Unsupervised Document Clustering based on the Genres of the Documents

Anubhab Mandal
Indian Institute of Technology, Kharagpur

April 8, 2022

1 Introduction

This is the task 2 of the **Senseloaf Technologies** task round. Here we have some documents which have content based on different topics and genres. We need to be able to form proper clusters of the documents by Unsupervised Algorithms in an attempt to group together those documents such that each group contains documents with similar content and different groups are based on different topics. With this lets get started with the methodology.

2 Relevant Theory

2.1 Unsupervised Learning

Unsupervised machine learning, analyses and clusters unlabeled datasets using machine learning techniques. Without the need for human intervention, these algorithms help uncover patterns or data groupings. Unsupervised learning models are utilized for three main tasks—clustering(KNN, K-Means Clustering), association(mainly Apriori Algorithms), and dimensionality reduction(PCA, SVD).

2.2 Clustering

Clustering is a data-mining technique that organizes unlabeled data into groups based on similarities and differences. Clustering techniques are used to organize raw, unclassified data items into groups that are represented by information structures or patterns. There are several types of clustering algorithms, including exclusive, overlapping, hierarchical, and probabilistic.

2.2.1 K-Means Clustering

The K-means algorithm starts with a set of randomly chosen centroids that serve as the starting points for each cluster, and then performs iterative (repetitive) calculations to maximise the centroids' positions. It stops forming and optimising clusters when either: the centroids have stabilised — their values have not changed as a result of successful clustering or the specified number of iterations has been reached.

2.2.2 DBSCAN

DBSCAN stands for Density-based spatial clustering of applications with noise. It is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers. Here, we don't have to specify the number of clusters to use it. All we need is a function to calculate the distance between values and some information for what amount of distance is considered "close". DBSCAN also produces more reasonable results than k-means across a variety of different distributions.

2.3 Hierarchical Clustering

Data is grouped into a tree of clusters in the hierarchical clustering approach. Every data point is treated as a separate cluster in hierarchical clustering. After that, it repeats the following steps: Identify two clusters that are the most similar, and merge the two clusters that are the most similar. We must repeat these processes until all of the clusters have been blended together. The goal of hierarchical clustering is to create a succession of nested clusters in a hierarchical order.

It is of 2 types:

1. **Agglomerative**: Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. This is a bottom-up approach.

2. **Divisive**: here we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable. This is a top-down approach.

2.4 Evaluation Metrics

We have used 2 famous evaluation metrics for evaluation of the performance of unsupervised Clustering. They are:

1. **Sihouette Score:** The Silhouette Score and Silhouette Plot are used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighbouring clusters. This measure has a range of [-1, 1] and is a great tool to visually inspect the similarities within clusters and differences across clusters. The greater the value, better is the clustering.

2. **Calinski-Harabasz Index:** The score is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. The C-H Index is a great way to evaluate the performance of a Clustering algorithm as it does not require information on the ground truth labels. The higher the Index, the better the performance.

3 Methodology

3.1 Dataset

The dataset in use is provided on Kaggle by the organization. It is named as **Article Classification - Assignment**. This dataset contains 2225 .txt files where each file has a short paragraph on the topic mentioned in the Title of the document.

A sample text from a file in the dataset is as follows:

Title: Row threatens Hendrix museum plan

Text: Proposals to open a museum dedicated to Jimi Hendrix are flailing because of a row over the home of his late father.

The run-down house in Seattle has already been moved wholesale once and local authorities are now demanding it be moved to another site. Hendrix supporters hoped to turn the home into a museum for the guitarist. "The mayor is going to go down as the mayor who destroyed Jimi Hendrix's house," said Ray Rae Marshall of the James Marshall Hendrix Foundation. The foundation moved the building, in which Al Hendrix lived between 1953 and 1956, when the land it was built on was to be developed for housing in 2002. Now the City of Seattle wants its new plot to be used for development, giving a deadline of 22 February for the home to be moved. Mr Goldman said the authority had promised the house could remain on its new site and be turned into a memorial and community centre. Seattle officials said no such deal had been offered.

"We never said, 'You can own this property,'" said John Franklin, chief of its operations department. "From our perspective, it was a temporary situation. We have not threatened to demolish the house. We've simply asked that they have to move it." Now Mr Goldman is calling for the authority to pay to

move the building to Seattle's central district, where Hendrix grew up. Janie Hendrix, the guitarist's stepsister, said the family were still hoping the guitarist would be honoured by having a road named after him. "That's something my father really wanted to see," she said. "It would be nice if we didn't have to fight for everything to get it." Hendrix was widely considered one of the most important guitarists of his time. He died of drug overdose in 1970 at the age of 27.

3.2 Creating the Dataframe and Preprocessing the Data

I extracted the Titles and the Texts from each of the text files in the form of lists and then created a Dataframe using *Pandas* library where the 2 columns are "Title" and "Text".

