

Text Document Clustering using K-Means and Dbscan by using Machine Learning



T.H. Feiroz khan, N.Noor Alleema, Narendra Yadav, Sameer Mishra, Anshuman Shahi

Abstract: With the growth of today's world, text data is also increasing which are created by different media like social networking sites, web, and other informatics and sources e.t.c . Clustering is an important part of the data mining. Clustering is the procedure of cleave the large & similar type of text into the same group. Clustering is generally used in many applications like medical, biology, signal processing, etc. Algorithm contains traditional clustering like hierarchal clustering, density based clustering and self-organized map clustering. By using k-means features and dbscan we can able to cluster the document. dbscan a part of clustering shows to a number of standard. The data sets will automatically evaluate the formulation of each and every part data through by the use of dbscan and k-means that will shows the clustering power of the data. document consists of multiple topic. Document clustering demands the context of signifier and form ancestry. Descriptors are the expression used to describe the satisfied inside the cluster.

Keywords: Text document clustering; k-means; Dbscan;

I. INTRODUCTION

It is an unsupervised learning technique. When we collect the data and when we apply for the k-means algorithm, it basically used to find the point in between the data and make a group of the similar data. Grouping similar body together give us discernment in to underlying patterns of different groups. For example, we can able to identify the group by their levels and could be able to maintain the data accordingly as per the clustering purpose we can able to use in e-commerce purpose and can be able to identify in most accurate format Most popular and widely used clustering algorithm is k-mean clustering, hierarchical clustering. Document clustering is relevant in areas such as search engine, web mining and information retrieval. Stratification Clustering algorithm can be stratified on the basis of:

a. Flat clustering

- b. Hard clustering
- c. Soft clustering
- a. Flat clustering: It don't have any plain or simple structure but it is a make even cluster. . It mainly
- b. points the difficult suitation of maintaining clusters automatically.
- c. Hard clustering : It deals with the tough work. Each and every text file is a part of the cluster..
- d. Soft clustering : it mainly uses the document for the distribution purpose on the basis of the object.

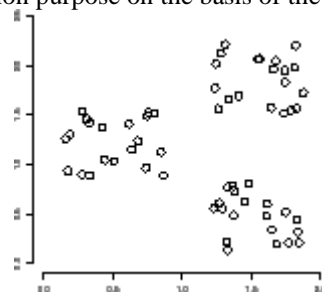


Fig.1: sample of clustering

Algorithm

Accumulation (hierarchical clustering)

K-Means (flat clustering, hard clustering)

EM Algorithm (flat clustering, soft clustering)

Hierarchical accumulation: we have directly uses the k-means algorithm as per the usable purpose.

II. LITERATURE SURVEY

Rupesh kumar Mishra, kanika saini, sakshi bagri; clustering is done on the basis of passage wise of each and every document by using k-means algorithm and the implementation part is done through by the use of the similarities purpose.

Sanjivani Tushar deokar , international journal of technology and engineering science, vol 1 (4),pp282-286, July 2013. Clustering was done by the mean and use of the new advancement of the technology and even they use the new technique for the clustering of large document and make in to the small document.

Li xinwu [electronic business department, jiangxi university of finance and economics, Nanchang, Jiangxi, 330013, they use the similar document clustering and the clustering is done through by the use of advantages and removing the disadvantage .

Shanshank paliwal and Vikram pudi [centre of data engineering, international institute os information technology, Hyderabad] MLDM 2012,LNAI 7376, PP.555-565. They use the clustering purpose by collecting the words and collecting it in the most appropriate format.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

T.H.Feiroz khan*, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: feirozkh@srmist.edu.in

N.Noor Alleema, Information Technology, SRM Institute of Science and Technology, Chennai, India. Email: noor25nrs@gmail.com

Narendra Yadav, B.Tech Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: ny344826@gmail.com

Sameer Mishra, B.Tech Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: sameermishra1999@gmail.com

Anshuman Shahi, B.Tech Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: ashahi7787@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Andrea tagarelli, george karypis DTD, MN 55455, USA , they use the clustering by using the large document and by using the process name as levarging and at the same time they maintain the multitopic format of the clustering purpose .

Sylvanin lamprier, tassadit amghar, Bernard levrat and Frederic saubion [AISMA] 2008, LNAI 5253,PP.69-82, they basically uses the clustering purposes by the use of information retrieval purpose and the purpose is done through by the use of the and make the implement of the non-relevant purpose.

