A.2: Assess Clustering and Classification Outputs

Gurjus Singh

MSDS 453: Natural Language Processing Sec 56

Professor Jennifer Sleeman

Northwestern University

Introduction & Problem Statement

In this research multiclass classification was the main focus with the movie reviews data set that a group of students from an NLP course collected. The students collected 7 reviews each for a total of 78 movie reviews which was labeled manually, and then K-Means was used to also label the dataset using a TF-IDF matrix created. The created TF-IDF matrix was used as features and labels to assess several models for classification. Metrics such as accuracy, confusion matrix were used to assess the models. The Models used were K-Nearest Neighbors, Support Vector Machines, Random Forests, Multinomial Naïve Bayes and Dense Neural Networks.

Literature Review

A paper related to the research came from Adam Nyberg (2018) who also did genre classification. He used the IMDB Movie Reviews Dataset for his research (pg. 1) . The author used key techniques such as removing stop words, lemmatization, and generating a TF-IDF matrix (Nyberg, 2018, pg. 1). The models that the author used to evaluate classification was Multi-Layer Perceptron and K-Nearest Neighbor (Nyberg, 2018, pg. 3). He describes the Multi-Layer Perceptron as having three layers with non-linear activation functions which are input, hidden and output layers (Nyberg, 2018, pg. 3). Nyberg describes KNN model as having a K parameter which signifies nearest neighbors where similarity is computed by Minkowski Distance or Euclidean Distance (Nyberg, 2018, pg. 3). The author used precision, recall and accuracy as metrics to his classification problem (Nyberg, 2018, pg. 4). The accuracy was quite low for his two models getting 55.4 percent on KNN and 37 percent with Multi-Layer Perceptron (Nyberg, 2018, pg. 7).

Data

The dataset used came from a CSV file that a set of students obtained by researching movie reviews online. The CSV contains columns such as **Doc_ID** used to distinguish each classmate document as my classmates obtained a total of seven movie reviews. It also contains **the title** for the text of each review, and the text itself. The data consists of 78 movie reviews. Below in 1-1 is a view of the CSV file imported as a DataFrame in Python.

Data Prep for the text itself consisted of tokenizing of the text into words. It also consisted of stop word removal, and it also consisted of removing punctuation. A custom function Clean Doc seen in 1-2 was used to prepare the text for TF-IDF vectorization. The generated TF-IDF matrix was also used for classification and to help label the dataset with K-Means as seen in 1-7 in the Appendix. In 1-7, intuitively, the expected outcome would be that each student's seven documents would get clustered together, but in analyzing 1-7, this is not the output and is surprising. One explanation to this could be that there are overlapping similarities between all of the student's documents. Another explanation to is analyzing the part of the movie review the student decided to cut out as some of the main words explaining the movie review could have been lost in the process. There are also document instances in the clusters where several of one student's documents are clustered together, but in its entirely the K-Means clustering algorithm does not perform as well as what a human would expect to see.

The dataset was also manually labeled by genres. In all there were 13 genres as seen in 1-3. The data labeling was based on using human intuition, and each human's labeling could be different from person to person. The final preprocessed text is in Data Frame in 1-4.

Out[51]:

	Doc_ID	DSI_Title	Text
0	0	BJL_Doc1_So-when-I.txt	So when I say that Tron: Legacy had me on the ...
1	1	BJL_Doc2_An-exploration-of.txt	TL;DR – An exploration of a film that effortle...
2	2	BJL_Doc3_Having-received-a.txt	Having received a mysterious signal emanating ...
3	3	BJL_Doc4_The-good-news.txt	The good news for that cult is that Tron: Lega...
4	4	BJL_Doc5_The_addition_of.txt	The addition of that stately "legacy" to the t...
...	...	...	...
73	73	SCFIDRJ_Doc3_The_Matrix.txt	"The Matrix," with Keanu Reeves, Laurence Fish...
74	74	SCFIDRJ_Doc4_I_Robot.txt	'I, Robot" takes place in Chicago circa 2035, ...
75	75	SCFIDRJ_Doc5_Ex_Machina.txt	After a recent run ofoverly-hyped, butultimate...
76	76	SCFIDRJ_Doc6_Her.txt	Spike Jonze's "Her" plays like a kind of mirac...
77	77	SCFIDRJ_Doc7_The Terminator.txt	On Oct. 26, 1984, a newsci-fifranchise was lau...

