Data Science Applications

Clustering Assignment

Students: Group 7

1. Gehad Hisham Hassan Abdelghany.
2. Kareem Atif Mohamed Bakli.
3. Kareem Khaled Waly.
4. Mostafa Nofal.

Overview:

The goal of this project is to produce clusters from our books, compare the different models we used; analyze each one and come out the best model which is most efficient in this problem.

The dataset:

1. We imported 5 books with different authors and different genres from Gutenberg.

- Austen-emma
- Bible-kjv
- Chesterton-brown
- Shakespeare-caesar
- Blake-poems

	Title	Text
0	austen-emma	[Emma by Jane Austen 1816]\n\nVOLUME I\n\nCHAP...
1	bible	[The King James Bible]\n\nThe Old Testament of...
2	chesterton-brown	[The Wisdom of Father Brown by G. K. Chesterto...
3	shakespeare-caesar	[The Tragedie of Julius Caesar by William Shak...
4	blake-poems	[Poems by William Blake 1789]\n\n \nSONGS OF I...

2. Created 200 partitions of each book text, each partition contains 150 words.
3. We come out with a DataFrame contains:

- Partitions columns
- The label of the book column

	0	1	2	3	4	5	6	7	8	9
0	[passed, thoroughly, distressed, state, mind, ...	[shall, scold, leave, reflections, trust, flat...	[room, want, upon, dear, said, shall, laid, up...	[happen, meet, well, miss, woodhouse, like, ex...	[spent, anywhere, home, confers, making, agree...	[satisfy, woman, good, sense, quick, feelings,...	[companion, emma, early, foreseen, useful, mig...	[know, family, name, sure, cried, father, bate...	[confined, foot, pace, much, anger, would, des...	[proved, erroneous, friends, deserted, pleasur...
1	[neither, regardeth, crying, driver, range, mo...	[friend, lazarus, sleepeth, awake, sleep, said...	[fury, upon, thee, accomplish, mine, anger, up...	[took, every, clean, beast, every, clean, fowl...	[deliver, depart, city, woman, said, unto, joa...	[lord, pitched, rephidim, water, people, drink...	[strengthened, weak, hands, words, upholden, f...	[truth, expedient, away, away, comforter, come...	[people, abner, saying, answerest, thou, abner...	[peace, thee, friends, salute, thee, greet, fr...
2	[priest, fancy, visiting, parish, cobhole, goi...	[easy, walk, rough, stones, green, slippery, s...	[already, laying, white, tables, insignia, ear...	[burglar, simply, blind, said, cray, stubborml...	[front, house, moments, later, roar, popular, ...	[take, ropes, asked, girl, stubbornly, hood, r...	[bronzed, also, sake, assure, story, cause, cu...	[silver, shouldn, pleased, think, tossed, tous...	[smooth, brow, became, suddenly, constricted, ...	[think, turned, little, instrument, water, sho...
3	[gods, incenses, send, destruction, thing, won...	[stay, home, feare, caesar, shall, danger, kno...	[greeuous, fault, greeuously, hath, caesar, an...	[faint, lucius, commend, lord, merry, come, ag...	[caesar, addrest, presse, neere, second, caska...	[dignities, onely, patient, till, haue, appeas...	[caesar, addrest, presse, neere, second, caska...	[seem, open, brest, heauen, present, selfe, eu...	[good, messala, messa, dyed, master, strato, s...	[dignities, onely, patient, till, haue, appeas...
4	[head, bowd, weeping, infant, life, exhald, mi...	[life, spring, fades, lotus, water, fade, chil...	[tears, didst, bind, nostrils, eyes, ears, did...	[grove, love, care, round, golden, tent, like,...	[life, spring, fades, lotus, water, fade, chil...	[heard, among, rushes, dank, weeping, weeping,...	[sees, whose, ears, heard, holy, word, walked,...	[upon, heath, smiled, among, winter, snow, clo...	[head, bowd, weeping, infant, life, exhald, mi...	[poems, william, blake, songs, innocence, expe...

	...	192	193	194	195	196	197	198	199	Label	Author
0	...	[everywhere, telling, grandmama, jane, little,...	[weston, score, elton, kindly, sent, jane, not...	[gainer, well, said, emma, willing, pass, want...	[suggesting, attending, happy, enjoyment, futu...	[gentility, emma, formerly, ready, vouch, like...	[always, first, company, others, well, papa, e...	[determined, look, confident, often, listening...	[circumstances, knightley, really, wished, mar...	a	austen-emma
1	...	[vessels, house, lord, made, zedekiah, brother...	[everlasting, name, deep, horse, wilderness, s...	[shall, upon, ground, seven, women, shall, tak...	[take, pity, none, comforters, found, none, ga...	[unto, name, declare, love, wherewith, thou, h...	[jehoiakim, gave, silver, gold, pharaoh, taxed...	[days, shall, fathers, eaten, sour, grape, chi...	[bondwoman, shall, heir, freewoman, brethren, ...	b	bible
2	...	[meet, chauvinist, officer, dubosc, sitting, s...	[western, bidden, take, butterfly, lugubrious,...	[beings, used, inappropriate, things, accustom...	[legal, records, newspapers, lawsuit, threaten...	[strong, massive, head, sort, head, door, chau...	[mistake, italy, sixteenth, century, tuscans, ...	[giant, brandishing, eiffel, tower, shock, sha...	[todhunter, reiterated, hood, quietly, happen,...	c	chesterton-brown
3	...	[brutus, trouble, brut, haue, beene, howre, aw...	[know, plaine, blunt, loue, friend, know, full...	[haue, climb, walles, battlements, towres, win...	[sennit, exeunt, caesar, traine, cask, cloake,...	[shout, flourish, another, generall, shout, be...	[place, world, furnish, well, flesh, blood, ap...	[tarrying, heere, farewell, volumnius, strato,...	[haue, thought, good, metellus, along, loues, ...	d	shakespeare-caesar
4	...	[burning, bright, forest, night, immortal, han...	[grove, love, care, round, golden, tent, like,...	[poems, william, blake, songs, innocence, expe...	[dream, weave, shade, angel, guarded, emmet, l...	[poems, william, blake, songs, innocence, expe...	[sees, whose, ears, heard, holy, word, walked,...	[dolors, lamentations, waiting, beside, dewy, ...	[dolors, lamentations, waiting, beside, dewy, ...	e	blake-poems

5 rows × 202 columns

Preprocess the data:

- Converted the text to lower case.
- Removed any special characters.
- Used RegexpTokenizer to tokenize the text.
- Created our stop words list and removed from our text.
- Remove single char, and chars with size 2.
- Label Encoder.

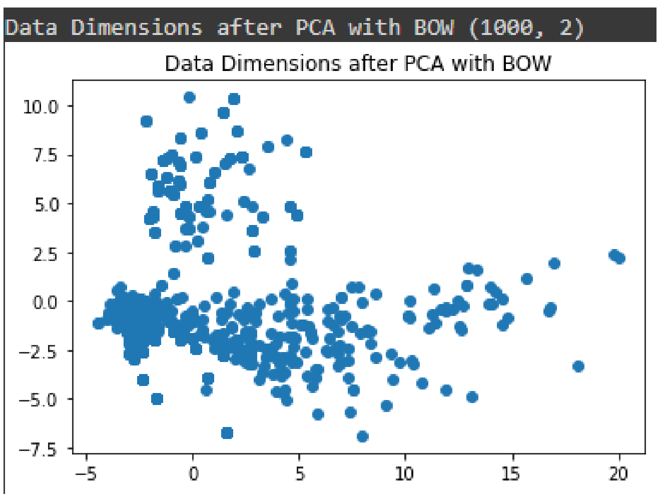
	Label	Author	value
0	0	austen-emma	[passed, thoroughly, distressed, state, mind, ...
1	1	bible	[neither, regardeth, crying, driver, range, mo...
2	2	chesterton-brown	[priest, fancy, visiting, parish, cobhole, goi...
3	3	shakespeare-caesar	[gods, incenses, send, destruction, thing, won...
4	4	blake-poems	[head, bowd, weeping, infant, life, exhald, mi...
...	...	...	...

Text Transformations:

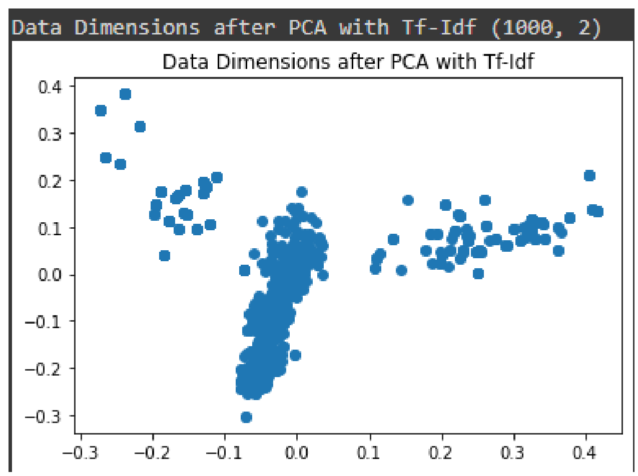
- BOW
- TF-IDF
- LDA
- Word-Embedding.

We used **PCA** to reduce the number of features of every one of the four vectorizers to plot them in 2d.