	Title	Text
0	Steady job growth continues in US	The US created fewer jobs than expected in Dec...
1	GSK aims to stop Aids profiteers	One of the world's largest manufacturers of HI...
2	Games firms 'face tough future'	UK video game firms face a testing time as the...
3	Tory candidate quits over remark	A Conservative election challenger is quitting...
4	Global digital divide 'narrowing'	The "digital divide" between rich and poor nat...

Next, I dropped the duplicate datapoints and ended up with a Dataframe of length 2105. After that I used the *texthero* and *nltk* libraries to perform text cleaning and preprocessing. I removed the **stopwords, whitespaces, NaN value columns** and **turned everything to lower case**. Finally I performed **Stemming** to finish up with the Text Preprocessing part.

3.3 Applying Clustering

I decided to perform the experiment in two parts using 3 clustering algorithms for each experiment. So, the plan was to perform:

1. K-Means Clustering Algorithm
2. DBSCAN Algorithm
3. Hierarchical Agglomerative Clustering Algorithm.

first considering the Bag of Words vector from the Text only and then using the Bag of Words vector from both the Text and the Title concatenated together.

3.3.1 K-Means Clustering

I created the bag of words vector using **countVectorizer** library from *sklearn* library.

vec1 contains the text only bag of words. **vec2** contains the text+title bag of words.

Next we apply K-Means Clustering using *sklearn* library. Here, I test on various number of clusters from 3 to 11 and plotted the graph of mean squared loss vs Number of Clusters, to find the optimal number of clusters at which the loss is minimum using the Elbow Method. I found that the optimal number of clusters is 9 for both **vec1** and **vec2**.

After applying K-Means clustering on the 2 bag of words, I used the **Silhouette Score** and the **Calinski-Harabasz Index** as the evaluation metrics. I got a Silhouette Score of 0.01149 and a CH Index of 9.57 on vec1 while I got a Silhouette Score of 0.0118 and a CH Index of 8.64 when I used K-Means Clustering on vec2. As we can see the results are not very good using K-Means Clustering Algorithm.

From this, we can see that its a bad idea to use the text+title bag of words vector since its giving inappropriate clustering results. Hence, I dropped the idea of applying clustering on text+title bad of words vector due to poor clustering results.

3.3.2 DBSCAN

Next, I applied DBSCAN algorithm on the **vec1** with the eps set to 3 and minimum number of samples to consider set to 4. I got 2 clusters labeled as -1 and 0 which contain 2006 and 4 datapoints respectively. I got a negative Silhouette Score of -0.1259 and a CH Index of 1.107199, which is even worse than the results of K-Mean clustering.

3.3.3 Hierarchical Agglomerative Clustering

Here I first constructed the Dendogram from which we can inference the number of optimal clusters required. It is used to show relationships between similar sets of data.

From the Dendogram I obverse that the height of the tree is 6 and hence the optimal number of cluster is 6.

Now I apply hierarchical agglomerative clustering with complete linkage from *sklearn* library.

Now here, on evaluation, I get a good Sihouette Score of 0.46 and a huge CH Index of 4797.87 which shows that we have a very well separated tight clusters.

Here, the clusters have 837, 400, 86, 437, 79 and 266 datapoints each.

Next, I created the word clouds for the clusters to visualise the most frequent words in the clusters. Before that I also removed the be verbs such that they do not appear as the most frequent words in the clusters and we can get a real essence of the genres indicated by the clusters.

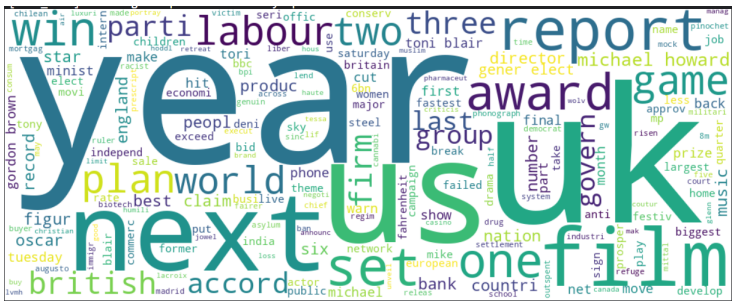
here is what we can infer from the outputs of the Word clouds:

Cluster 0: Films, Music and Entertainment

Cluster 1: Europe, United Kingdom

Cluster 2: Finance and Economy

Cluster 3: News and Job Market in different Nations



Cluster 4: European Union, World Politics and Brexit

Cluster 5: Sports and Games

4 Conclusion

What we can see here from the different clustering algorithms is that Hierarchical Clustering and Topic Modelling is the best possible choice for clustering the documents in NLP. I have used various techniques to optimise the clustering to make it better - applying 3 different clustering algorithms and then optimising on them using Elbow Method and Dendogram diagram.

Here are the final evaluation results:

Algorithm	Clusters	Silhouette Score	Calinski-Harabasz Index
K-Means$_1$	9	0.01149	9.57
K-Means$_2$	9	0.0118	8.64
DBSCAN	2	-0.1259	1.1072
Hierarchical Agglomerative	6	0.46	4797.87