Table with titles of documents and Text before cleaning 1-1

Before using the clean doc function, to clean the text up, the text has characters, syntax and quotes which do contribute to classification of the text. Therefore, it is better to remove these unwanted elements of the text. For example, in 1-5 in the Appendix, there are numbers that get removed from the clean doc function, syntax that such as quotations, commas. Stop words such ‘and’, ‘the’, ‘to’. These are the unwanted text elements.

After removing the unwanted text, in 1-6 in the Appendix, the text appears to be in a list of significant terminology from the movie reviews. A few terms that are significant are ‘mysterious’, ‘synapsefrying’ and ‘esoteric’ could give the genre of this film.

```
def clean_doc(doc):
    #split document into individual words
    tokens=doc.split()
    re_punc = re.compile('[%s]' % re.escape(string.punctuation))
    # remove punctuation from each word
    tokens = [re_punc.sub('', w) for w in tokens]
    # remove remaining tokens that are not alphabetic
    tokens = [word for word in tokens if word.isalpha()]
    # filter out short tokens
    tokens = [word for word in tokens if len(word) > 4]
    #lowercase all words
    tokens = [word.lower() for word in tokens]
    # filter out stop words
    stop_words = set(stopwords.words('english'))
    tokens = [w for w in tokens if not w in stop_words]
    # word stemming
    # ps=PorterStemmer()
    # tokens=[ps.stem(word) for word in tokens]
    return tokens
```

Clean Doc Function 1-2

```
{'Comedy': 0,
 'Disney': 1,
 'Hero': 2,
 'Horror': 3,
 'India': 4,
 'Justice': 5,
 'Love': 6,
 'Magic': 7,
 'Science-Fiction': 8,
 'Sports': 9,
 'Thriller': 10,
 'Time': 11,
 'War': 12}
```

Manual Labeling 1-3

	Doc_ID	DSI_Title	Text	processed_text	Labels
0	0	BJL_Doc3_Having-received-a.txt	Having received a mysterious signal emanating ...	[received, mysterious, signal, emanating, arca...	1
1	1	BJL_Doc2_An-exploration-of.txt	TL;DR – An exploration of a film that effortle...	[exploration, effortless, blends, music, visua...	1
2	2	BJL_Doc6_TRON-Legacy-is.txt	'TRON: Legacy' is the sequel to the 1982 film ...	[sequel, stars, bridges, garrett, hedlund, oli...	1
3	3	BJL_Doc1_So-when-I.txt	So when I say that Tron: Legacy had me on the ...	[legacy, impressed, special, effects, action, ...	1
4	4	BJL_Doc5_The_addition_of.txt	The addition of that stately "legacy" to the t...	[addition, stately, legacy, title, strains, co...	1
...	...	...	...	...	...
72	72	SCFIDRJ_Doc3_The_Matrix.txt	"The Matrix," with Keanu Reeves, Laurence Fish...	[matrix, keanu, reeves, laurence, fishburne, c...	8
73	73	SCFIDRJ_Doc4_I_Robot.txt	'I, Robot" takes place in Chicago circa 2035, ...	[robot, takes, place, chicago, circa, spectacu...	8
74	74	SCFIDRJ_Doc5_Ex_Machina.txt	After a recent run ofoverly-hyped, butultimate...	[recent, ofoverlyhyped, butultimatelydisappoin...	8
75	75	SCFIDRJ_Doc6_Her.txt	Spike Jonze's "Her" plays like a kind of mirac...	[spike, jonzes, plays, miracle, first, around,...	8
76	76	SCFIDRJ_Doc7_The Terminator.txt	On Oct. 26, 1984, a newsci-fifranchise was lau...	[newscififranchise, launched, theatrical, debu...	8