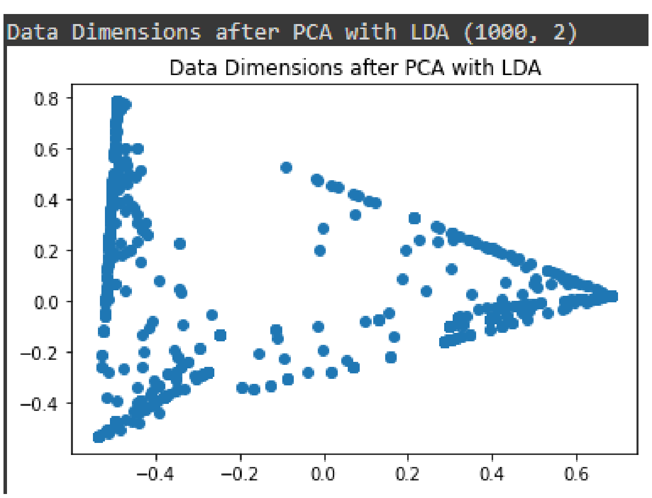
BOW



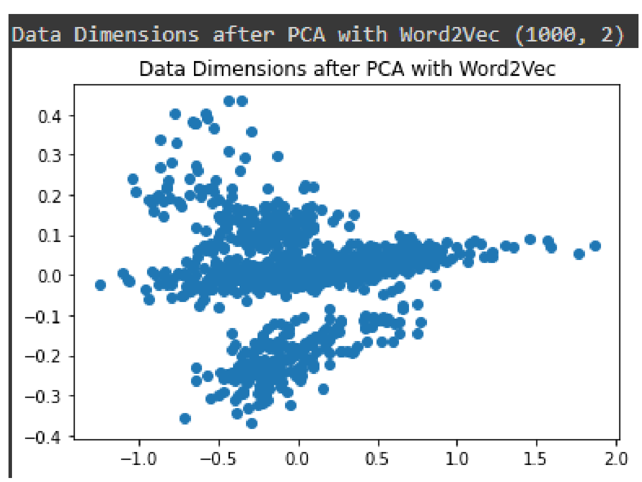
TF-IDF



LDA



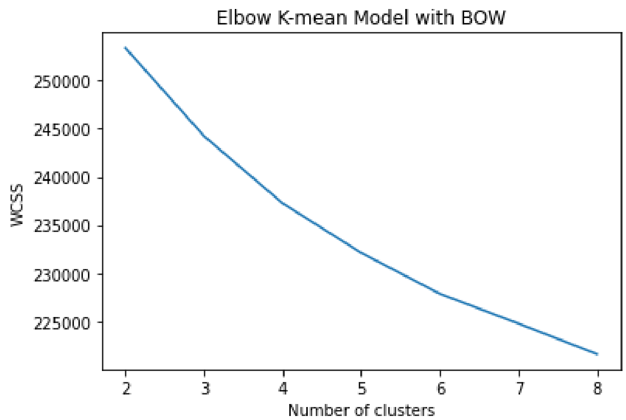
Word2Vec



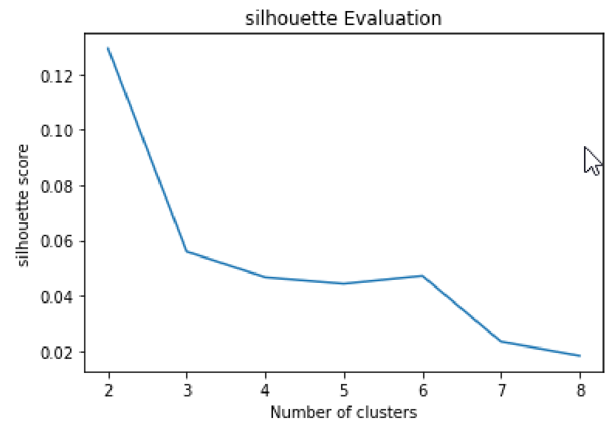
Clustering algorithms:

K-means with BOW:

Best k from 2 to 8



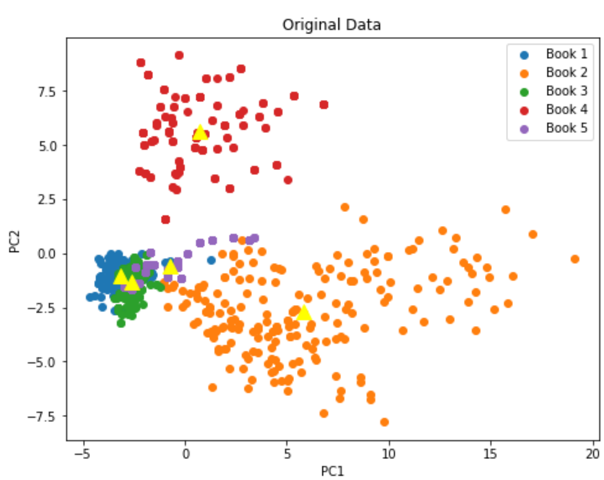
Best silhouette from 2 to 8

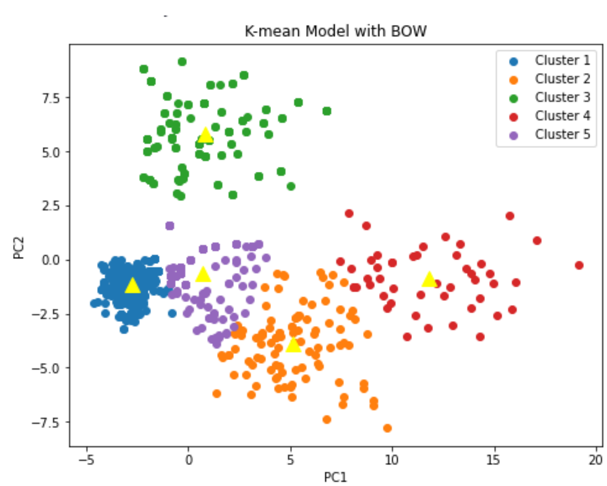


Evaluations for 5 number of clusters

```
Kappa for the model at n_clusters= 5 is  0.71
Best Value for n cluster is  = 2 The average silhouette_score : 0.1294
For n_clusters = 5 The silhouette_score : 0.045
For n_clusters = 5 The homogeneity_score : 0.7338
For n_clusters = 5 The v_measure_score : 0.8093
```

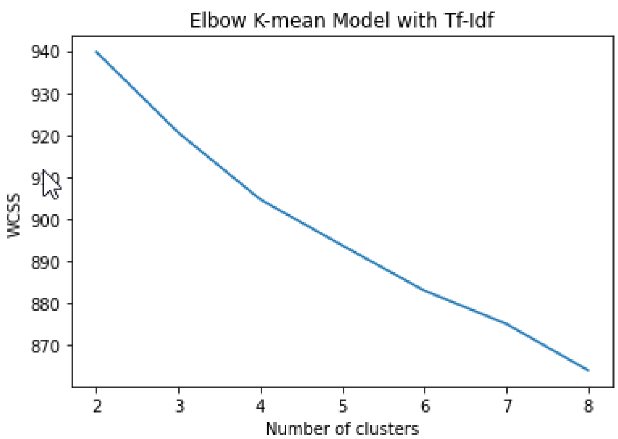
Plot original data vs K-mean model with BOW



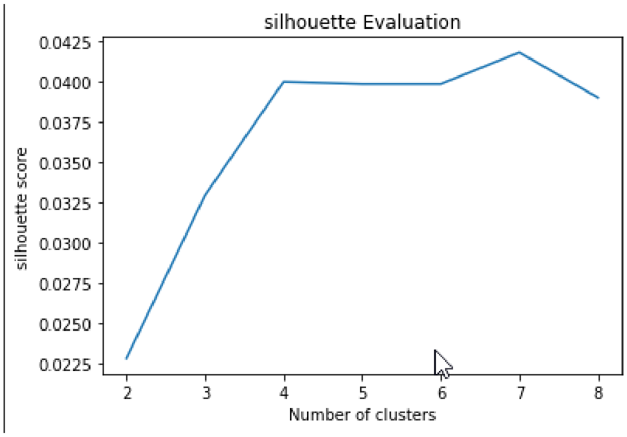


K-means with TF-IDF:

Best k from 2 to 8



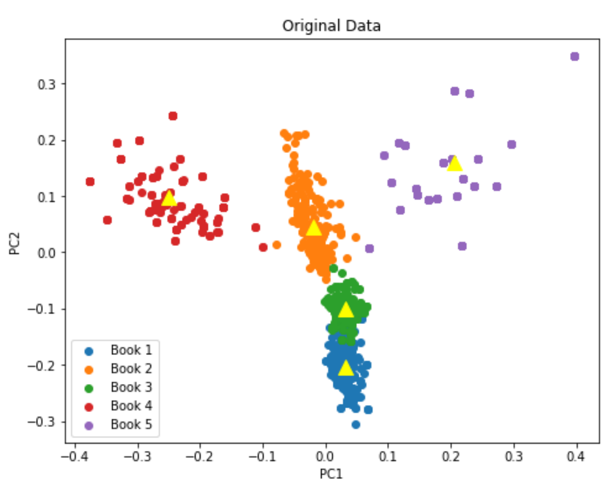
Best silhouette from 2 to 8

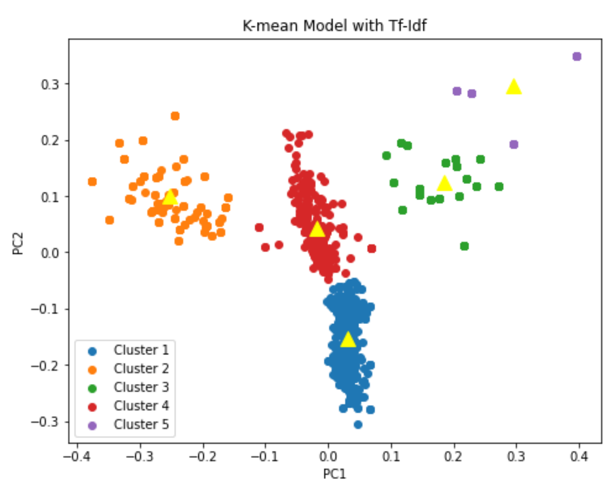


Evaluations for 5 number of clusters

```
Kappa for the model at n_clusters= 5 is  0.7488
Best Value for n cluster is  = 7 The average silhouette_score : 0.0418
For n_clusters = 5 The silhouette_score : 0.04
For n_clusters = 5 The homogeneity_score : 0.8234
For n_clusters = 5 The v_measure_score : 0.9012
```

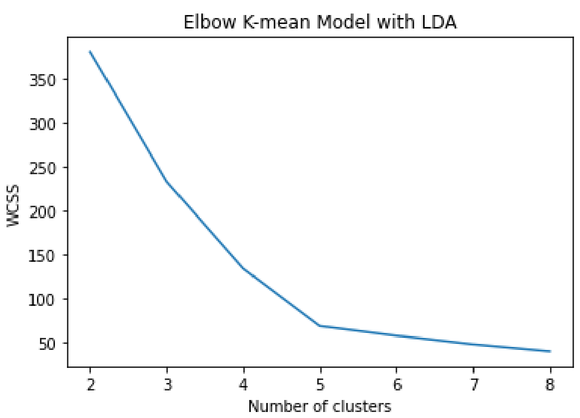
Plot original data vs K-mean model with Tf-IDF



