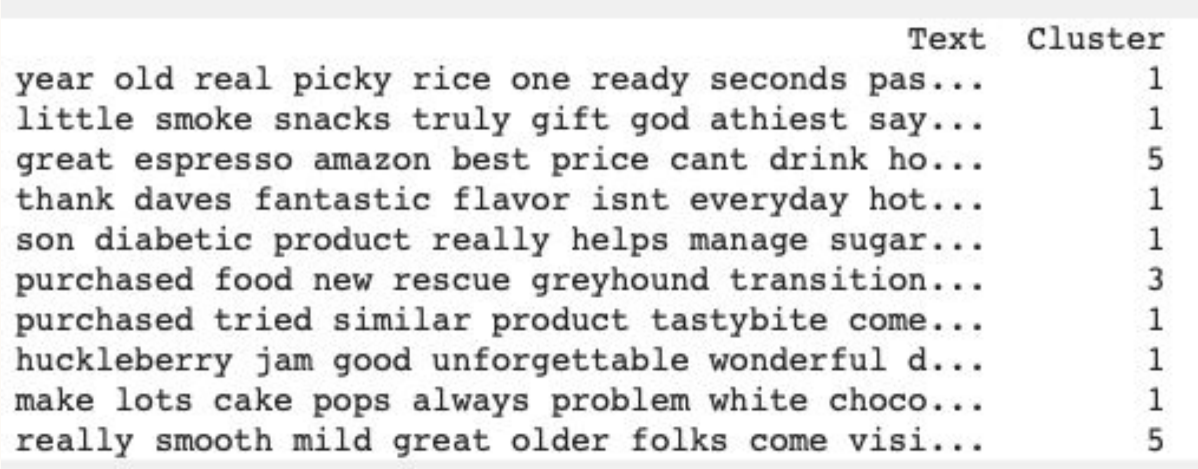


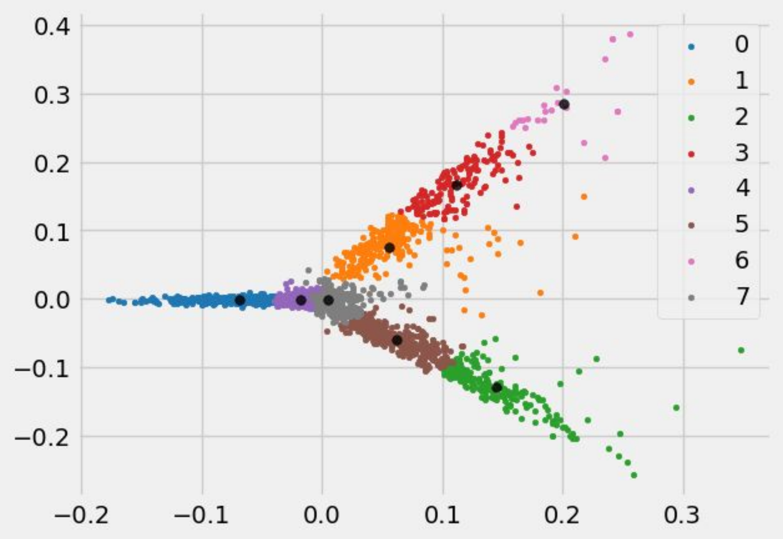


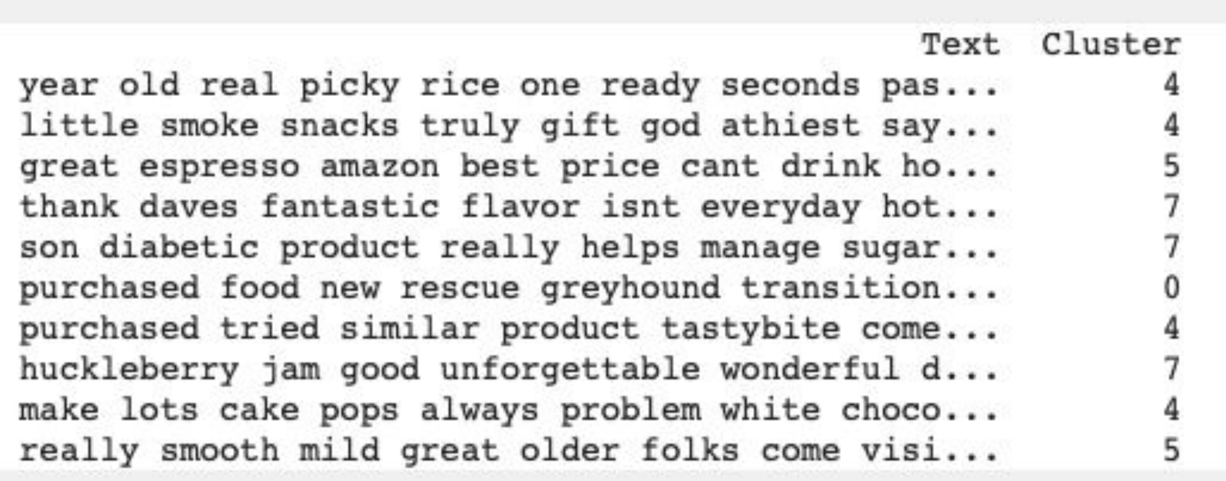
```
                                                   Text  Cluster
year old real picky rice one ready seconds pas...        1
little smoke snacks truly gift god athiest say...        1
great espresso amazon best price cant drink ho...        2
thank daves fantastic flavor isnt everyday hot...        1
son diabetic product really helps manage sugar...        1
purchased food new rescue greyhound transition...        3
purchased tried similar product tastybite come...        1
huckleberry jam good unforgettable wonderful d...        1
make lots cake pops always problem white choco...        1
really smooth mild great older folks come visi...        2
```