77 rows × 5 columns

Table with titles of documents and Text before cleaning and after cleaning 1-4

Research Design and Modeling Method

Key things in the research were there were a total 40 experiments. Parameters were changed while experimenting such as K in K-Means, Kernel in the Support Vector Machine, K in K-Nearest Neighbors, changing the percent split in **train_test_split** method and **n_estimators** in Random Forests. N-Gram parameter in the TF-IDF function was also experimented on. The TF-IDF matrix was used as features. TF-IDF is important as it as it helps to understand if a word is important to a document and corpus (Jurafsky & Martin, 2008, pg. 8) . If the word appears less

in the corpus it treats it as more important and its score improves (Jurafsky & Martin, 2008, pg. 890)

As mentioned above there have been models used such as the Support Vector Machine, Random Forests, Multinomial Naïve Bayes, K-Nearest Neighbors, and Dense Neural Networks. Support Vector involves a hyperplane which is involved in classifying the data. (Aylien, n.d. , section: What is hyperplane). The distance between nearest data and hyperplane is know as margin (Aylien, n.d. ,section: What is hyperplane). The goal of the SVM is to find a margin/distance that is the largest to help classify datapoints correctly (Aylien, n.d. ,section: What is hyperplane).

Random Forests are another algorithm that involves a single component decision tree (Yiu, 2021,Section: Decision Trees). These trees each separate data based on features (Yiu, 2021,Section: Decision Trees). In Random Forests, a collection of trees is involved each making a prediction that is not correlated by other trees (Yiu, 2021,Section:Random Forests). Once each tree comes to a prediction, the majority prediction wins (Yiu, 2021,Section:Random Forests).

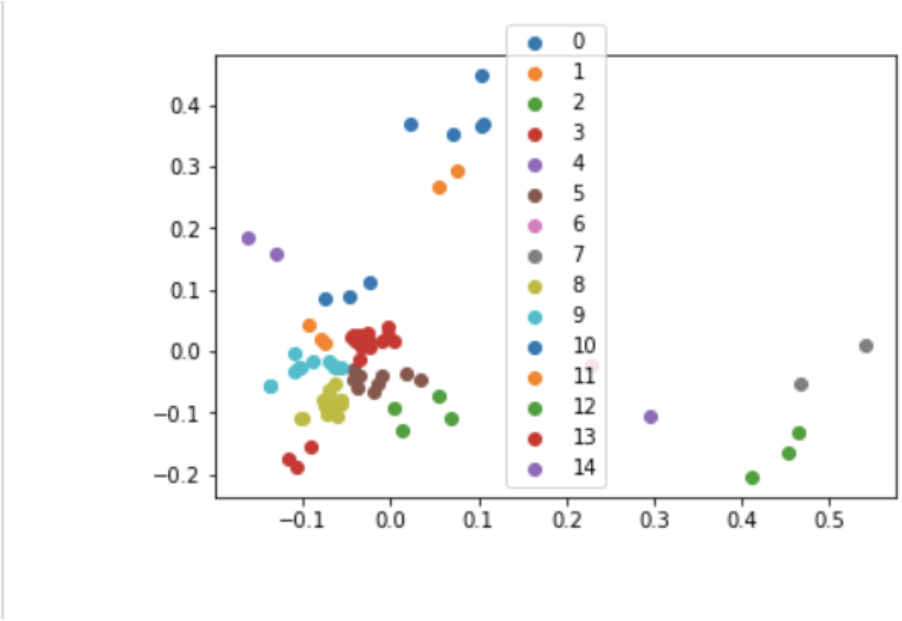
Multinomial Naïve Bayes is another model used. It is explained by using the feature vector and figuring out the probability of each event occurring based on multinomial distribution ("Geeksforgeeks", 2020, n.page). Multinomial Naïve Bayes is for document classification, but originates from Naïve Bayes original which is based on the Bayes' Theorem formula ("Geeksforgeeks", 2020, n.page). Bayes' Theorem finds probability of an event occurring "given the probability of another event already occurred" ("Geeksforgeeks", 2020, n.page).