Deng cai, shipeng yu, ji-rong wen,wei-yingma [SIGIR] 04 JULY 25-29, 2004, they basically uses the page segmentation method for the purpose of the clustering and make the document into the smaller document ..

Efficient phrase -based document indexing for web document clustering IEEE transaction on knowledge and data engineering] vol.16,NO.10, October 2004.kf. for the increase in the efficiency they basically uses the vector space model for the clustering purpose .

III. METHODOLOGY

k-MEANS ALGORITHM:

The main approach behind this is that its an updating cluster as this measures the distance part that occur repeatedly. The k-means algorithm is repeatedly occurs in many format basis as numpy, scipy and matplotlib.

```
import pandas as pd
import numpy as np
import plotly.offline as plt
import plotly.graph_objs as go
```

One of the main thing that can be noted as this is that once the program is executed it can be evaluated correctly and same thing may occurs but the thing is that once we reevaluated the chances of result may occur is different.

Initially we first insert the formal basis of the formula through by the use program and get the current and previous means.

TF-IDF:

It defines the repetation-inverse document repetation. Where the numerical statistics reflects how important a word in to the document to be cluster. The method is used to describe the how important the document is in the form of model called vector space. The term frequency define the number of times the words present in the document to the total number of words in the document.

The frequency having inverse document is defined by the dividing the total number of documents by the number of document having in the terms.

The $tf*idf =$

$$tf-idf_{t,d} = tf_{t,d} \times idf_t$$

KMEANS IMPLEMENTATION:

1. First we need to prepare the document cluster and need to collect the data sets.
2. After that we need to initialize the cluster centre for the next density as the counting the variable from each document. It holds the value that is already defined by the user.

3. We need to find the closest cluster centre by identifying the closeness of the document. The measure hold is the array and the obj.
4. We need to identify the new position of the cluster centre again we need to calculate the mean value finding the exact cluster process which shows the new positon of the cluster centre.

IMPLEMENTATION

DBSCAN:

In this work we want to give light for the implementation part by using dbscan algorithm. It is the kind of algorithm that mainly used in data clustering. This is basically deals to point to point connection where they deals with their neighbour points and give the result in more proper manner. It helps basically in many fields like medical science ,e-commerce e.t.c.

It takes two inputs. First one is the .csv file which contains the data (no headers). In 'main.py' change line 12 to:

`DATA = '/path/to/csv/file.csv'`

1.	The first and thirty-third books of Pliny's Natural history
2.	The Works of Josephus
3.	The History of the Decline and Fall of the Roman Empire Vol 5
4.	The Annals
5.	The Genuine Works of Flavius Josephus, the Jewish Historian
6.	Livy, Vol 3
7.	The History of the Decline and Fall of the Roman Empire Vol 1
8.	Livy, Vol 5
9.	The History of Rome, Vol 1
10.	The History of the Decline and Fall of the Roman Empire Vol 6
11.	The Works of Flavius Josephus, Vol 2
12.	The History

Fig.2: Sample data collection of title of books.

IV. WORK DONE

At first we collected the 24 books and with that books we seperate the name of the tile and the author name and we prepare it in ms excel. After finalizing the data ,we started our techniques called k-means algorithm. We imported each and every standard library of the python and we coded it in more properly through the using of TF-IDF as it calculate manually and gave us the proper result with their proper output and when it cluster the we got the separate level as level 0,level1,level2,level3, level4. So after ward we got the particular results with the plots.



So as by importing each and everything we started the coding for the implementation through by the use of the output of the coding part we just imported the structured data in the most format manner and we got the result of dbscan as we uses the algorithm like k-means algorithm , vector space algorithm and SVM.

V. FUTURE SCOPE

As we did by collecting the datasets of around 24 books and we got success in the particular field through by the implementation of k-means and dbscan but in futher future where technology is growing around so by that moment we will having not only 24 datasets instead of that we will have more than 24 datasets and that can be implemented by the same process of k-measns and dbscan and we can even apply in the field of the ecommerce as we can able to increase the demand of the datasets as when everyone has applied the products but it is an unstructured way so when we apply for the or counting the number of products we can able apply the number of out comes in the most accurate format and we will get the most appropriate result .