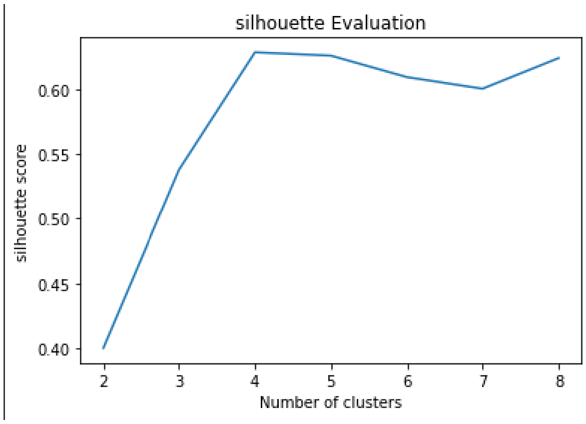


K-means with LDA:

Best k from 2 to 8



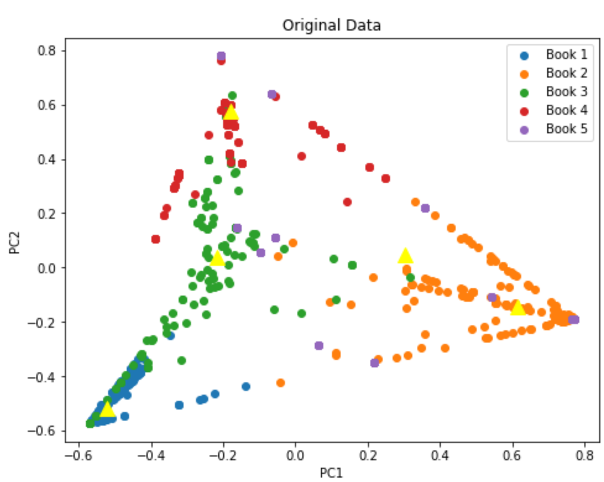
Best silhouette from 2 to 8

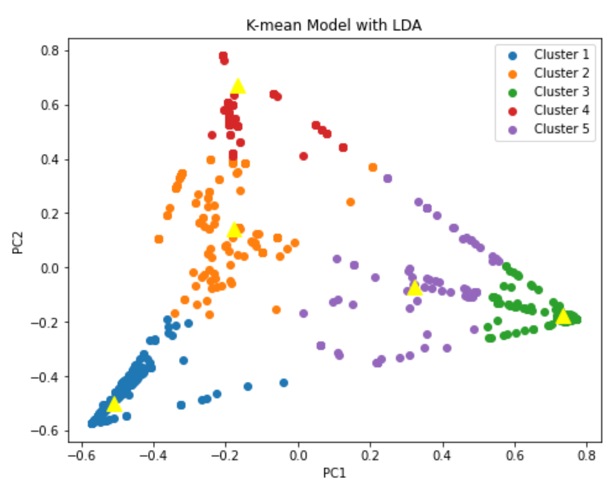


Evaluations for 5 number of clusters

```
Kappa for the model at n_clusters= 5 is  0.66
Best Value for n cluster is  = 4 The average silhouette_score : 0.6283
For n_clusters = 5 The silhouette_score : 0.6257
For n_clusters = 5 The homogeneity_score : 0.5878
For n_clusters = 5 The v_measure_score : 0.6332
```

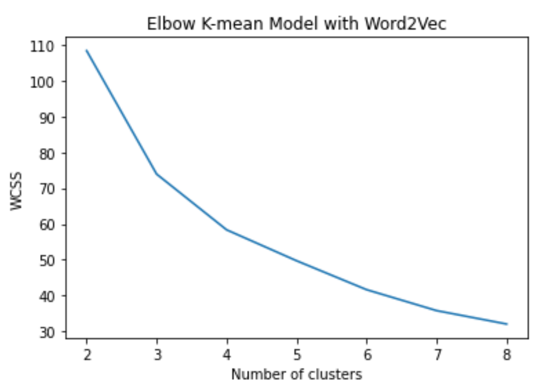
Plot original data vs K-mean model with LDA



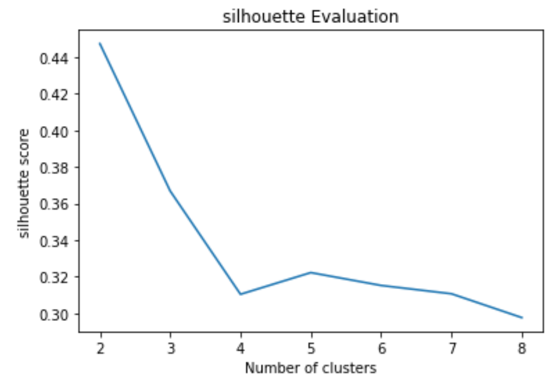


K-means with Word2Vec:

Best k from 2 to 8



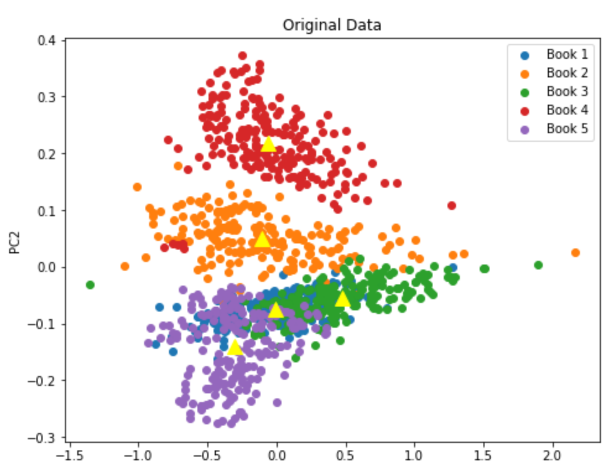
Best silhouette from 2 to 8

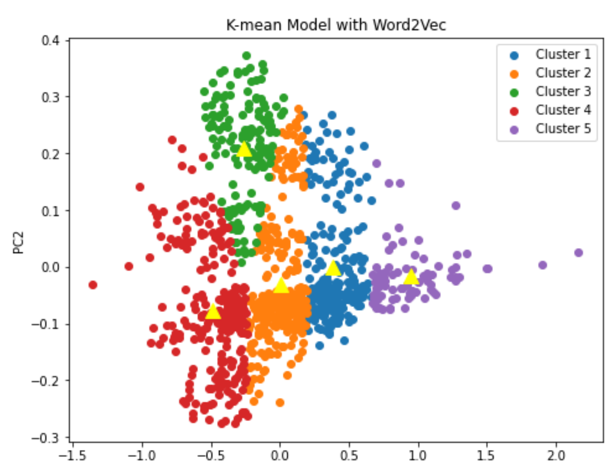


Evaluations for 5 number of clusters

```
Kappa for the model at n_clusters= 5 is  0.3712
Best Value for n cluster is  = 2 The average silhouette_score : 0.4473
For n_clusters = 5 The silhouette_score : 0.324
For n_clusters = 5 The homogeneity_score : 0.2648
For n_clusters = 5 The v_measure_score : 0.2902
```

Plot original data vs K-mean model with **Word2Vec**

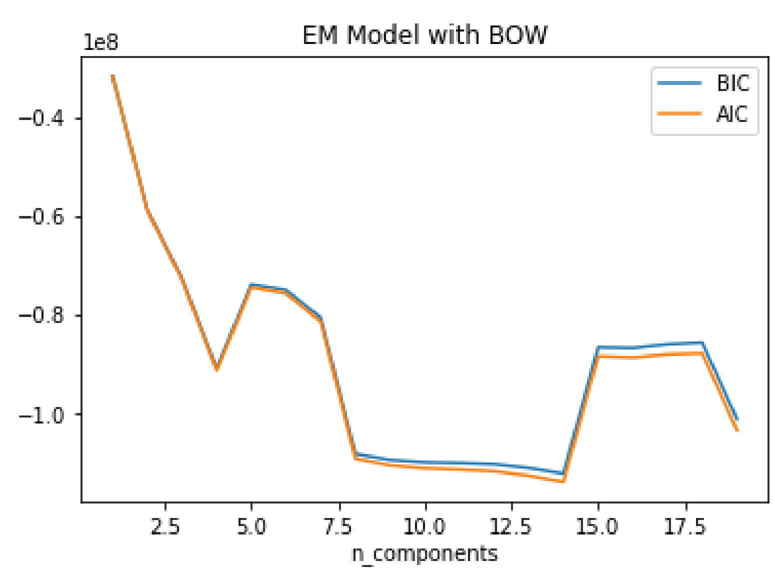




Best transformarion with K-means algorithm is Tf-Idf

- Kappa for the model at n_clusters= 5 is 0.745
- Best Value for n cluster is = 5
  The average silhouette_score : 0.0438
- For n_clusters = 5 The silhouette_score : 0.0406
- For n_clusters = 5 The homogeneity_score : 0.8138
- For n_clusters = 5 The v_measure_score : 0.8913

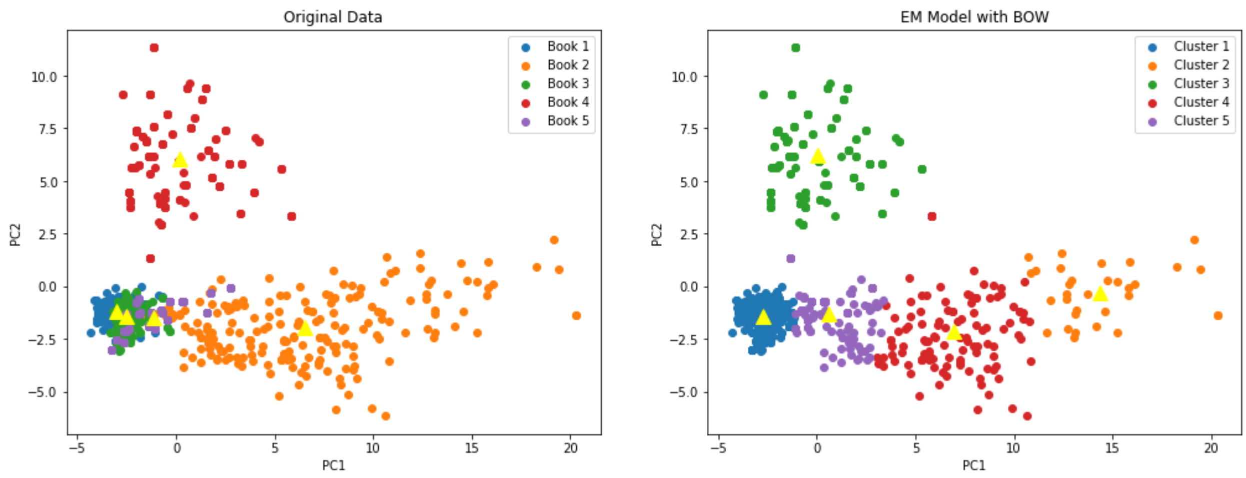
Expectation Maximization with BOW:



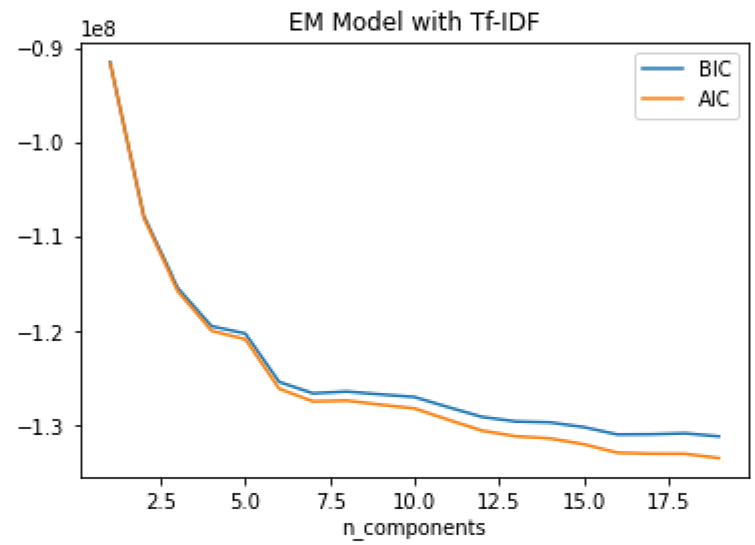
Evaluations for 5 number of clusters

```
Kappa for the EM model at number of Cluster 5 is  0.725
For n_clusters = 5 The average silhouette_score : 0.0339
For n_clusters = 5 The average homogeneity_score : 0.7348
For n_clusters = 5 The v_measure_score : 0.7941
```

Plot original data vs EM model with **BOW**



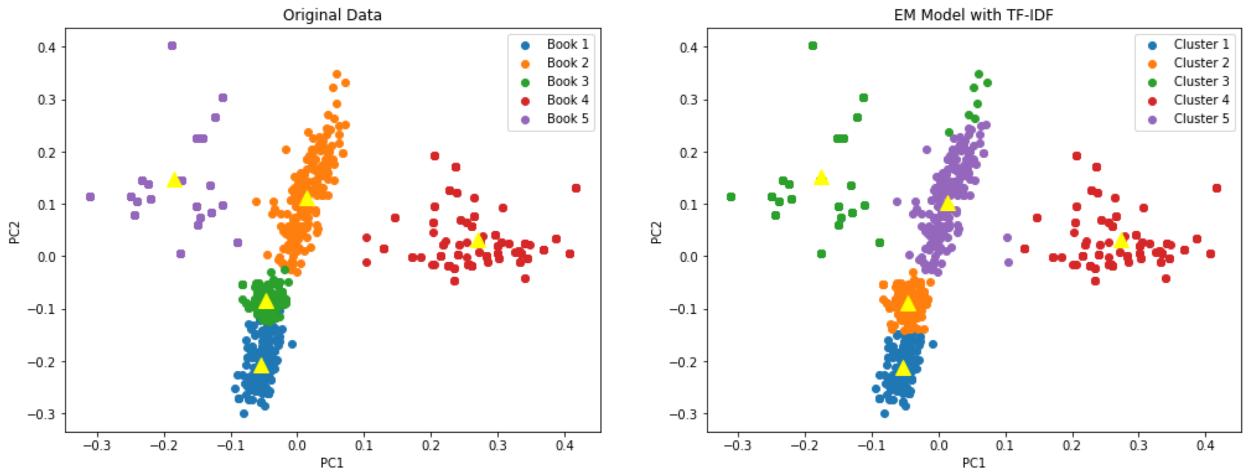
Expectation Maximization with Tf-IDF:



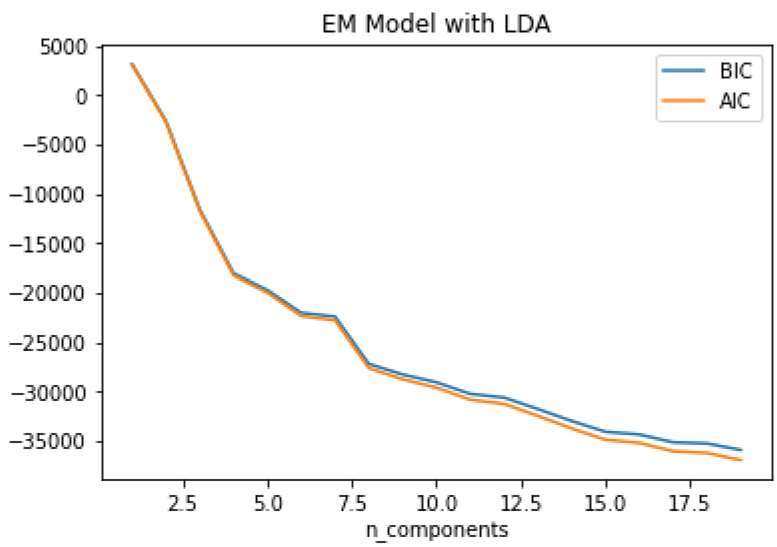
Evaluations for 5 number of clusters

```
Kappa for the EM model at number of Cluster 5 is  0.7463
For n_clusters = 5 The average silhouette_score : 0.0406
For n_clusters = 5 The average homogeneity_score : 0.8167
For n_clusters = 5 The v_measure_score : 0.8944
```

Plot original data vs EM model with **Tf-IDF**



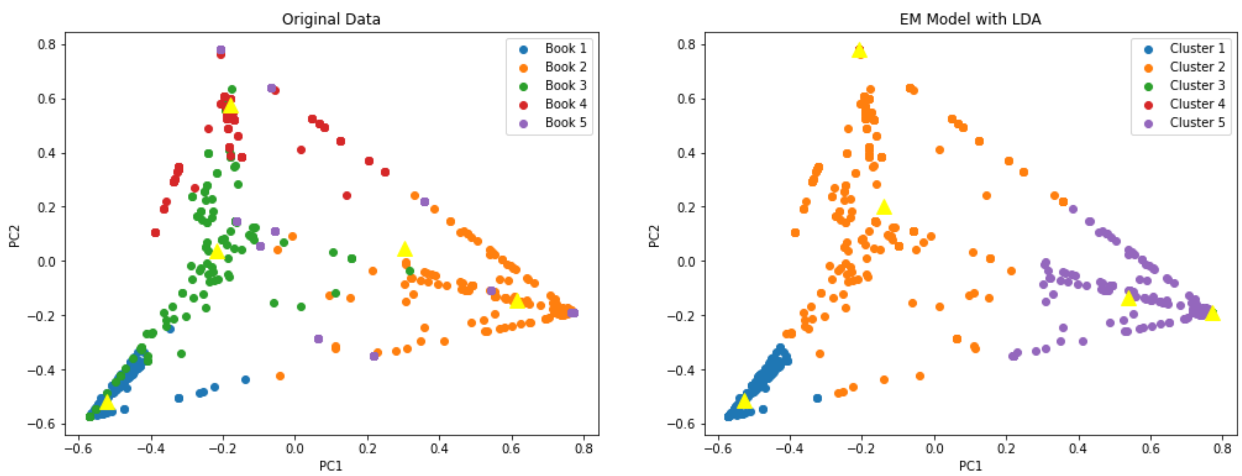
Expectation Maximization with LDA:



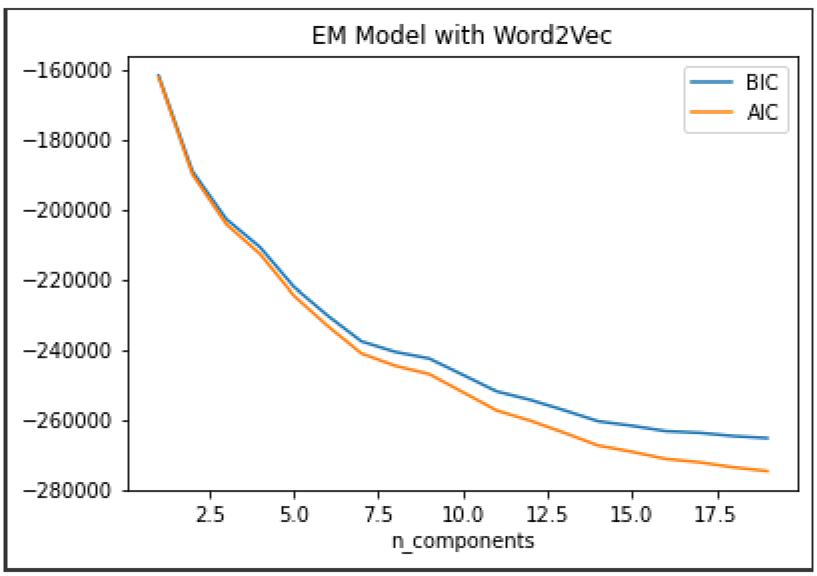
Evaluations for 5 number of clusters

```
Kappa for the EM model at number of Cluster 5 is  0.305
For n_clusters = 5 The average silhouette_score : 0.1537
For n_clusters = 5 The average homogeneity_score : 0.2492
For n_clusters = 5 The v_measure_score : 0.3033
```