K-Nearest Neighbors is an algorithm which relies on K, which signifies the number of neighbors that are related to a given point (Srivastava, 2020, n.page). K can also be used to find

the boundaries (Srivastava, 2020, n.page). In regards to where a datapoint belongs K-Nearest Neighbor uses Euclidean Distance for this (Schott, 2019, n.page).

The Dense Neural Network is a little bit more sophisticated from the other models mentioned. It is composed of layers like in an actual brain, and nodes that at like neurons (Kyrykovych, 2020, n.page).

Results



PLOT K MEANS ALGORITHM CLUSTERS; K = 15 1-8

	A	B	C	D	E	F
1	MODEL	PARAMETERS TRIED	TRAIN ACCURACY	TEST ACCURACY	LABELED	TRAIN/TEST SPLIT
2	SVM	KERNAL = POLY	100%	25%	MANUALLY	90-10 SPLIT
3	Random Forest	n_estimators = 1000	100%	62.50%	MANUALLY	90-10 SPLIT
4	Multinomial Naïve Bayes	None	100%	62.50%	MANUALLY	90-10 SPLIT
5	Random Forest	n_estimators = 1000	100%	75%	K-MEANS K =15	90-10 SPLIT
6	SVM	KERNAL = LINEAR	100%	25%	K-MEANS K =15	90-10 SPLIT
7	Multinomial Naïve Bayes	None	100%	75%	K-MEANS K =15	90-10 SPLIT
8	Random Forest	n_estimators = 10000	100%	62.50%	K-MEANS K =15	90-10 SPLIT
9	Random Forest	n_estimators = 10000	100%	0.00%	K-MEANS K =15	95-5 SPLIT
10	Random Forest	n_estimators = 10000	100%	31.25%	K-MEANS K =20	80-20 SPLIT
11	SVM	KERNAL = LINEAR	100%	12.50%	K-MEANS K =20	80-20 SPLIT
12	SVM	KERNAL = POLY	100%	0.00%	K-MEANS K =20	80-20 SPLIT
13	SVM	KERNAL = RBF	100%	0.00%	K-MEANS K =20	80-20 SPLIT
14	SVM	KERNAL = RBF	100%	0.00%	K-MEANS K =20	80-20 SPLIT
15	Multinomial Naïve Bayes	None	100%	31%	K-MEANS K =20	80-20 SPLIT
16	Random Forest	n_estimators = 10000	100%	62.50%	K-MEANS K =10	80-20 SPLIT
17	Multinomial Naïve Bayes	None	100%	62.50%	K-MEANS K =10	80-20 SPLIT
18	SVM	KERNAL = RBF	100%	25.00%	K-MEANS K =20	80-20 SPLIT
19	TRIED CHANGING NGRAMS BUT ACCURACY REMAINED SAME					
20	Random Forest	n_estimators = 100	100%	62.50%	K-MEANS K =15	90-10 SPLIT
21	KNN	n_neighbors =	73.90%	50%	MANUALLY	90-10 SPLIT
22	KNN	n_neighbors =	100.00%	38%	MANUALLY	90-10 SPLIT
23	KNN	n_neighbors =	100.00%	75%	K-MEANS K =15	90-10 SPLIT
24	KNN	n_neighbors =	100.00%	63%	K-MEANS K =15	90-10 SPLIT
25	DNN 128 NEURONS - 2 LAYERS	NONE	13.16%	17.95%	K-MEANS K =15	90-10 SPLIT
26	SVM	KERNAL = POLY	100%	29.03%	MANUALLY	60-40 SPLIT
27	KNN	n_neighbors =	100.00%	45%	MANUALLY	60-40 SPLIT
28	Random Forest	n_estimators = 1000	100%	45%	MANUALLY	60-40 SPLIT
29	Multinomial Naïve Bayes	None	100%	45%	MANUALLY	60-40 SPLIT
30	SVM	KERNAL = POLY	100%	19.35%	K-MEANS K =15	60-40 SPLIT
31	KNN	n_neighbors =	100.00%	48%	K-MEANS K =15	60-40 SPLIT
32	Random Forest	n_estimators = 1000	100%	32%	K-MEANS K =15	60-40 SPLIT
33	Multinomial Naïve Bayes	None	100%	32%	K-MEANS K =15	60-40 SPLIT