VI. RESULT

k-means:

x	y	label	title
-0.275	-0.59091	0	Livy, Vol 3
-0.13302	-0.31861	0	The Description of Greece
-0.27225	-0.59039	0	Livy, Vol 5
-0.2504	-0.59284	0	The History of Rome, Vol 1
-0.1617	-0.40072	0	The Historical Annals of Cornelius Tacitus
-0.22345	-0.51109	0	Roman History
0.210214	-0.15981	1	The first and thirty-third books of Pliny's Natural history
1.089988	0.339877	1	The History of the Decline and Fall of the Roman Empire Vol 5
0.43874	0.128323	1	The History of the Decline and Fall of the Roman Empire Vol 1
0.89583	0.385861	1	The History of the Decline and Fall of the Roman Empire Vol 6
1.241602	0.503294	1	The History of the Decline and Fall of the Roman Empire Vol 3
1.080956	0.351631	1	Gibbon's History of the decline and fall of the Roman empire
0.593261	0.122712	1	The History of the Decline and Fall of the Roman Empire Vol 2
-0.73693	0.834124	2	The Works of Josephus
-0.73974	0.867467	2	The Genuine Works of Flavius Josephus, the Jewish Historian
-0.88173	0.957245	2	The Works of Flavius Josephus, Vol 2
-0.86168	0.871688	2	The Works of Flavius Josephus, Vol 3
-0.25339	-0.41657	3	The History of the Peloponnesian War Std
-0.2682	-0.28756	3	The History of the Peloponnesian War Vol 1
-0.07074	-0.33822	4	The Annals
-0.10122	-0.37804	4	The History
-0.1116	-0.33566	4	The Histories of Gaius Cornelius Tacitus
-0.06901	-0.03586	4	Dictionary of Greek and Roman Geography
-0.14056	-0.40595	4	The History of Rome, Vol 3

Fig.3. The result of k-means

As here the datasets are arrange in the format manner with particular distinguish of the different levels. Plot of k-means:

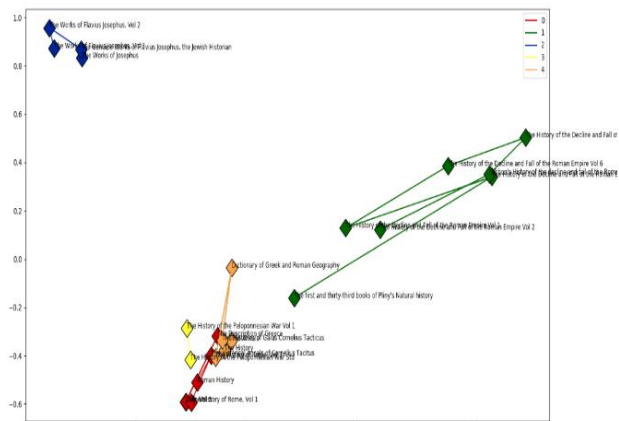


Fig.4: the plot of the k-means

The plots of k-means has shown in many different ways as when we applied the k-means clustering program then it applied in many different colour on the basis of their levels as red color is shows the level 0, green color is shows in the

level 1, blue color is shown in the level 2, yellow colour is shown the level 3, orange color shows the level 4. The plot representation is the main important thing for detailed information and can be clearly defined with the proper object. The plot can be fully defined with the proper information .

DBSCAN:

x	y	label	title
0.643991	-0.27535	0	Livy, Vol 3
-0.6776	0.261309	0	The Description of Greece
0.685437	-0.11807	0	Livy, Vol 5
0.451166	-0.10607	0	The History of Rome, Vol 1
-0.03297	0.053741	0	Dictionary of Greek and Roman Geography
-0.15114	0.71375	0	The History of the Peloponnesian War Std
-0.02161	0.681523	0	The History of the Peloponnesian War Vol 1
0.030937	0.322024	0	Roman History
0.200268	-0.68505	0	The History of Rome, Vol 3
0.380297	0.359433	1	The Works of Josephus
0.512752	0.327892	1	The Genuine Works of Flavius Josephus, the Jewish Historian
0.489494	0.44636	1	The Works of Flavius Josephus, Vol 2
0.426801	0.496061	1	The Works of Flavius Josephus, Vol 3
-0.63711	-0.00826	2	The first and thirty-third books of Pliny's Natural history
-0.53143	-0.34066	2	The History of the Decline and Fall of the Roman Empire Vol 5
-0.09956	-0.39635	2	The History of the Decline and Fall of the Roman Empire Vol 1
-0.37255	-0.2587	2	The History of the Decline and Fall of the Roman Empire Vol 6
-0.43844	-0.32822	2	The History of the Decline and Fall of the Roman Empire Vol 3
-0.43528	-0.4716	2	Gibbon's History of the decline and fall of the Roman empire
-0.1963	-0.59913	2	The History of the Decline and Fall of the Roman Empire Vol 2
0.21701	-0.37124	3	The Annals
-0.43228	0.331249	3	The History
-0.4314	0.50355	3	The Histories of Gaius Cornelius Tacitus
0.419525	-0.53819	3	The Historical Annals of Cornelius Tacitus