AGGLOMERATIVE CLUSTERING

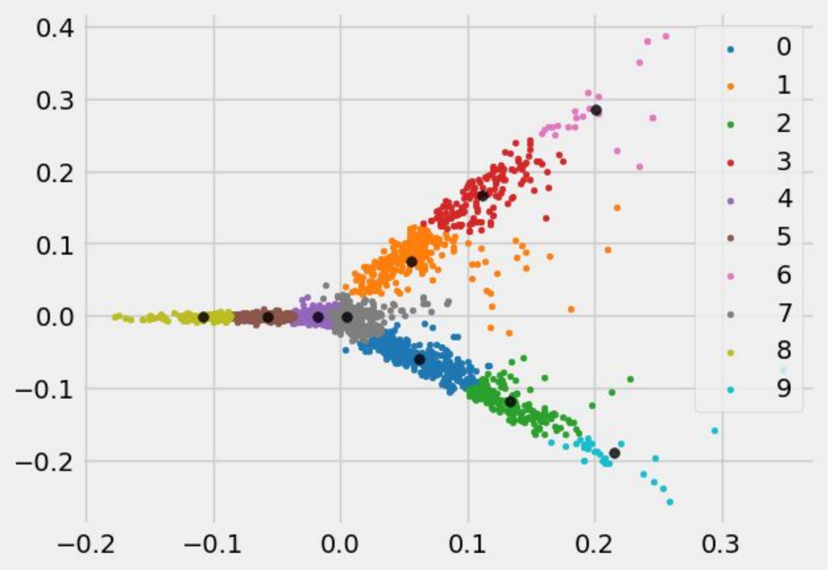


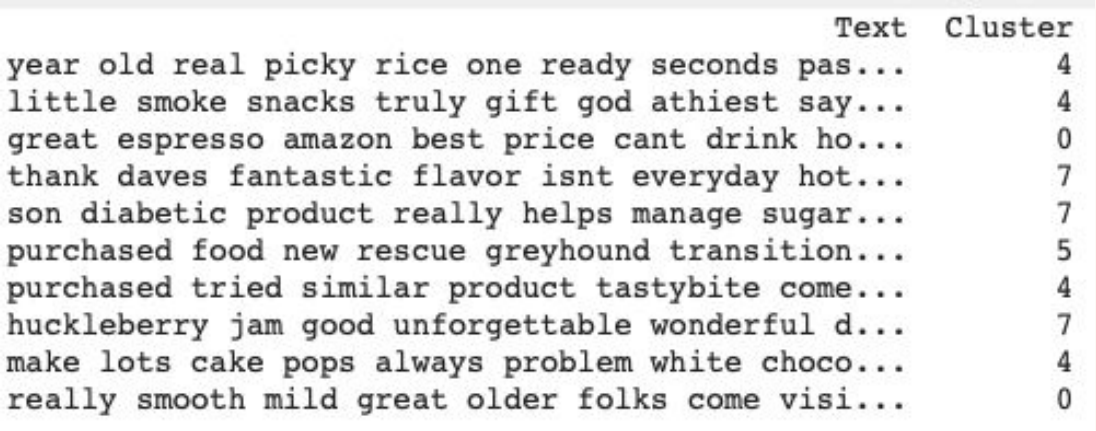

```
                                                Text  Cluster
year old real picky rice one ready seconds pas...        1
little smoke snacks truly gift god athiest say...        1
great espresso amazon best price cant drink ho...        5
thank daves fantastic flavor isnt everyday hot...        1
son diabetic product really helps manage sugar...        1
purchased food new rescue greyhound transition...        3
purchased tried similar product tastybite come...        1
huckleberry jam good unforgettable wonderful d...        1
make lots cake pops always problem white choco...        1
really smooth mild great older folks come visi...        5
```




```
                                                Text  Cluster
year old real picky rice one ready seconds pas...        4
little smoke snacks truly gift god athiest say...        4
great espresso amazon best price cant drink ho...        5
thank daves fantastic flavor isnt everyday hot...        7
son diabetic product really helps manage sugar...        7
purchased food new rescue greyhound transition...        0
purchased tried similar product tastybite come...        4
huckleberry jam good unforgettable wonderful d...        7
make lots cake pops always problem white choco...        4
really smooth mild great older folks come visi...        5
```