40 EXPERIMENTS of Machine Learning Models; Please see attached Excel File for better view

1-9

```
array([[1, 0, 0, 0, 0, 0, 0],
       [1, 0, 0, 0, 0, 0, 0],
       [0, 0, 1, 0, 0, 0, 0],
       [0, 0, 0, 2, 0, 0, 0],
       [0, 0, 0, 0, 1, 0, 0],
       [0, 0, 0, 0, 1, 0, 0],
       [0, 0, 0, 0, 0, 0, 1]])
```

Confusion Matrix for high test accuracy 75% which includes Random Forest, Multinomial Naïve Bayes, and KNN 1-10

Analysis and Interpretation

In 1-8, is a plot of 15 clusters for K-Means. Several of the data points in the K-Means clusters seem to be large in distance to other points in its cluster. This could be part which is not intuitive when looking at the clustered documents in 1-7 of the Appendix. As mentioned in the Research Design and Modeling Method section there were 40 experiments done. In 1-9, the 40 experiments can be seen. Some observations were the models were overfitting as the train set accuracy was significantly higher than the test set accuracy. KNN, Multinomial Naïve Bayes and Random Forests performed the best throughout the experiments done with a test accuracy around 75 percent as seen in 1-9 and in the Excel Sheet. These models performed the best when adjusting the split to 90-10, using K-Means to label the dataset, and setting K=15. The Confusion Matrix for these models is shown 1-10, but it can be inferred there was not a lot of data in the Test Set.

These models did not perform as well when the manual labels were used as seen in 1-3. This could be explained because the manual labels could be unevenly split. With K-Means there is a balanced split of the labels as the algorithm is finding more similarities than what a human could intuitively find, so the models are finding better patterns as they have more labels to see. Parameters such as increasing, decreasing K in K-Means resulted in much lower test accuracy.

The K in KNN was set to 2 and performed the best at that value. There were also experimentations done with NGRAMS = (1,2), (1,10), but it resulted in the same accuracies as if NGRAMS was set to (1,1). Surprisingly, the Neural Network Model did not perform as well as expected with an accuracy around 17 percent. Overall, since there were only 78 data instances, there is an accuracy problem and more data needs to be used to train on, for testing accuracy to increase.

Conclusions

This research involved exploring the movie corpus student dataset with classification models. As explained above several parameters were taken into account including train test split size, K in K-Means, K in KNN, n_estimators in Random Forests, and kernel in Support Vector Machines and NGRAMS in TF-IDF. With these parameters taken into account, the models seemed to overfit and the best model performed with 100 percent and 75 percent as seen in 1-9 and in the Excel Sheet. Therefore, it would be best to choose either KNN, Random Forest, or Multinomial Naïve Bayes each set to the parameters seen in 1-9 and in the Excel Sheet which are K in K-Means = 15, with a 90-10 split, K = 2 in KNN and n_estimators set to 1000 in Random Forests.

Directions for future work

More experimentation will be needed in how to get a better accuracy with less data as this movie corpus only involves 78 rows of text. More experimentations will have to be done with how to balance the train, test sets, so every different instance is trained on, which will only improve the test sets accuracy. There should also be more experimentations done on Neural Networks as the accuracy when using DNN was surprising. LSTM/RNN models are believed to perform better with Text, so more exploration will have to be done with these models.