Plot original data vs EM model with **LDA**



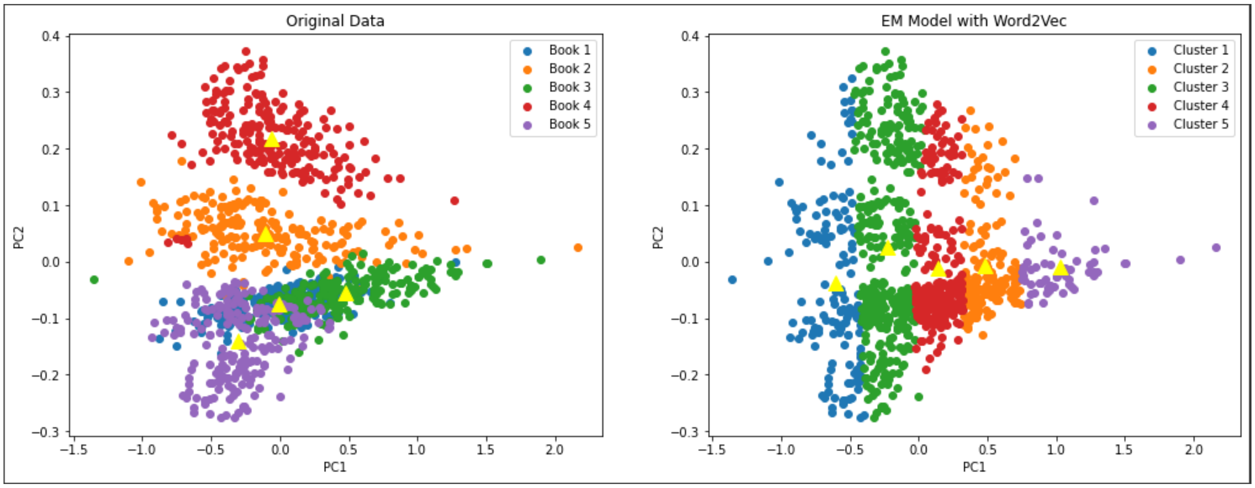
Expectation Maximization with Word2Vec:



Evaluations for 5 number of clusters

```
Kappa for the EM model at number of Cluster 5 is  0.37
For n_clusters = 5 The average silhouette_score : 0.3016
For n_clusters = 5 The average homogeneity_score : 0.2629
For n_clusters = 5 The v_measure_score : 0.2903
```

Plot original data vs EM model with **Word2Vec**

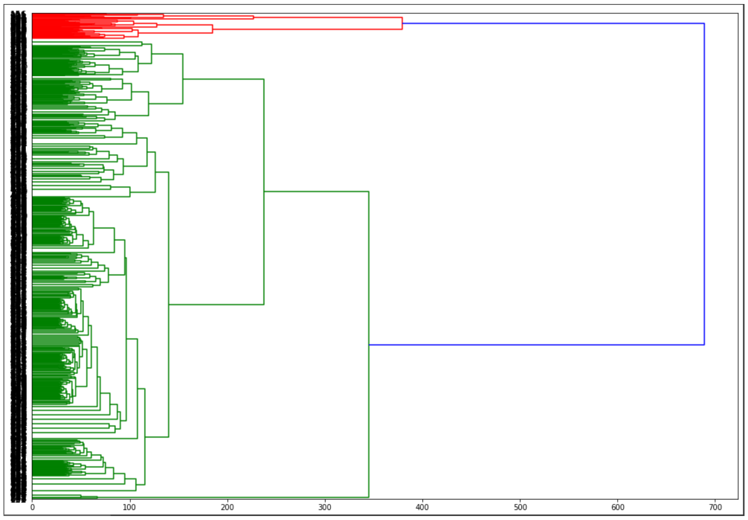


Best transformarion with EM is Tf-Idf

- Kappa for the EM model at number of Cluster 5 is 0.7463
- For n_clusters = 5 The average silhouette_score : 0 4
- For n_clusters = 5 The homogeneity_score : 0.8167
- For n_clusters = 5 The v_measure_score : 0.8944

Hierarchial model with BOW:

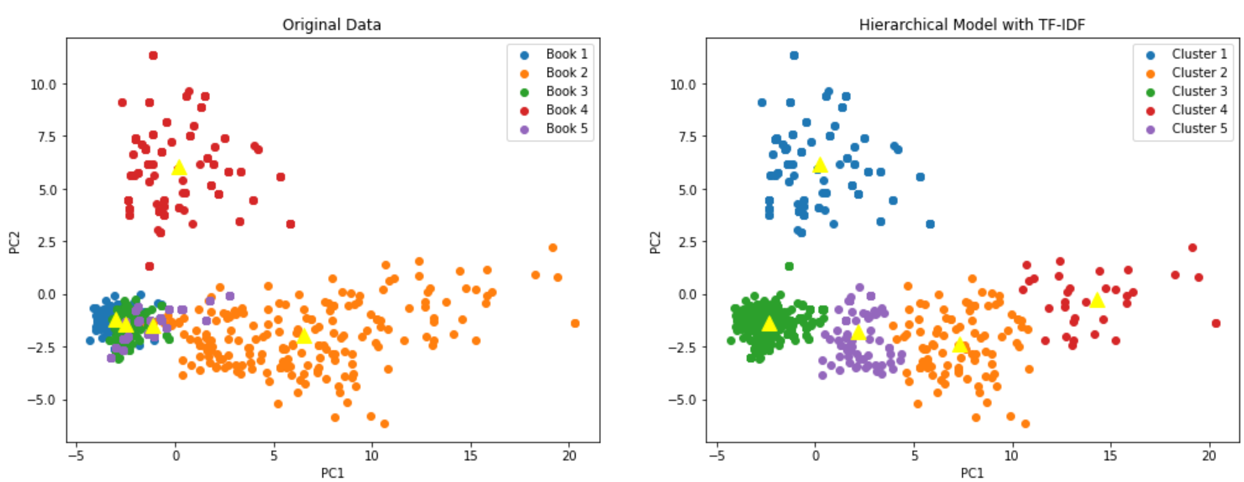
Dendrogram plot



Evaluations for 5 number of clusters

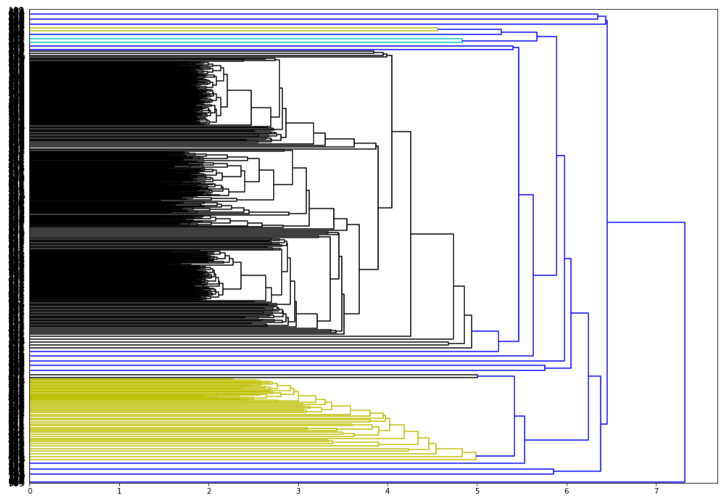
```
Kappa for the Hierarchy model at number of Cluster 5 is  0.6875
For n_clusters = 5  Silhouette Coefficient is: 0.0436
For n_clusters = 5  The average homogeneity_score : 1
For n_clusters = 5 The v_measure_score : 0.7779
```

Plot original data vs hierarchial model with **BOW.**



Hierarchial model with TF-IDF:

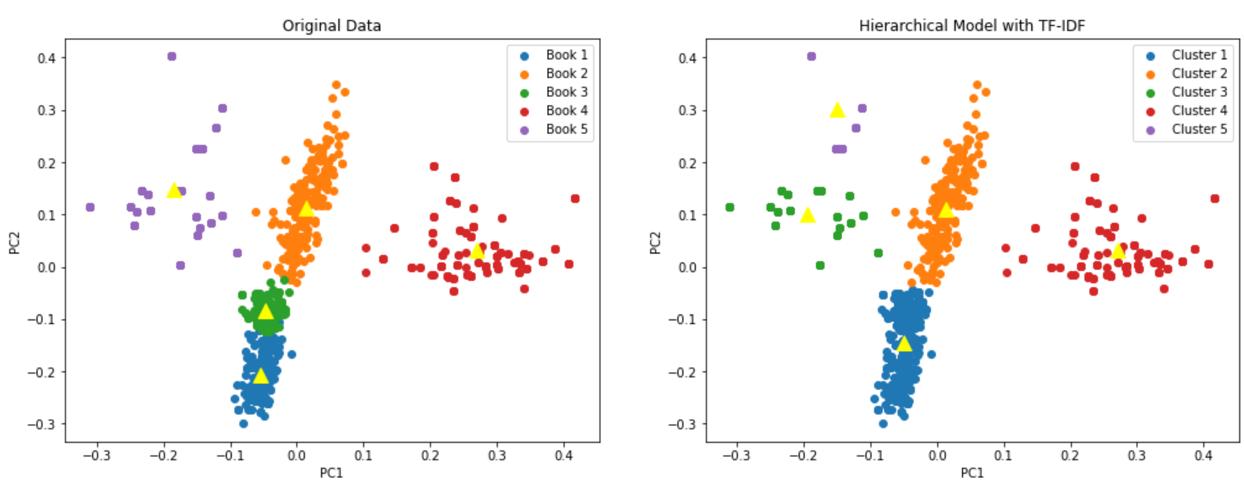
Dendrogram plot



Evaluations for 5 number of clusters

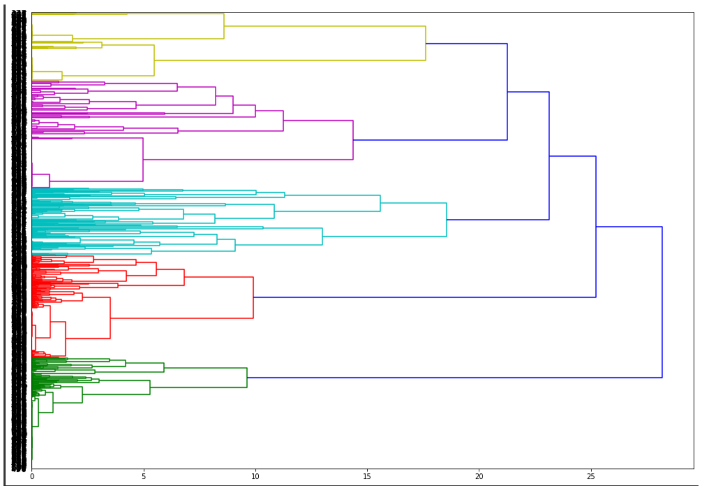
```
Kappa for the Hierarchy model at number of Cluster 5 is  0.73375
For n_clusters = 5  Silhouette Coefficient is: 0.0403
For n_clusters = 5  The average homogeneity_score : 1
For n_clusters = 5 The v_measure_score : 0.8693
```

