Volume 4, Issue 4, April 2019 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Distance Based Similarity Between Text Documents Using Euclidean Distance

Devendra Patil
M Tech. (CSE) Semester
Lord Krishna College of Technology
Indore M.P. India

Vijay Kumar Verma
Asst. Professor Dept. C.S.E.
Lord Krishna College of Technology
Indore M.P. India

Abstract: ***There are several parameters by which similarity can be evaluated. The first category of similarity evaluation is based on the document size and structures the length of the document, the number of paragraphs, number of sentences, average number of characters per word, average number of words per sentence. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system. The main problem is to calculate the distance between the different documents. In the proposed work we calculate distance between different documents. So we have to discover out which of the distance gives output (that is the recommendation of the items) in least time and efficiently. And what are the advantages and disadvantages of different similarity measure. The aim of this work was to find the most optimum distance value. Euclidean distance method with only terms and Euclidean distance method with terms and tokens. By comparing these two distances find which method gives more correct result. Results from this study are the Euclidean distance method gives the best value of distance matrix***

Keywords: Similarity, Distance, Text, Euclidean, Term, Tokens

I. INTRODUCTION

There are several parameters by which similarity can be evaluated. The first category of similarity evaluation is based on the document size and structure the length of the document, the number of paragraphs, number of sentences, average number of characters per word, average number of words per sentence etc. The second category is based on "style", whether the contents have been written in the first person conversational style or in thethird person and so on. Thirdly, similarity can be based on the set of wordsused in the document. For example the original text of the novel "A Tale of two cities" written by Charles Dickens may contain 20,000 distinct words,whereas the same novel rewritten for seventh standard students may contain only a set of 1000 words. The fourth category of similarity is "content Similarity" which reflects to what extent the contents of the two documents is alike. This category is adopted throughout this thesis wherever similarityis talked of hereafter.26The similarity between two documents is computed by any one ofthe several similarity measures based on the two corresponding featurevectors, e.g. cosine, dice, and jaccard measure. The common framework forthe document clustering model starts with the representation of any documentas a feature vector of the terms (words) that appear in the documentcollection. Let D = (D_1,D_2, . . .,D_n) denote the collection of documents,where 'n' is the number of documents in the collection. Let T = (T_1, T_2, . ..T_m) represent all the terms that occurred in the document collection 'D'.Here 'm' is the number of unique terms in the document collection. In mostclustering algorithms, the dataset to be clustered is represented as a set ofvectors, where each vector corresponds to a single object and is called the feature vector.

The data stored in the computer can be in any one of the form (i) structured (ii) semi structured and (iii) unstructured. The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Huge amount of data today are stored in text databases and not in structured databases. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining.
Approximately 80% percent of the corporate data is in unstructured format. The information retrieval from unstructured text is very complex as it contains massive information which requires specific processing methods and algorithms to extract useful patterns. As the most likely form of storing information is text, text mining is considered to have a high value than that of data mining. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, information extraction, computational linguistics and natural language processing

II. TEXT MINING PROCESS

Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The great deal of studies done on the modeling and implementation of

semi structured data in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and has been written by someone else. The problem is in result as it is not relevant to users need. This is the goal of text mining to discover unknown information which is not known and yet not written down [16].
Text mining process starts with a document collection from various resources. Text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