References

Aylien, N. B. (n.d.). *Support Vector Machines: A Simple Explanation*. KDnuggets. Retrieved May 1, 2021, from https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

GeeksforGeeks. (2020a, May 15). *Naive Bayes Classifiers*. https://www.geeksforgeeks.org/naive-bayes-classifiers/

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing, 2nd Edition* (2nd ed.). Prentice Hall.

Kyrykovych, A. (2020, February). *Deep Neural Networks*. KDnuggets. https://www.kdnuggets.com/2020/02/deep-neural-networks.html

Nyberg, A. (2018, February 4). *Classifying movie genres by analyzing text reviews*. ArXiv.Org. https://arxiv.org/pdf/1802.05322.pdf

Schott, M. (2020, February 27). *K-Nearest Neighbors (KNN) Algorithm for Machine Learning*. Medium. https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26

Srivastava, T. (2020, October 18). *Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

Yiu, T. (2021, March 28). *Understanding Random Forest - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Appendix

Having received a mysterious signal emanating from his father's arcade, Sam sets off to investigate, and following a few brief key presses on an old computer terminal, he too finds himself locked inside the Grid.

Almost like an atonement for the lengthy scene-setting that forms Legacy's opening, Kosinski has Sam kinkily stripped out of his civilian wear, zipped into a skin-tight costume made of rubber and glowsticks, and throws him into a series of Tron-referencing gladiatorial games, beginning with a high-velocity tournament with deadly discs, and then a dizzying Light Cycle chicken run, before he's whisked away by slinky heroine Quorra.

Quorra, played with feline innocence by Olivia Wilde, is a warrior program who, like Sam, is dangerously multi-talented, and as at home talking about 20th century literature as she is driving an off-road vehicle or smashing opponents into cubes with fists and feet. Whisking Sam off the Grid, Quorra ushers Sam to a refuge in the world's digital wilderness, where his father Kevin sits in a zen-like trance.

It's here we learn a little more of Kevin's backstory – his most advanced program, Clu, has gone rogue, turning the Grid into a fascist state, subjugating its populace and trapping Kevin inside the world's neon confines.

It also feels, at the same time, that Kosinski is a little impatient with some of the trappings he's inherited – the disc battles and Light Cycle duels are dispensed with early on and never revisited, which is a pity because, as muddled and synapse-frying as these sequences are, they're never topped elsewhere.

There's an aerial dogfight that attempts to dazzle with sheer light and noise, but is too obviously reminiscent of its analogues in the Star Wars movies to really satisfy, and gives way to a denouement that could be seen as muted,

we haven't yet reached the point where we can create an entirely convincing human head out of pixels.

It's this aspect that will be looked upon least fondly in another 28 years, I fear. But at the same time, there's an awful lot in Kosinski's take on Tron that is enormously enjoyable.

Clu may have all the presence of a Texas Instruments calculator, but the real Jeff Bridges is an absolute delight as Kevin Flynn. Playing up to his Baby Boomer equivalent of Obi Wan Kenobi, his every utterance of "far out" and "I'm gonna go and touch the sky" are perfectly delivered – they're lines that could only work coming from his mouth, and had he wandered around the Grid with a carton of milk in his hand, this could just as easily have been Tron: Lebowski.

```
Garett Hedlund is merely okay in his everyman hero role, but Olivia
Wilde is great as Quorra, and the film could have benefited from more of
her.

As a big-screen experience, and as pure eye-candy, Tron: Legacy is a
riot of colour and esoteric design, and visually, Legacy is a confident
debut from Kosinski.
```