Plot original data vs hierarchial model with **TF-IDF.**



Hierarchial model with LDA:

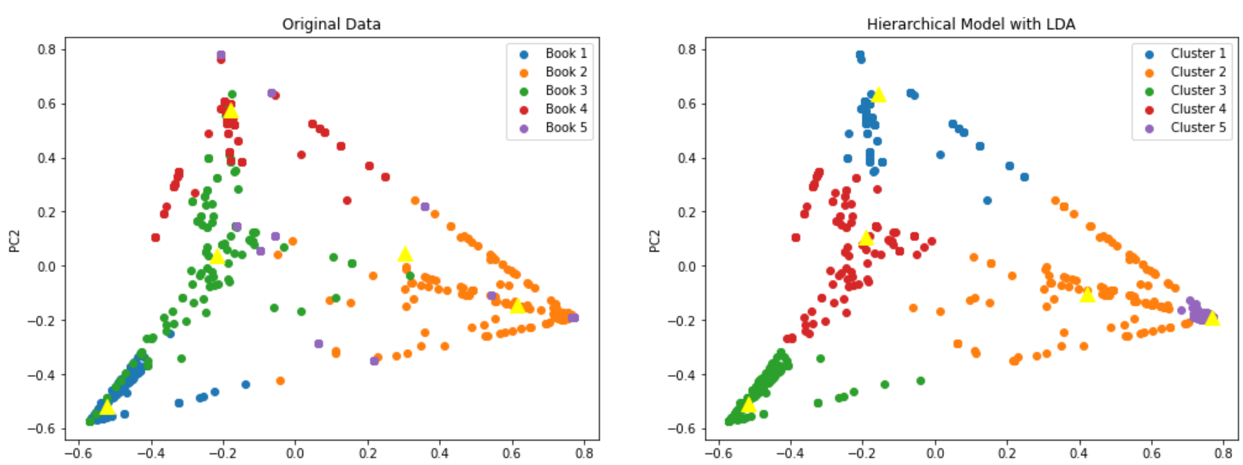
Dendrogram plot



Evaluations for 5 number of clusters

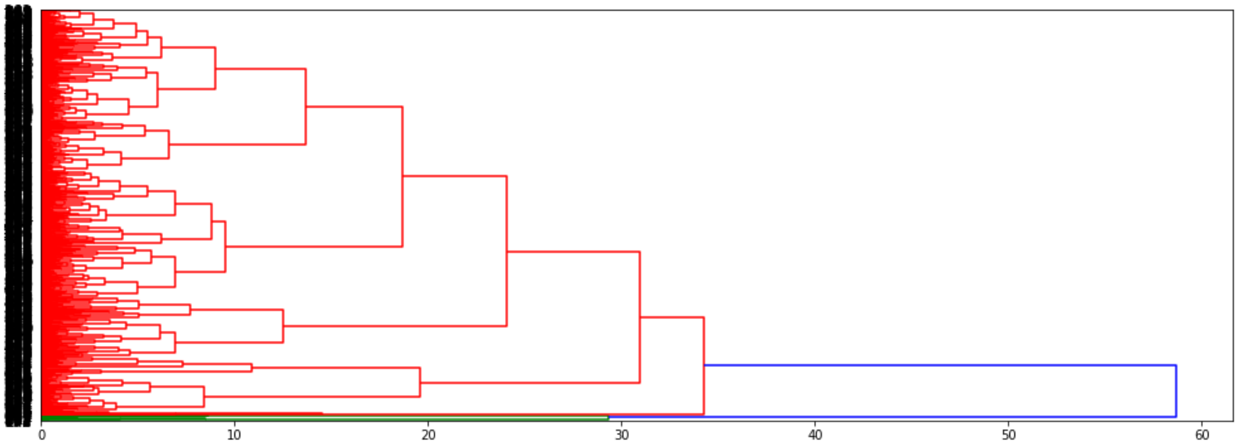
```
Kappa for the Hierarchy model at number of Cluster 5 is  0.665
For n_clusters = 5  Silhouette Coefficient is: 0.6126
For n_clusters = 5  The average homogeneity_score : 1
For n_clusters = 5 The v_measure_score : 0.6309
```

Plot original data vs hierarchial model with **LDA.**



Hierarchial model with Word2Vec:

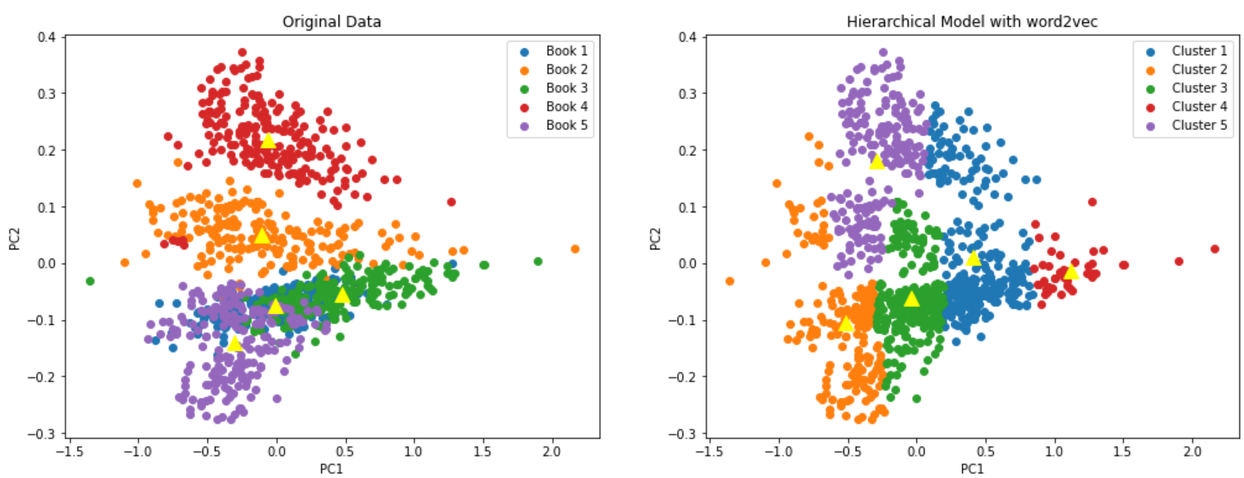
Dendrogram plot



Evaluations for 5 number of clusters

```
Kappa for the Hierarchy model at number of Cluster 5 is  0.24375000000000002
For n_clusters = 5  Silhouette Coefficient is: 0.2733
For n_clusters = 5  The average homogeneity_score : 0
For n_clusters = 5 The v_measure_score : 0.1476
```

Plot original data vs hierarchial model with **Word2Vec.**

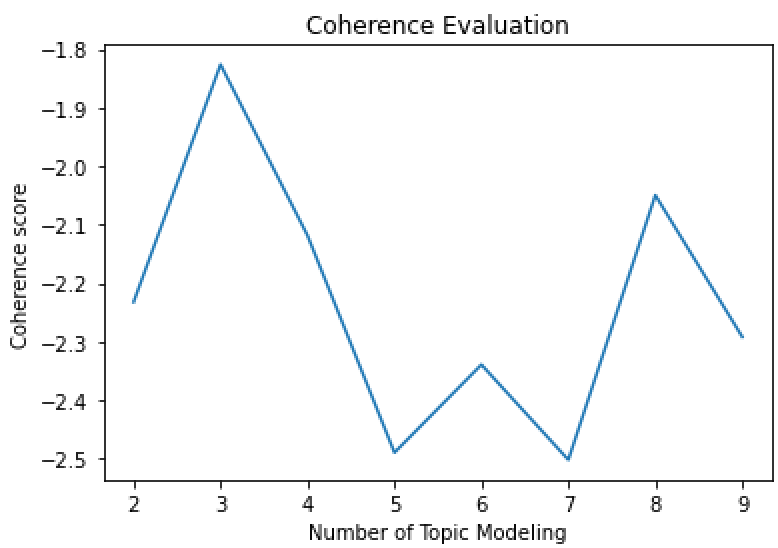


Best transformarion with Hierachy is Tf-Idf

- Kappa for the Hierarchy model at number of Cluster 5 is 0.73375
- For n_clusters = 5 Silhouette Coefficient is: 0.0403
- For n_clusters = 5 The homogeneity_score : 1
- For n_clusters = 5 The v_measure_score : 0.8693

Calculate coherence for LDA from gensim:

Selecting the best number of topic modeling for coherence.



Best coherence value is: -1.82 at 3 number of topic modeling.

For consistency, we used homogenity and v_score to measure the similarity in cluster for every model.

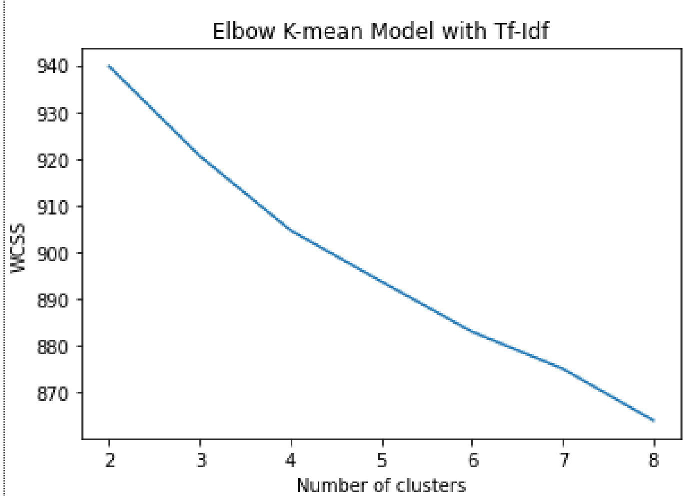
Champion model:

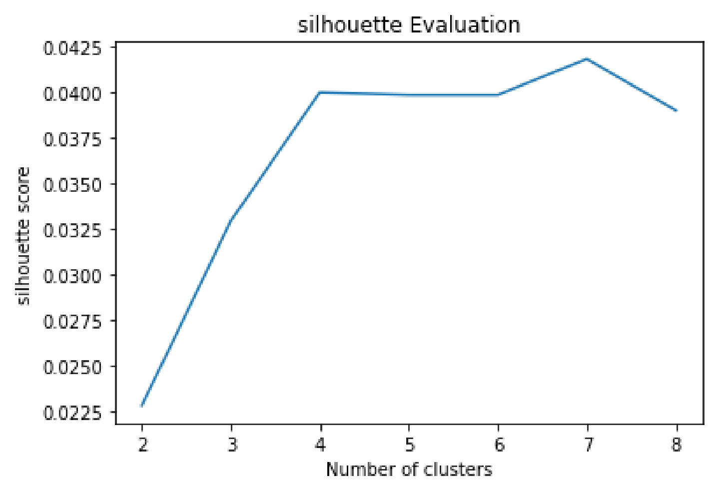
We compared our models by its Kappa scores, coherence, and silhouette.