AGGLOMERATIVE CLUSTERING

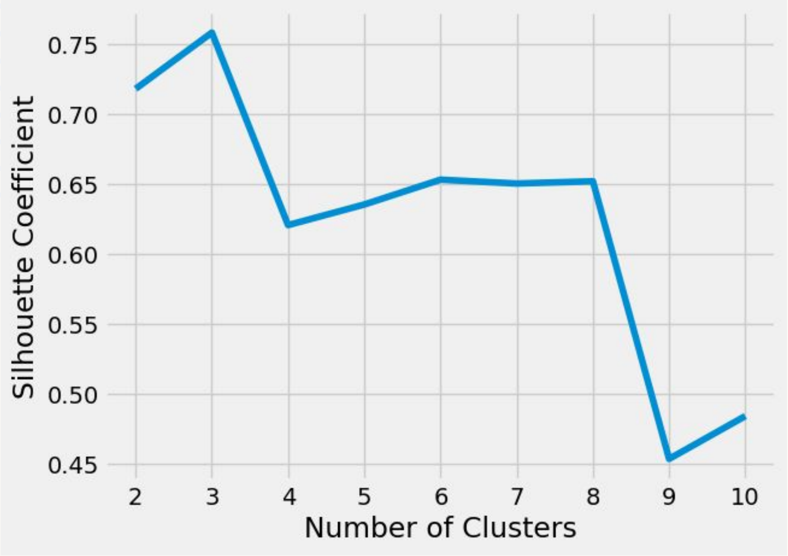


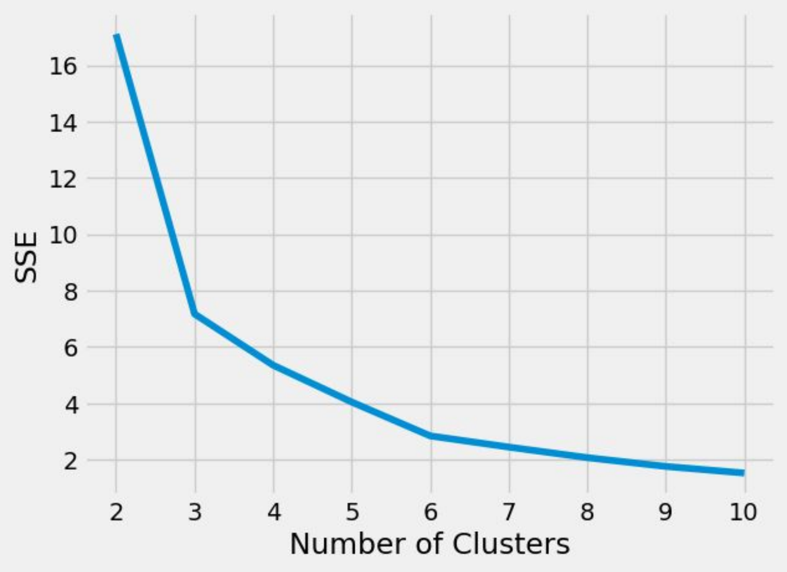


```
                                                 Text  Cluster
year old real picky rice one ready seconds pas...        4
little smoke snacks truly gift god athiest say...        4
great espresso amazon best price cant drink ho...        0
thank daves fantastic flavor isnt everyday hot...        7
son diabetic product really helps manage sugar...        7
purchased food new rescue greyhound transition...        5
purchased tried similar product tastybite come...        4
huckleberry jam good unforgettable wonderful d...        7
make lots cake pops always problem white choco...        4
really smooth mild great older folks come visi...        0
```