Example of Unclean Text 1-5

```
['received', 'mysterious', 'signal', 'emanating', 'arcade',
'investigate', 'following', 'brief', 'presses', 'computer', 'terminal',
'finds', 'locked', 'inside', 'almost', 'atonement', 'lengthy',
'scenesetting', 'forms', 'opening', 'kosinski', 'kinkily', 'stripped',
'civilian', 'zipped', 'skintight', 'costume', 'rubber', 'glowsticks',
'throws', 'series', 'tronreferencing', 'gladiatorial', 'games',
'beginning', 'highvelocity', 'tournament', 'deadly', 'discs',
'dizzying', 'light', 'cycle', 'chicken', 'whisked', 'slinky', 'heroine',
'quorra', 'quorra', 'played', 'feline', 'innocence', 'olivia', 'wilde',
'warrior', 'program', 'dangerously', 'multitalented', 'talking',
'century', 'literature', 'driving', 'offroad', 'vehicle', 'smashing',
'opponents', 'cubes', 'fists', 'whisking', 'quorra', 'ushers', 'refuge',
'digital', 'wilderness', 'father', 'kevin', 'zenlike', 'trance',
'learn', 'little', 'backstory', 'advanced', 'program', 'rogue',
'turning', 'fascist', 'state', 'subjugating', 'populace', 'trapping',
'kevin', 'inside', 'confines', 'feels', 'kosinski', 'little',
'impatient', 'trappings', 'inherited', 'battles', 'light', 'cycle',
'duels', 'dispensed', 'early', 'never', 'revisited', 'muddled',
'synapsefrying', 'sequences', 'never', 'topped', 'elsewhere', 'aerial',
'dogfight', 'attempts', 'dazzle', 'sheer', 'light', 'noise',
'obviously', 'reminiscent', 'analogues', 'movies', 'really', 'satisfy',
'gives', 'denouement', 'could', 'muted', 'reached', 'point', 'create',
'entirely', 'convincing', 'human', 'pixels', 'aspect', 'looked',
'least', 'fondly', 'another', 'years', 'awful', 'enormously',
'enjoyable', 'presence', 'texas', 'instruments', 'calculator',
'bridges', 'absolute', 'delight', 'kevin', 'flynn', 'playing', 'boomer',
'equivalent', 'kenobi', 'every', 'utterance', 'gonna', 'touch',
'perfectly', 'delivered', 'lines', 'could', 'coming', 'mouth',
'wandered', 'around', 'carton', 'could', 'easily', 'lebowski', 'garett',
'hedlund', 'merely', 'everyman', 'olivia', 'wilde', 'great', 'quorra',
'could', 'benefited', 'bigscreen', 'experience', 'eyecandy', 'legacy',
'colour', 'esoteric', 'design', 'visually', 'legacy', 'confident',
'debut', 'kosinski', 'storytelling', 'terms', 'legacy', 'seems',
'unusually', 'timid']
```