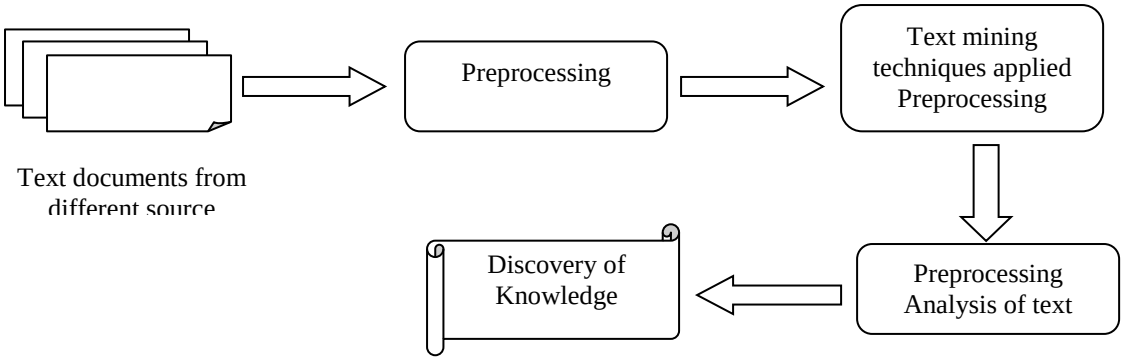

Figure 1 Text mining process

III. LITERATURE SURVEY

2011 James Thomas,John McKnightproposed "Applications of text mining withinsystematic reviews". They describe the application of four text mining technologies, namely, automatic termrecognition; document clustering, classification and summarization, which support the identificationof relevant studies in systematic reviews. They showed that text mining technologies to improvereviewing efficiency are considered and their strengths and weaknesses explored.They conclude that these technologies do have the potential to assist at various stages of the reviewprocess. They are relatively unknown in the systematic reviewing community, and substantialevaluation and methods development are required before their possible impact can be fully assessed.Text mining technologies offer ways forward in reducing the amount of time systematic reviews: to; includethe automatic identification of concepts and related documents, thus reducing the time of searching, filteringand categorizing relevant documents and last supporting the reviewers in the description of research identifiedby using automatic text summarization. Applying these technologies in an interactive manner entails a moveaway from identifying every relevant paper from a limited search, to identifying a proportion of relevant papersfrom a much wider base.As these technologies are now deployed in different areas, further methodological and evaluative work isneeded to develop methods and an evidence base for their use[2].

In 2012 Su Gon Cho and Seoung Bum Kim proposed "Identification of Research Patterns and Trends throughText Mining".They crawled the keywords from theabstracts in IIE Transactions, one of the representative journalsin the field of Industrial Engineering from 1969 to 2011. They applied a low-dimensional embedding method, clusteringanalysis, association rule, and social network analysis to findmeaningful associative patterns of the keywords frequentlyappeared.Theyrevealed research trends andpatterns of the Industrial Engineering field from one of therepresentative journals in Industrial Engineering by usingtext mining. They employed the dimensional reduction method,clustering analysis, and social network analysis to draw out meaningful patterns of the keywords and research trends overtime. Social network analysis visualizes and summarizesassociation patterns of the keywords. They stimulate further investigation in applying appropriate textand data mining tools to various applications in both academia and industry [3].

In 2013 K. L. Sumathy M. Chidambaram proposed "Text Mining: Concepts, Applications, Tools and Issues – An Overview".Theygive an overview of concepts, applications, issues and tools used for text mining. Due to the rapid growth of digital data made available in recent year's knowledge discovery and data mining have attracted great attention with a forthcoming need for turning data into useful information and knowledge. Consequently there is growing research interest in the topic of text mining. In general text mining consists of analyzing large amount of text documents by extracting key phrases; concepts etc., and prepare the text processed for further analysis with data mining techniques. In this paper an overview of concepts, applications, tools and issues of text mining is presented to give the researchers to carry it to the next level[4].

In 2014 Sonali Vijay Gaikwad proposed "Text Mining Methods and Techniques". They presented survey paper we discuss such successful techniques and methods to give effectiveness over information retrieval in text mining. These types of situations where each technology may be useful in order to help users are also discussed. They presented overview techniques, methods and challenging issue in text mining. The focus has been given on fundamental methods for conducting text mining. They addressed the most challenging issue in developing text mining systems. Four methods of text mining term based, phrase based, concept based and pattern taxonomy model discussed. Term based approach suffer from polysemy and synonymy while phrase based approach performs better as phrase carries more

semantics like information and is less ambiguous. Two terms can have same frequency fromstatistical analysis this problem can be solved by concept based approach by finding term contributing more meaning. In pattern based approach pattern taxonomy is formed to solve low frequency problem and misinterpretation problem. Then in next half Text Mining is discussed with its various techniques and usages. To extract structured information from the unstructured text Information Extraction is used. In information extraction data mining techniques can be applied for getting useful patterns or knowledge from the documents. To produce the relevant information from the corpus is known as summarization. Classification is a supervised technique because before it can be used to classify the newly arrived document it has all the input output patterns which are used to train the model. Clustering is unsupervised learning technique because no pre-defined input-output patterns are there. According to summery of documents text is clustered, this process is known as clustering. To provide improved understandable information for mining the documents Graphical Visualization is used[5].

In 2014 K. Thilagavathi, V. ShanmugaPriya proposed **"**A Survey on Text Mining Techniques".They presented and discussed different method of text categorization and cluster analysis or text documents. In additionof that a new text mining technique is proposed for future implementation. They discuss various techniques and methods for efficient and accurate text mining. In addition of thatthe efficient algorithms are also learned. Due to observation a promising approach is obtained given in. Accordingto the analyzed methods an improvement over this is suggested. Infuture the proposed technique isimplemented using JAVA technology and the comparative results are provided [6].