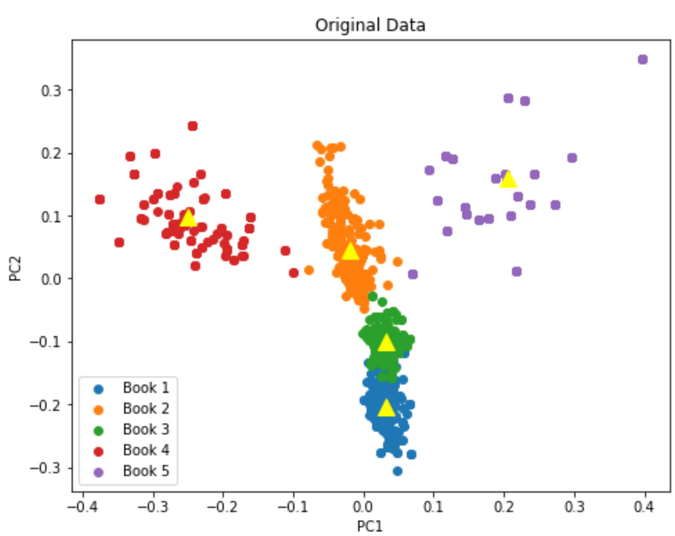
Clustering result that is the closest to the human labels.

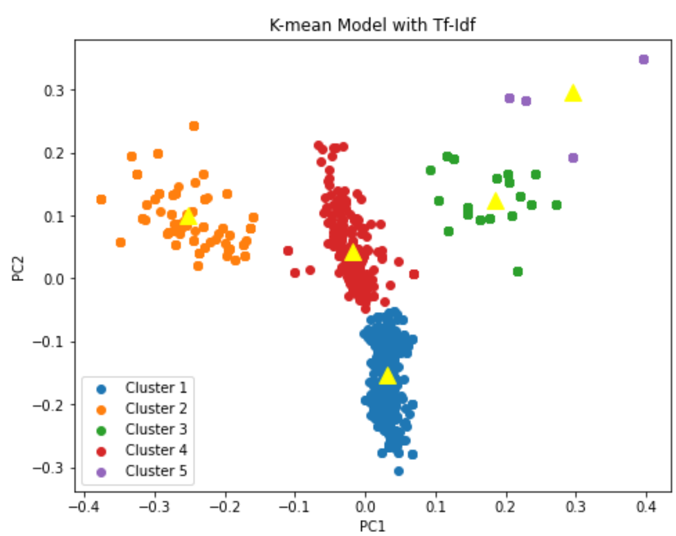
We choose K-means Algorithm with TF-IDF:

- ✓ Kappa for the model at n_clusters= 5 is 0.7488
- ✓ Best Value for n cluster is = 7 The average silhouette_score : 0.0418
- ✓ For n_clusters = 5 The silhouette_score : 0.04
- ✓ For n_clusters = 5 The homogeneity_score : 0.8234
- ✓ For n_clusters = 5 The v_measure_score : 0.9012









This is obvious that this model is the most accurate one, as it separates each paragraph related to each book categories in separate cluster.

.

Cluster	Num of paragraph
0	196
1	194
2	210
3	200
4	200

Label	a	b	c	d	e
Kmeans_TF_IDF					
0	196	0	0	0	0
1	0	194	0	0	0
2	4	6	200	0	0
3	0	0	0	200	0
4	0	0	0	0	200

Error Analysis

Idea:

- In the error analysis process, we looked at each cluster in our chosen (champion) model, and visualize how they cluster the books, and how much data they managed to cluster, and tried to find the most frequent words (10 words) in each cluster.
- We compared the most frequent words in each cluster and find the most similar ones.
- We found that the clusters separate the docs well, and in the most occurred words, there is no big conflict between the clusters.
- So, we searched for the paragraphs that the model failed to cluster right and print the most occurred words.
- We want to reduce the similarity between each cluster, the more the cluster are far from each other, the more the model is good
- The words that appear in many clusters are the ones that confuse our model, and that led to increase the error in it.

Model:

- We used this error analysis technique with our champion model: Kmeans with TFIDF.
- We made a data frame contains each word with its TF IDF values and what the number of the cluster it was belong to.
- We get the most occurred words in each cluster (0:4).

	accepted	_adair_	_alone_	_and_	_any_	_at_	_be_	_broke_	_cause_	_compassion_	...	zilpah	zion	ziphion	zippor	zoan	zohar	zorah	zuar	zurishaddai	Cluster
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	0.0	0.0	0
1	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	0.0	1
2	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	0.0	2
3	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	0.0	3
4	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	0.0	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	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	0.0	0
996	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	0.0	1
997	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	0.0	2
998	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	0.0	3
999	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	0.0	4

1000 rows × 12319 columns

Each cluster and the most occurred words

Cluster 0

	word_freq	Words
0	18.233063	emma
1	12.436194	miss
2	12.079241	would
3	11.659534	harriet
4	10.593390	weston
...	...	...
12313	0.000000	hatred
12314	0.000000	hats
12315	0.000000	hatted
12316	0.000000	hatter
12317	0.000000	zurishaddai

12318 rows × 2 columns

Cluster 1

	word_freq	Words
0	26.506424	unto
1	24.107254	lord
2	21.029456	shall
3	13.995123	thou
4	10.479593	israel
...	...	...
12313	0.000000	greatcoat
12314	0.000000	greatness
12315	0.000000	greatnesse
12316	0.000000	greefe
12317	0.000000	zurishaddai

12318 rows × 2 columns

Cluster 2

	word_freq	Words
0	12.302770	brown
1	9.891546	like
2	9.731789	said
3	7.272611	flambeau
4	6.900820	father
...	...	...
12313	0.000000	philosophy
12314	0.000000	entertaine
12315	0.000000	philistines
12316	0.000000	philistine
12317	0.000000	_accepted_

12318 rows × 2 columns

Cluster 3

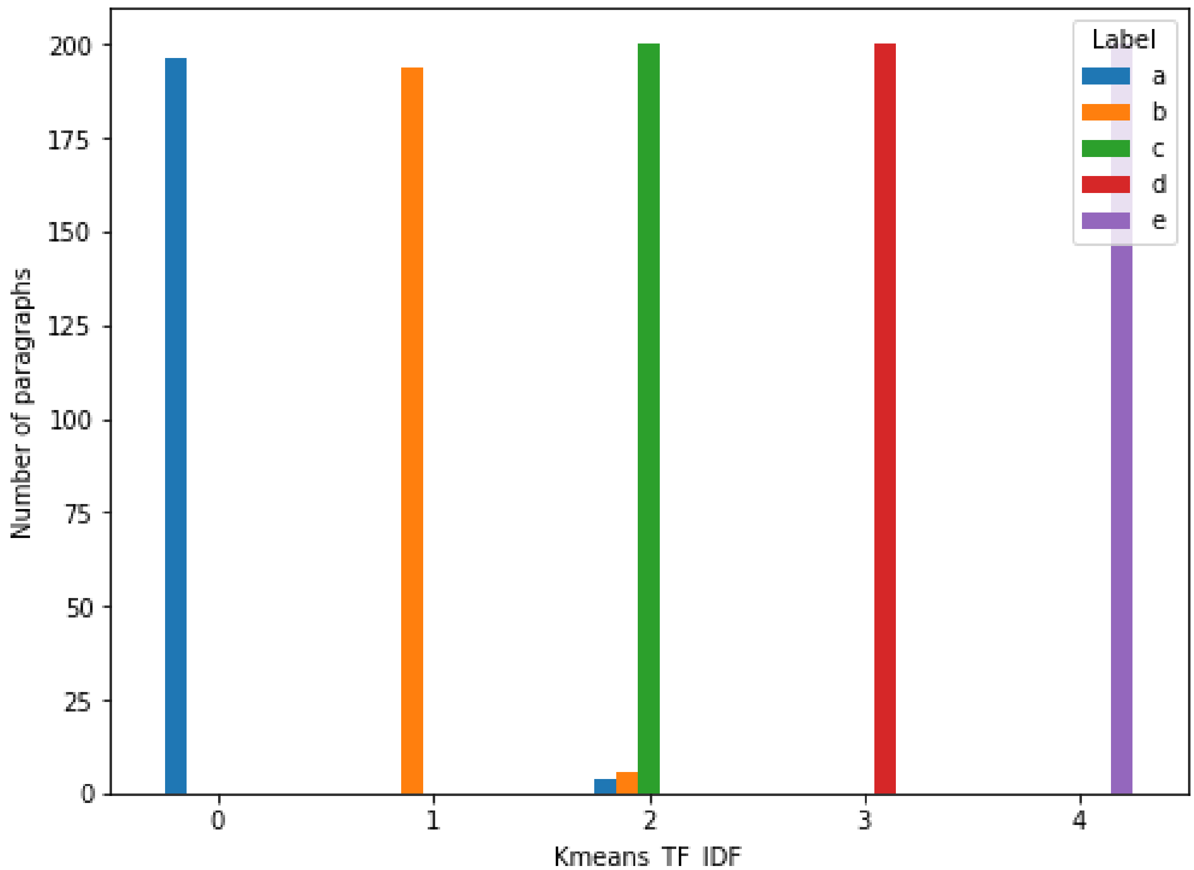
	word_freq	Words
0	26.589467	caesar
1	20.771001	brutus
2	18.686161	haue
3	16.761804	cassi
4	12.870834	cassius
...	...	...
12313	0.000000	fulfilled
12314	0.000000	fulham
12315	0.000000	fullest
12316	0.000000	fullness
12317	0.000000	zurishaddai

12318 rows × 2 columns

Cluster 4

	word_freq	Words
0	12.153468	weep
1	11.230576	little
2	10.948309	thou
3	10.792630	love
4	10.493045	thee
...	...	...
12313	0.000000	flat
12314	0.000000	flattered
12315	0.000000	flatterer
12316	0.000000	flatterers
12317	0.000000	zurishaddai