Fig.5: The result of implementation dbscan

After applying the k-means algorithm we got some results ,in that results we applied the algorithm of dbscan and we applied it in most accurate format. Plots of dbscan:

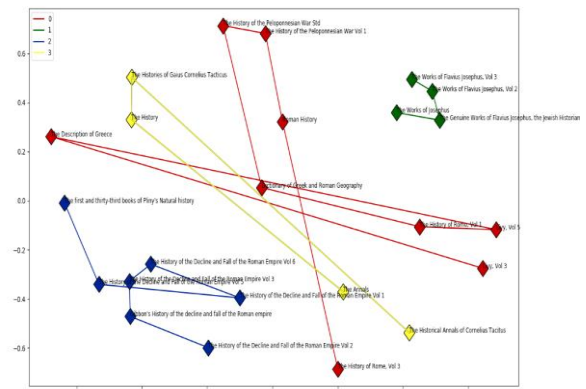


Fig.6. The plot of dbscan

VII. DISCUSSION AND CONCLUSION

In this paper, it is discussed about the document clustering. Document clustering is process of cluster the document and finds the similar document in one levels and get the division in between the levels. I have mentioned about the k-means and their useful with the classification. I have also discussed about the k-means algorithm with the basis of the algorithm. I have also discussed about the implementation of the k-means algorithm and use of the algorithm.



It discussed about the getting the better similarity of the results in the document clustering. As per the section it is totally applied in many sectors like in e-commerce and medical science as the clustering part is one of the most valuable and make to get the work better and faster.

REFERENCE

1. Rupesh kumar Mishra, kanika saini, sakshi bagri : Text document clustering on the basis of inter passage approach by using k-means, IEEE xplore,2015
2. Sanjivani Tushar deokar international journal of technology and engineering science, vol 1 (4),pp282-286, July 2013.
3. Li xinwu [electronic business department, jiangxi university of finance and economics, Nanchang,Jiangxi, 330013, china liyue7511@163.com
4. Shanshank paliwal and Vikram pudi [centre of data engineering, international institute os information technology, Hyderabad] MLDM 2012,LNAI 7376, PP.555-565.
5. Andrea tagarelli, george karypis DTD, MN 55455, USA
6. Sylvanin lamprier, tassadit amghar, Bernard levrat and Frederic saubion [AISMA] 2008, LNAI 5253,PP.69-82
7. Deng cai, shipeng yu, ji-rong wen,wei-yingma [SIGIR] 04 JULY 25-29, 2004
8. Efficient phrase -based document indexing for web document clustering[IEEE transaction on knowledge and data engineering] vol.16,NO.10, October 2004.kf

AUTHORS PROFILE



T.H.Feiroz Khan received his M.E. degree in Computer science Engineering from Annamalai University, Tamilnadu in 2004. Currently pursuing PhD in school of computing in the Domain of Wireless Networking. He is a currently Assistant Professor in SRM Institute of Science and Technology, Chennai. He has teaching work experience of 15 years in the Engineering and Technology. He is member of International Association of Engineers (IAENG).His research interest includes Wireless System and Machine Learning.



N. Noor Alleema obtained her Bachelor degree in Computer Science from Madras University, Chennai in 2000-2004 and Master degree in the Department of Computer Science and Engineering from Anna University, Chennai in 2006. Currently, she is a PhD candidate in the Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai. Her main research interest now on Mobile Ad Hoc networks



Narendra Yadav is currently pursuing his Bachelors in Science and Technology degree in SRM Institute of Science and Technology. He is also a member of various technical clubs of SRM IST. He is currently working on a ML projects on sentimental analysis and data science.



Anshuman Shahi is currently pursuing his Bachelors in Science and Technology degree in SRM Institute of Science and Technology. He is also a member of various technical clubs of SRM IST.



Sameer Mishra is currently pursuing his Bachelors in Science and Technology degree in SRM Institute of Science and Technology. He is also a member of various technical clubs of SRM IST.