In 2015 E. Alan Calvillo,Alejandro Padillaproposed "Searching Research Papers Using Clustering and Text Mining".They proposed a better classification of research papers,the architecture works with a database of knowledge related with the topics of programming, databases and operating systems.Theyevaluates a way to optimize the information to be located within a structured framework with an initial knowledge base that helps the easy categorization of information by implementing a clustering for fast search and location as well as a textual analysis entered by the user as a basis for consultation, as future work is to implement an automatic learning that allows the steady increase in the manipulated texts. That kind of techniques allows making a best search engine using database knowledge to work with filter, wrapper or even ontology. The use of text mining technologies are not used in web search or meta search, that kind of tools usually use only meta crawler to classify the information the current work shows how the search engine can be used and it should make a benchmark between the filter, wrapper and ontology to the next work. They shows as future work is consider to extend the search engine into another kind of devices using Android or IOS that's to generate a portable application to make searches in different kind of devices[7].

In 2015 Nadir Zaniniand VikasDhawanproposed "Text Mining: An introduction to theory and someApplications".The key advantage provided by TM is the opportunity to exploit textrecords, on a very large scale. They briefly described the techniques of TM and some of its applications.TM has a variety of potential applications in the field of education.In formative and summative assessment, for instance, it could be used tounderstand trends in vocabulary usage over time and the use of spellingand punctuation. To date, these applications have been carried out byteachers and assessment experts without using advanced techniquessuch as TM, but TM allows the possibility of implementing theseapplications on a more comprehensive scale. The developments in NLPallow educational professionals to analyze the language structure of avast amount of text documents in just a few minutes, plus the ongoingdevelopments in this field could result in an increase in the accuracy ofthe findings.The availability of novel data could lead, at least in principle, to novelmeasurement and research designs to address old and new researchquestions[8].

In 2016 Ramzan Talib, Muhammad KashifHanif proposed "Text Mining: Techniques, Applications and Issues". Theybriefly discuss and analyze the text mining techniques and theirapplications in diverse fields of life. They discuss the issues in thefield of text mining that affect the accuracy and relevance ofresults are identified. Theypresent a brief overview oftext mining techniques that help to improve the text miningprocess. Specific patterns and sequences are applied in orderto extract useful information by eliminating irrelevant detailsfor predictive analysis. Selection and use of right techniquesand tools according to the domain help to make the text miningprocess easy and efficient. Domain knowledge integration,varying concepts granularity, multilingual text refinement, andnatural language processing ambiguity are major issues andchallenges that arise during text mining process. In futureresearch work, they will focus to design algorithms which willhelp to resolve issues presented in this work [9]

in 2016 Abhishek Kaushik and SudhanshuNaithaniproposed "A Comprehensive Study of Text Mining Approach".They proposed a review of text mining techniques, tools and various applications Text mining is one of thefastest growing fields today. With the passage of time its importance is only going to increase because rate of data production is very high. Automatic text mining has a long way to go because it is not in the position to challenge the human's capabilities. From last few years text mining (sentiment analysis) is largely being used to predict the results of elections at national and state level which is most significant development in the field recently. Onaccount of growing interaction of text mining to some other fields, especially with machine learning, visualization and natural language processing, it is possible to design more effective and useful text mining system. Text mining is also being used by industry and it is generating the sheer amount of knowledge which cannot even consume by humans. They tried to present an overview of text mining approach with its techniques, tools and applications[10].

In 2016 R. Janani, Dr. S.Vijayarani proposed "Text Mining Research: A Survey".They discussed about the text mining techniques and its applications. Text mining is used to extract interesting informationor knowledge or pattern from the unstructured texts that are from different sources. It converts the words andphrases inunstructured information into numerical values which may be linked with structured information in database andanalyzed with ancient data mining techniques. There are many techniques used in text mining such as informationextraction, information retrieval, natural language processing (NLP), query processing, categorization and clustering Data Mining is the important as well as active research area helps to extract helpful patterns from the data.These patterns generated facilitate decision making in industries. Text mining is also crucial field that deals withunstructured or semi structured data. They delineated the

various text mining techniques such asInformation Extraction, Information retrieval, Natural Language processing, Categorization and Clustering. They alsodefined text mining processing flow, applications of text mining and issues in text mining. Mining text indifferent languages may be a major problem, since text mining tools and techniques ought to be able to work withseveral languages and multilingual languages. Integrating a domain knowledge base with text mining engine wouldincrease its efficiency, especially within the information retrieval and information extraction phase [11].