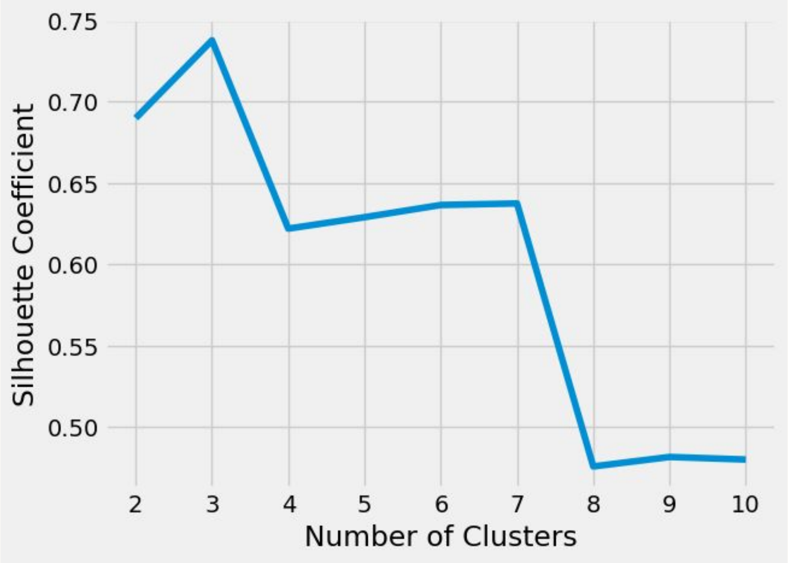
EVALUATION METRICS: Iteration 1

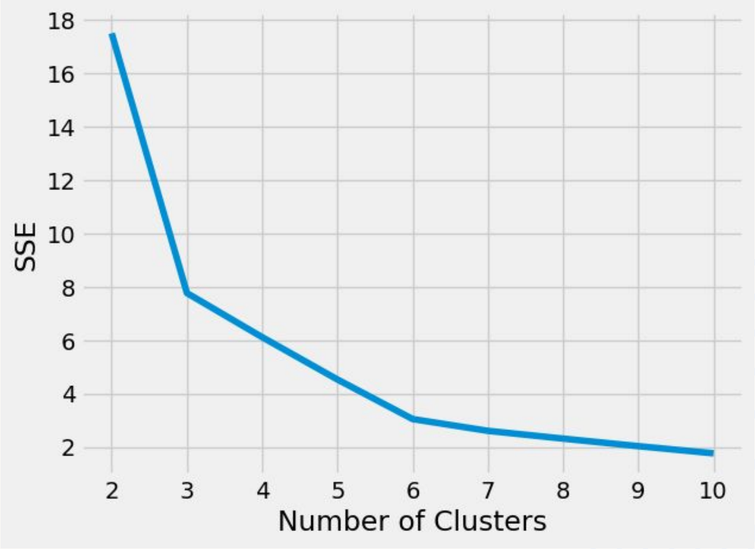
Kmeans





Agglomerative





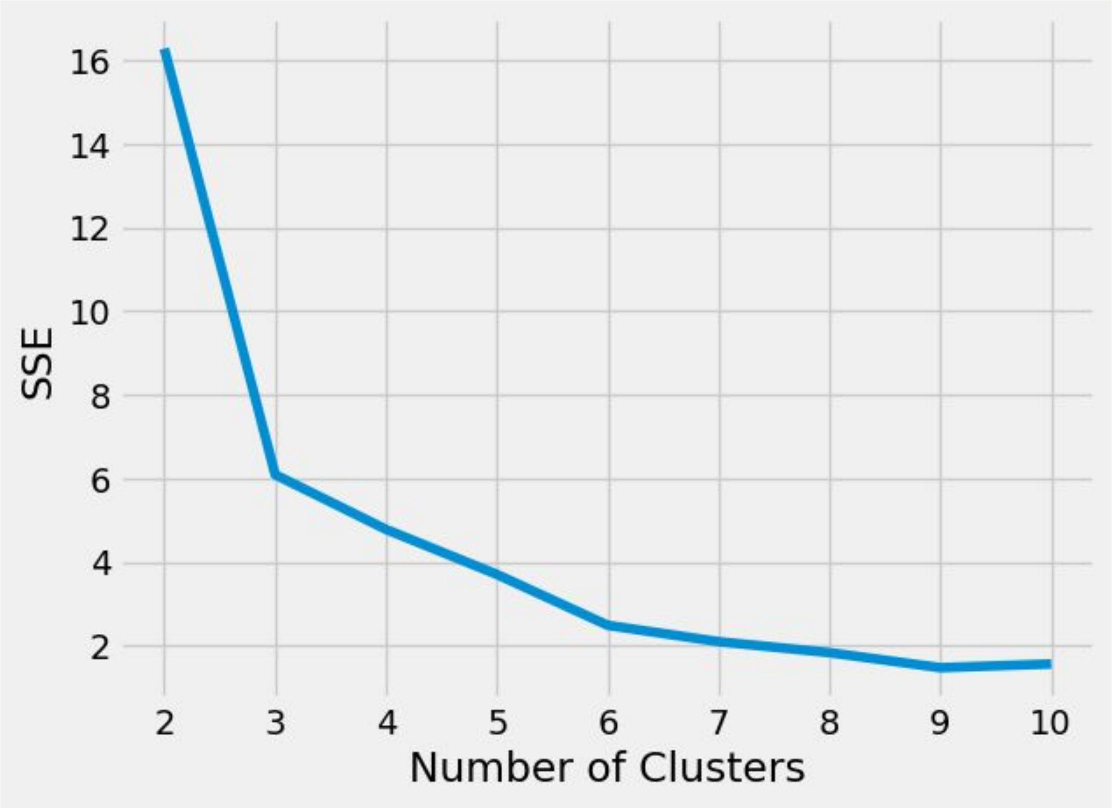
EVALUATION METRICS: Iteration 2