Example of Clean Text 1-6

```
CLUSTER 0

{0: ['MSS_Doc4_Troy-Is-Based.docx', 'EG_Doc2_This_Time_Dream.docx',
'MSS_Doc6_Having-Become-The.docx', 'CVN_Doc3_The_CIA_Agent.docx'],

CLUSTER 1

1: ['BJL_Doc3_Having-received-a.txt', 'BJL_Doc2_An-exploration-of.txt',
'BJL_Doc6_TRON-Legacy-is.txt', 'BJL_Doc1_So-when-I.txt',
'BJL_Doc5_The_addition_of.txt', 'BJL_Doc4_The-good-news.txt',
'BJL_Doc7_Both-Blade-Runner.txt', 'MSS_Doc3_That-Old-Time.docx',
'WS_DOC1_Tenet.docx', 'GS_DOC6_Hidden_Figures_Review.docx',
'EG_Doc1_When_Traveling.docx', 'MSS_Doc5_When-Dune-The.docx']

CLUSTER 2

2: ['CVN_Doc7_1917.docx', 'EG_Doc3_Epic_Intimacy_Arrival.docx',
'CVN_Doc5_Dunkirk_Film_Review.docx', 'CVN_Doc2_Film_Review_Panoramic.docx'],

CLUSTER 3

3: ['GS_DOC5_Delightfully_smart_exciting.docx', 'MSS_Doc2_As-Possibly-
Cinema_s.docx', 'SCFIDRJ_Doc4_I_Robot.txt', 'SCFIDRJ_Doc5_Ex_Machina.txt'],

CLUSTER 4

4: ['RC_Doc5_I_went_into.docx', 'RC_Doc3_The_witch_a.docx',
'RC_Doc4_Were_instinctively_afraid.docx', 'CVN_Doc1_To_Hell_With.docx',
'CVN_Doc4_Pearl_Harbor.docx', 'SIM_Doc5_Movie_MS_Dhoni.docx',
'RC_Doc7_A_fundamental_difference.docx', 'SCFIDRJ_Doc3_The_Matrix.txt'],

CLUSTER 5

5: ['WS_DOC3_Prestige.docx', 'MSS_Doc7_The-Latest-Entry.docx',
'EG_Doc7_The_Martian.docx', 'WS_DOC6_Martian.docx', 'SCFIDRJ_Doc1_2001 Space
Odyssey.txt'],


CLUSTER 6

6: ['SD_Doc5_A-Hollywood-Ending.docx', 'SD_Doc3_The-Mighty-Ducks.docx',
'SD_Doc2_Goon.docx', 'CVN_Doc6_Black_Hawk_Down.docx', 'SD_Doc7_Creed.docx'],

CLUSTER 7

7: ['SIM_Doc3_Movie_Kal_Ho_Naa_Ho.docx',
'SIM_Doc7_Movie_The_Pursuit_of_Happyness.docx', 'WS_DOC2_KalhoNaHo.docx',
'SIM_Doc4_Movie_Guru.docx', 'SIM_Doc6_Movie_Kutch_Kutch_Hota_Hai.docx'],

CLUSTER 8

8: ['GS_DOC3_Movie_Review_Superman.docx',
'GS_DOC1_Superhero_Sandbagged.docx',
```

```
'GS_DOC4_Aquaman_Complete_Bellyflop.docx',
'ADRJ_Doc1_Disneys_New_Mulan.txt'],

CLUSTER 9

9: ['WS_DOC7_CoachCarter.docx', 'SD_Doc4_Throwing-A-Digital.docx',
'SD_Doc1_Writing-a-Playbook .docx'],

CLUSTER 10

10: ['RC_Doc6_In_Rodney_Aschers.docx', 'RC_Doc1_Creepy_beyond_belief.docx',
'RC_Doc2_Insidious_is_an.docx', 'WS_DOC4_ImpracticalJokers.docx',
'WS_DOC5_Memento.docx'],

CLUSTER 11

11: ['SIM_Doc1_Movie_The_Notebook.docx', 'SIM_Doc2_Move_Titanic.docx',
'SCFIDRJ_Doc7_The Terminator.txt'],

CLUSTER 12

12: ['SD_Doc6_Steamrolling-Over-Lifes.docx', 'EG_Doc5_Stuck_Steerage.docx',
'ADRJ_Doc2_Cinderella_Review_Straight-Faced.txt',
'ADRJ_Doc3_Aladdin_Review_This.txt', 'ADRJ_Doc4_Review_Beauty_And.txt',
'ADRJ_Doc5_Film_Review_Tim.txt', 'ADRJ_Doc6_The_Jungle_Book.txt',
'ADRJ_Doc7_Whats_A_Nice.txt'],

CLUSTER 13

13: ['GS_DOC2_Wonder_Woman1984_Review.docx', 'MSS_Doc1_In-The-Middle.docx',
'SCFIDRJ_Doc2_Wargames.txt', 'SCFIDRJ_Doc6_Her.txt'],

CLUSTER 14

14: ['EG_Doc6_The_Inexorable_Pull.docx',
'EG_Doc4_Off_Stars_With_Grief.docx',
'GS_DOC7_Fantastic_Four_Spoilreview.docx']}
```

K-MEANS LABELS BY CLUSTERS with K = 15

1-7