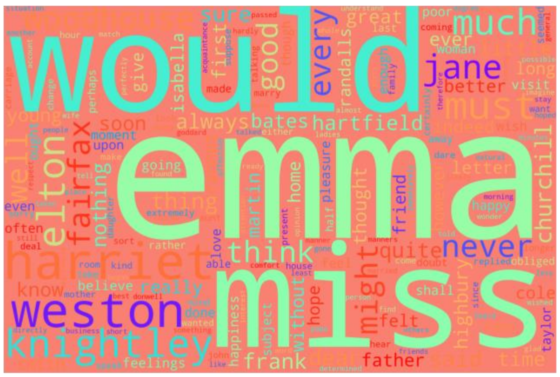
12318 rows × 2 columns

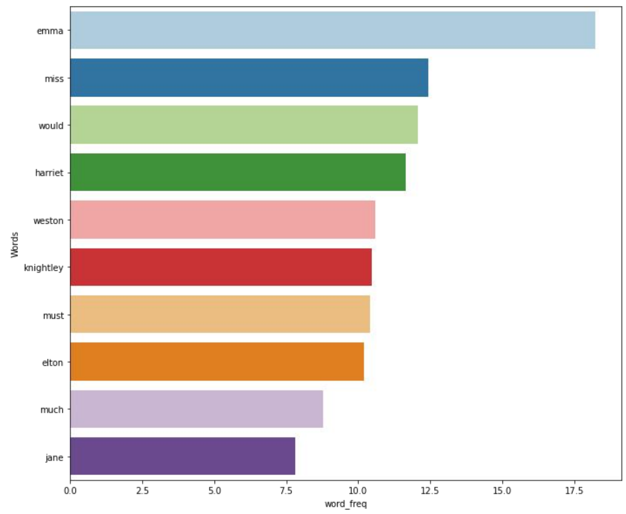


We can see that cluster two was confused with clustering, it clusters data that should be in cluster 1 and 0.

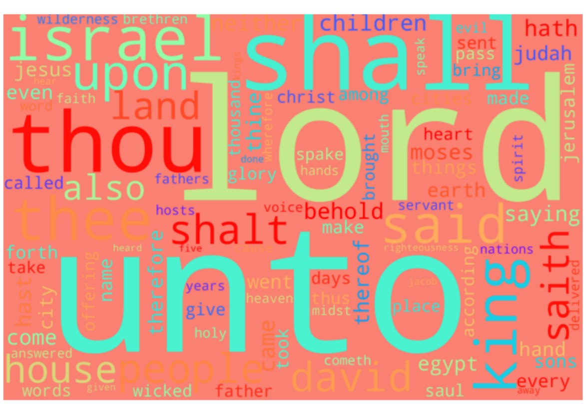
Most 10 Frequent words in each Cluster:

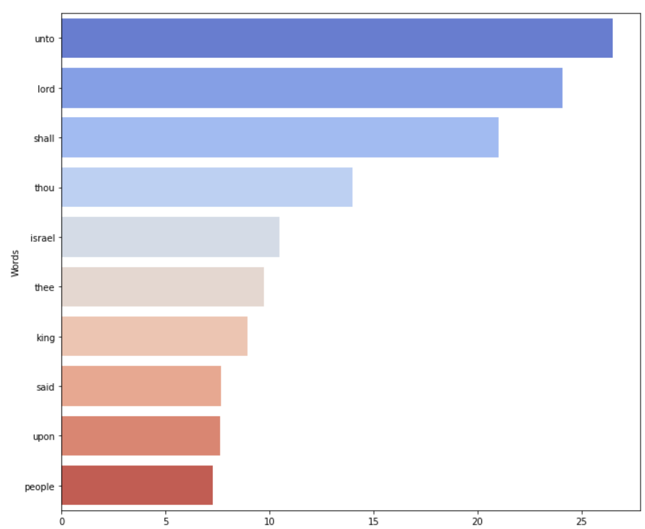
Cluster 0



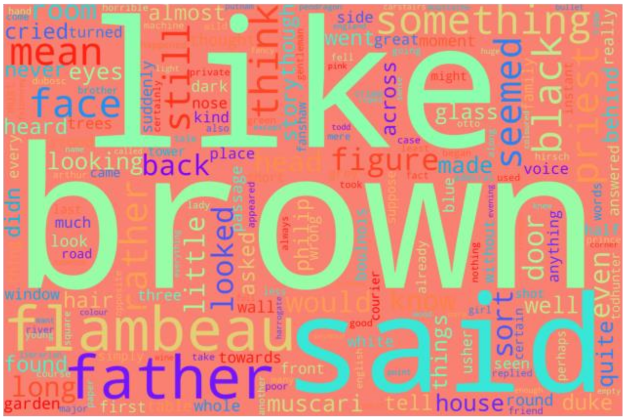


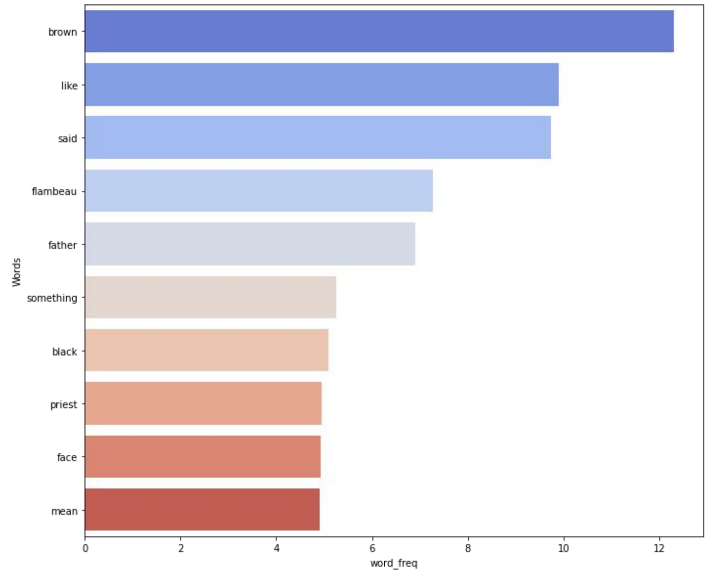
Cluster 1



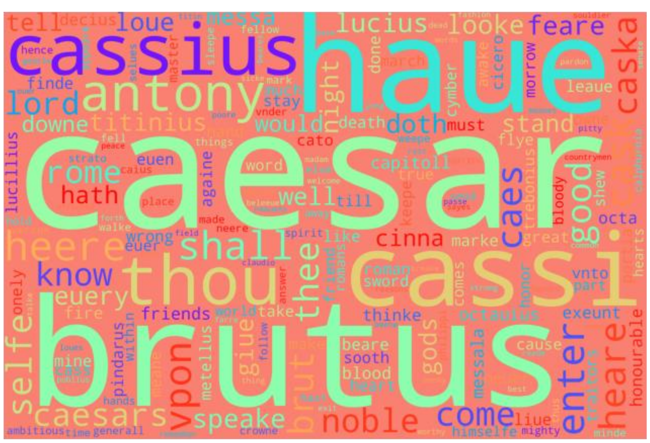


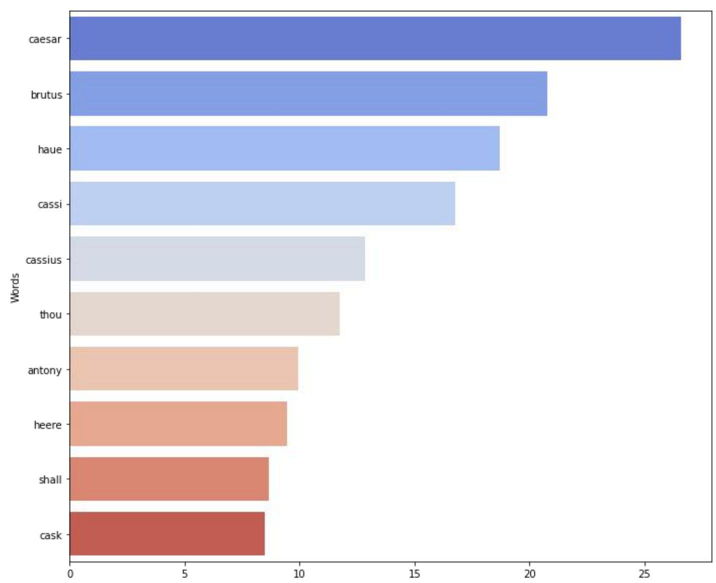
Cluster 2



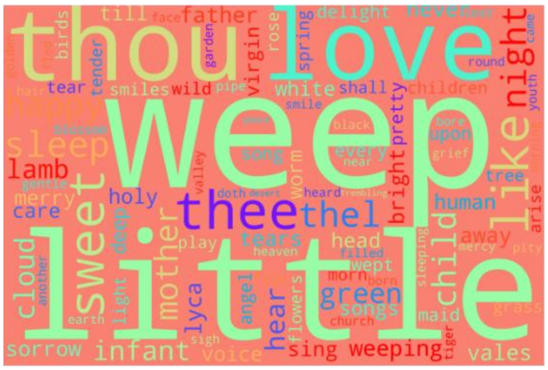


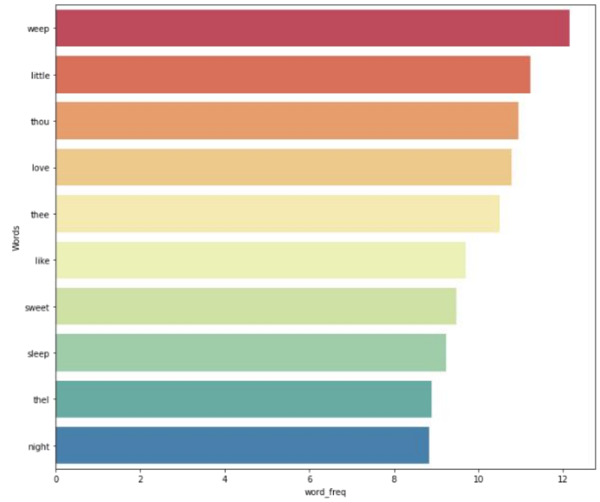
Cluster 3





Cluster 4





We can see that There is common words in cluster 2, 1, 0 that made the error drop down.