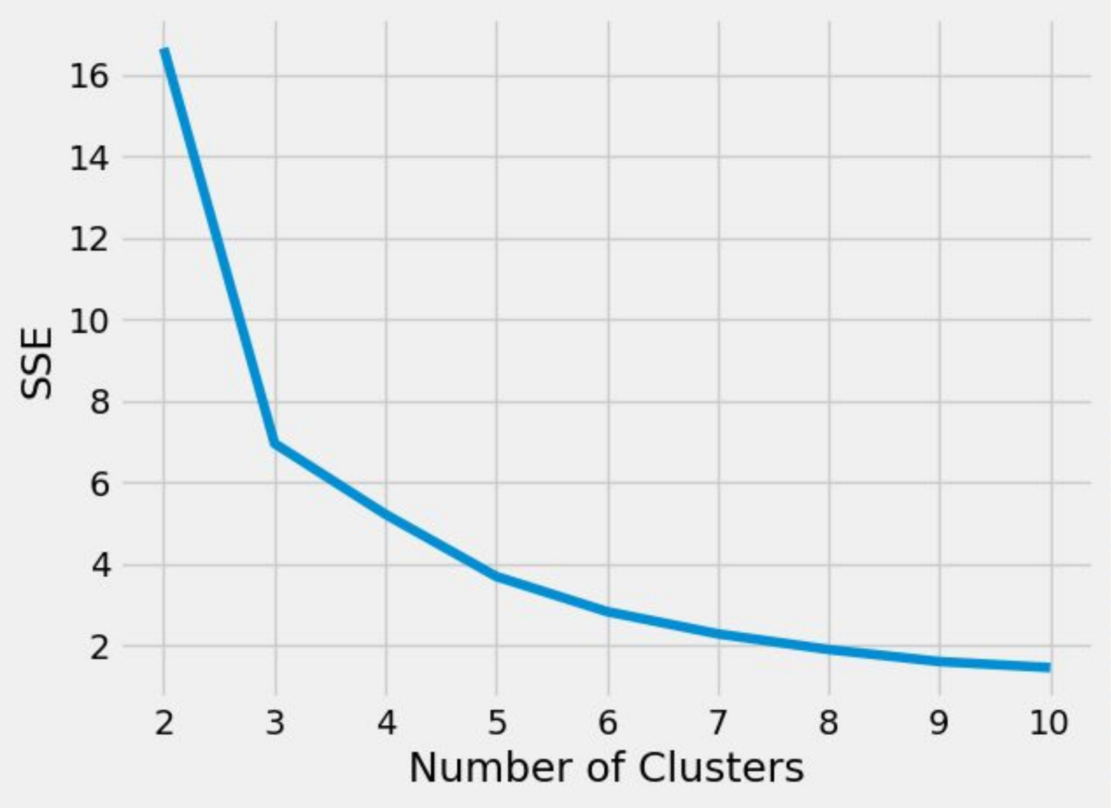
SSE Kmeans:

```
[16.27749691226183, 6.090974731032253, 4.782723881323039, 3.705984835132813, 2.489006181122323,
2.0978249969662626, 1.835643554053413, 1.4755836524332113, 1.5651367284412205]
```

SSE Agg:

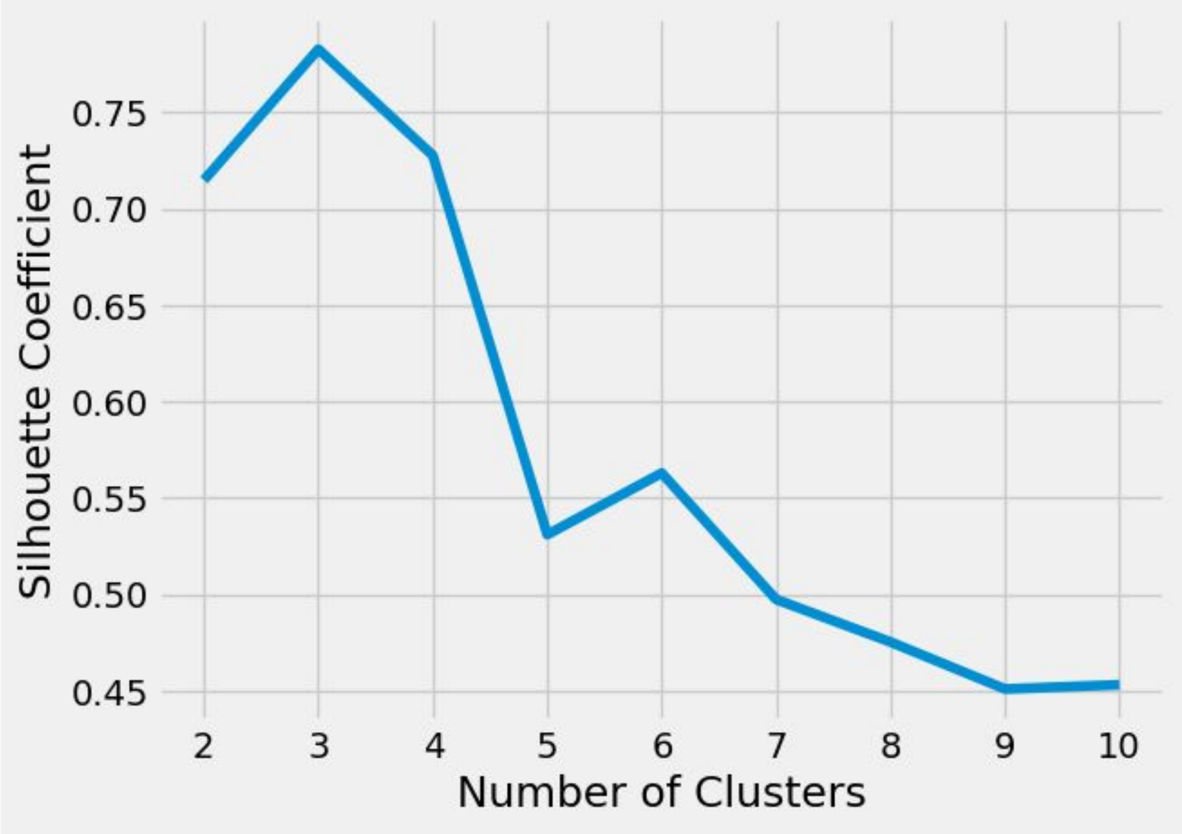
```
[16.652362545563605, 6.95132444671975, 5.214176953965005, 3.6879955806961062, 2.825471158710691,
2.2792622815619357, 1.8946710952500458, 1.5994249226224473, 1.4512095592783145]
```

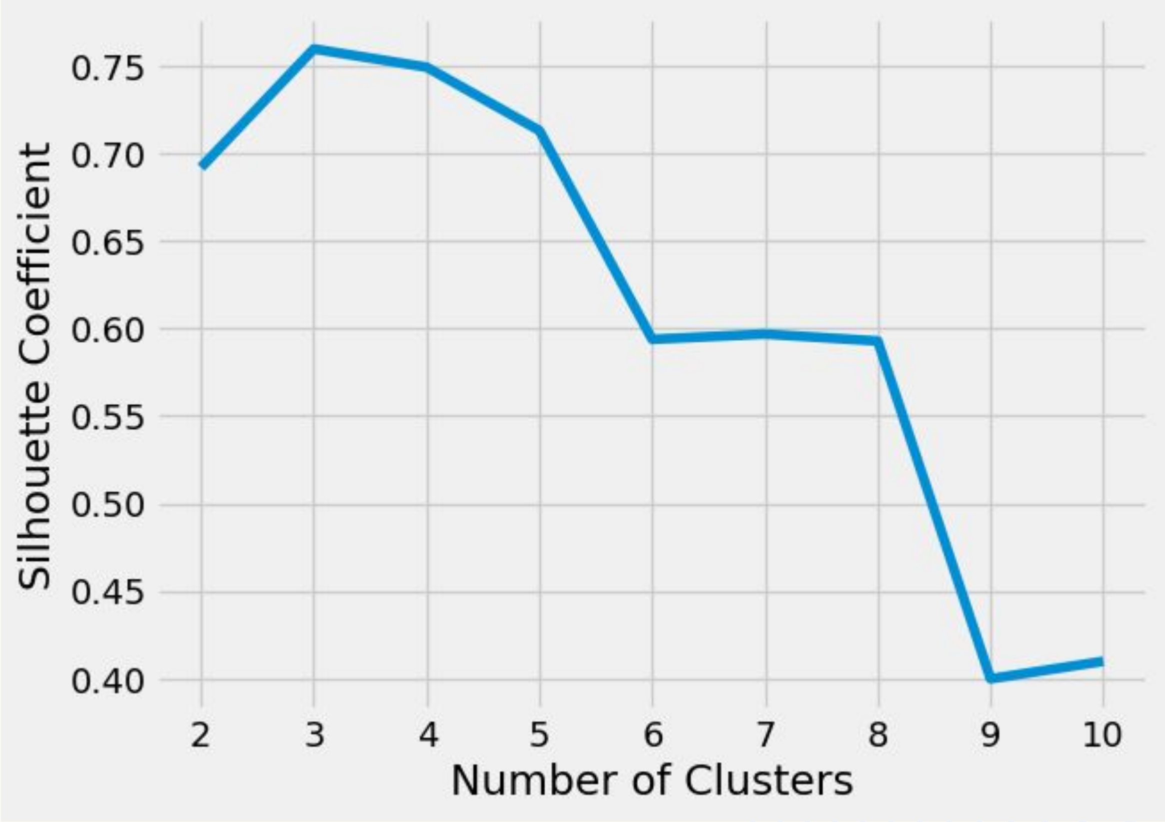


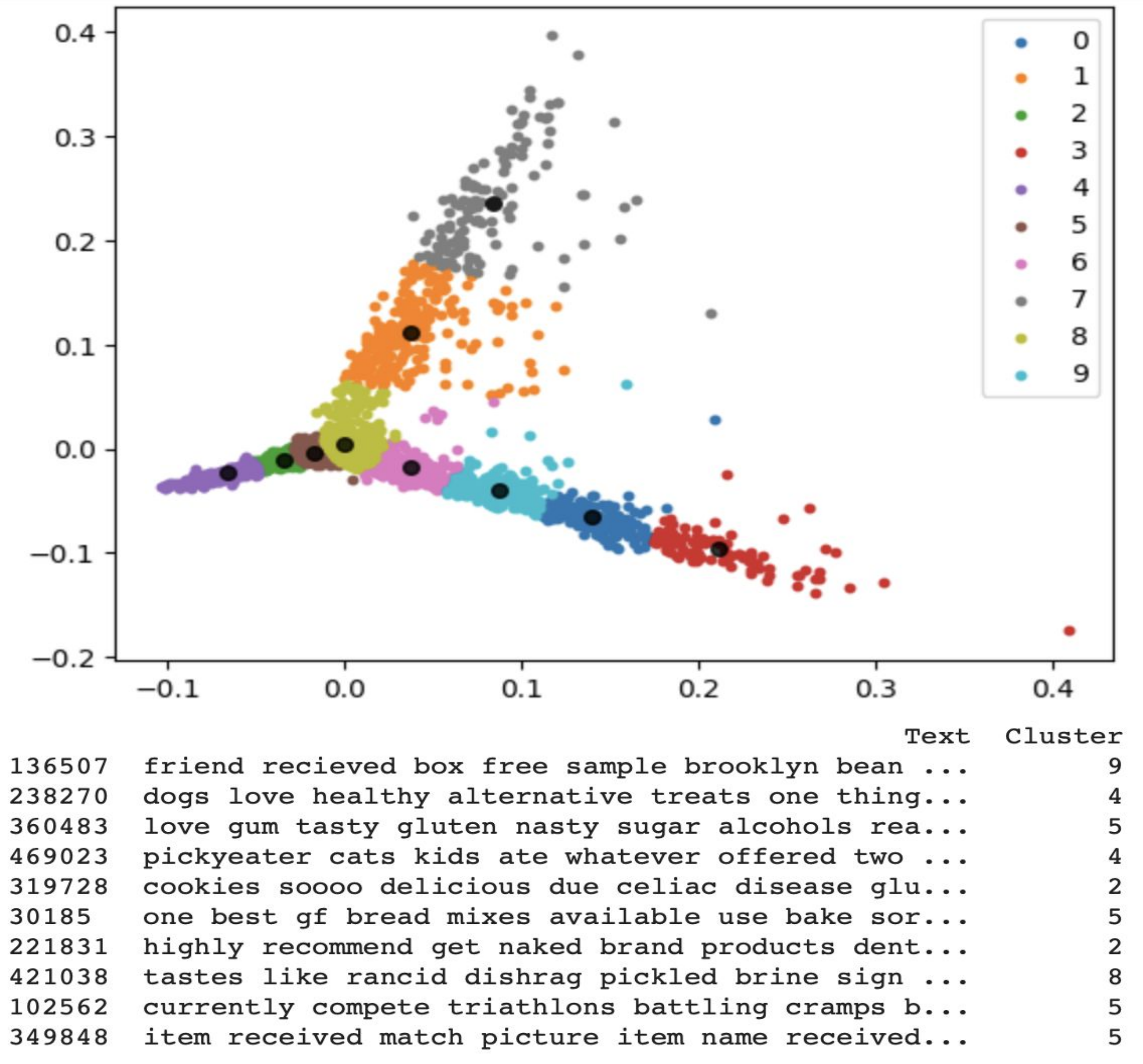


Silhouette score Kmeans:
[0.7149459014347481, 0.7828665588917019, 0.7277556995828209, 0.5310367591174248, 0.5629574522945204, 0.49740461582393647, 0.475314833811205, 0.4509377154528336, 0.4531237485766858]

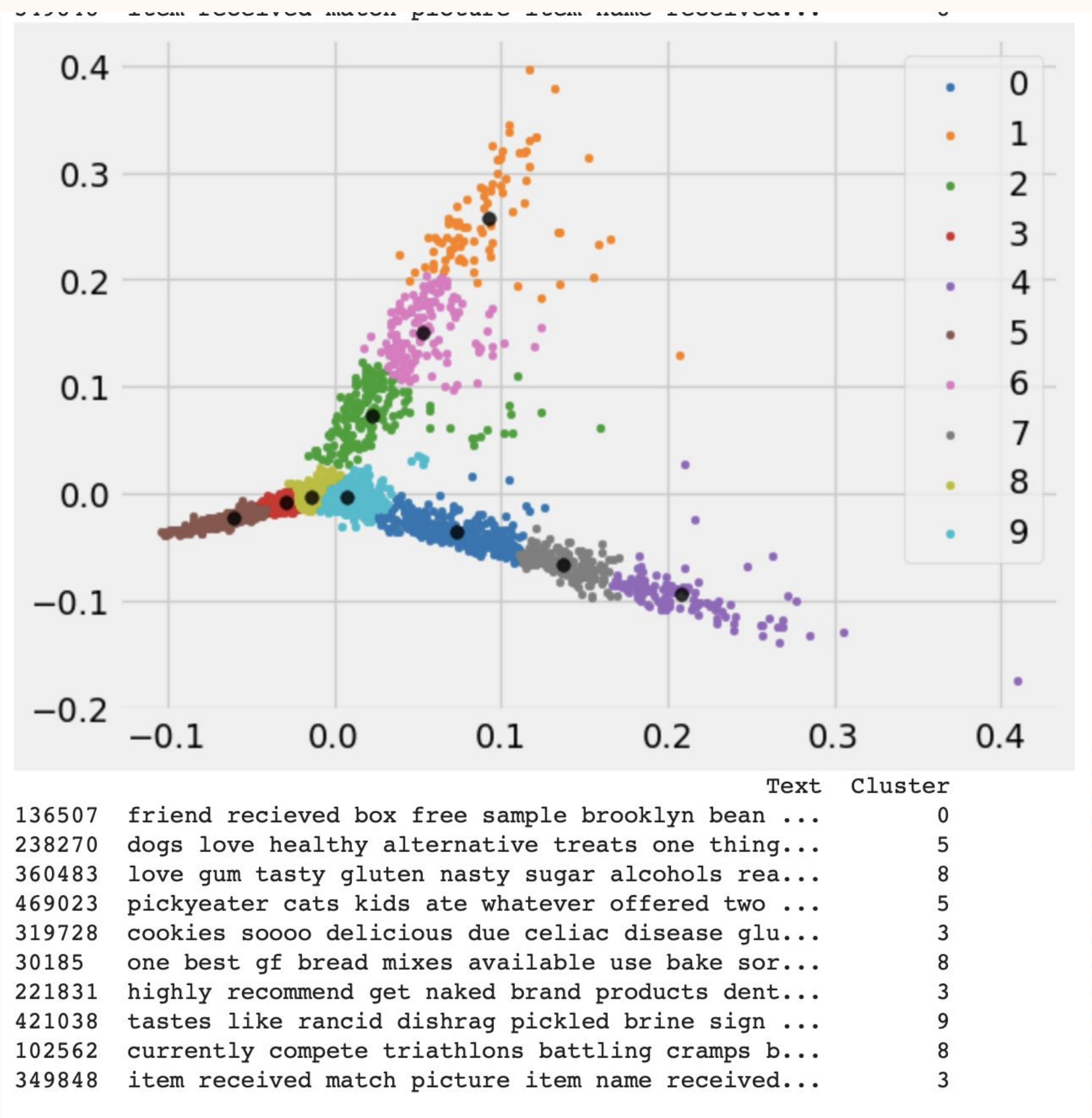
Silhouette score Agg:
[0.6918399420476093, 0.7595577282682838, 0.7491244089731679, 0.7129361809249849, 0.593718480754173, 0.5967226574434822, 0.5927606405026988, 0.39987115002227397, 0.4098520154988161]







```
                                                    Text  Cluster
136507  friend recieved box free sample brooklyn bean ...        9
238270  dogs love healthy alternative treats one thing...        4
360483  love gum tasty gluten nasty sugar alcohols rea...        5
469023  pickyeater cats kids ate whatever offered two ...        4
319728  cookies soooo delicious due celiac disease glu...        2
30185   one best gf bread mixes available use bake sor...        5
221831  highly recommend get naked brand products dent...        2
421038  tastes like rancid dishrag pickled brine sign ...        8
102562  currently compete triathlons battling cramps b...        5
349848  item received match picture item name received...        5
```



```
                                                    Text  Cluster
136507  friend recieved box free sample brooklyn bean ...        0
238270  dogs love healthy alternative treats one thing...        5
360483  love gum tasty gluten nasty sugar alcohols rea...        8
469023  pickyeater cats kids ate whatever offered two ...        5
319728  cookies soooo delicious due celiac disease glu...        3
30185   one best gf bread mixes available use bake sor...        8
221831  highly recommend get naked brand products dent...        3
421038  tastes like rancid dishrag pickled brine sign ...        9
102562  currently compete triathlons battling cramps b...        8
349848  item received match picture item name received...        3
```

THANK YOU

References -

1 .https://www.kaggle.com/code/lazaro97/clustering-sentiment-analysis