In 2017 Binling Nie and Shouqian Sun proposed "Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research".They giveinformation about scientificliterature in design research. A combination of clustering and metric analysis led toshaping four academic branches and summarizing each academic branch. Then, research trends andthe evolution for each academic branch are explored. We perform a two-dimensional text mining approach, including bibliometric and network analysis, in order to detect trends of major academicbranches. Specifically, the bibliometric characterization aims to assess design research area outputs,while the network analysis intends to reveal research trends in each academic branch of designresearch and the evolution of core research themes. They presents a bibliometric, network-theoretic and text-based analysis of the designresearch area during the last 12-year period

IV. PROBLEM STATEMENT

The main problem is to calculate the distance between the different documents. For calculating distance between different documents we used Euclidean distance,Murkowski distance, Manhattan Distance. So we have to discover out which of the distance gives output (that is the recommendation of the items) in least time and efficiently. And what are the advantages and disadvantages of different similarity measure.

OBJECTIVES

There are several algorithms and methods have been text document clustering. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are-

1) Apply Euclidean distance measures for distance calculation for one title with another title using terms only.
2) Apply Euclidean distance measures for distance calculation for one title with another title using terms and tokens.
3) find out which distance method is more accurate

V. PROPOSED APPROACH

When talking about patterns, the distance and similarity concepts are in a way reciprocal. In fact we often use the term distance to convey the idea of dissimilarity. Through the use of a association or similarity measure we try to quantify the likeness between patterns (Van Rijsbergen 1979). So when comparing patterns, it is very useful if they are represented in a space that has a metric. This is a property of any set of elements characterized by the distance function between all pairs of elements, denoted d(**x,y**) for elements **x** and **y** e.g. (Kohonen 2001). For the choice of the distance, the following conditions must hold.

d (x, y)≥0 where equality holds if and only if x=y
d (x, y)=d(y, x) symmetry
d (x, y)≤d(x, z)+d(z, y) triangle inequality

Euclidean distance coordinate system is calculated using the formula

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Murkowski distance: A popular distance measure
Where $i = (x_{i1}, x_{i2} \ldots x_{ip})$ and $j = (x_{j1}, x_{j2} \ldots x_{jp})$ are two p dimensional data objects, and h is the order.

$$d(i,j) = \sqrt[h]{(x_{i1} - y_{j1})^h + (x_{i2} - y_{j2})^h + \ldots + (x_{ip} - y_{jp})^h}$$

Special cases of Minkowaski Distance

- h=1 Manhattan Distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- h=2 Euclidean distance
-

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

ILLUSTRATE WITH EXAMPLE

Consider a simple table with 10 documents and words present in the doc

Table 3.1 terms and tokens

Doc/words	objects	memory	frequency	distance	metric	domain
Doc1	7/55	5/55	8/55	5/55	0	0
Doc2	0	4/60	6/60	0	6/60	8/60
Doc3	6/65	0	0	3/65	3/65	7/65
Doc4	0	0	0	4/43	0	5/43
Doc5	3/48	0	4/48	5/48	0	0
Doc6	0	5/60	8/60	3/60	0	7/60
Doc7	6/45	0	3/45	2/45	4/45	0
Doc8	0	8/49	6/49	5/49	7/49	0
Doc9	6/56	0	8/56	4/56	6/56	0
Doc10	5/68	0	6/68	0	7/68	6/68

3.5.1 Euclidean distance between documents using frequency of the words only

Calculate simple Euclidean distance between documents using number of term only consider first and seconds documents

Doc/words	objects	memory	frequency	distance	metric	domain
Doc1	7	5	8	5	0	0
Doc2	0	4	6	0	6	8

$$d(doc1, doc2) = \sqrt{(7-0)^2 + (5-4)^2 + (8-6)^2 + (5-0)^2 + (0-6)^2 + (0-8)^2}$$

$$d(doc1, doc2) = \sqrt{(7)^2 + (1)^2 + (2)^2 + (5)^2 + (6)^2 + (8)^2}$$

$$d(doc1, doc2) = \sqrt{49 + 1 + 4 + 25 + 36 + 64}$$

$$d(doc1, doc2) = \sqrt{179}$$

$$d(doc1, doc2) = 13.37$$

By using Euclidean distance with only term distance between documents 1 and documents 2 are 13.37.

Euclidean distance between documents using frequency divided by number of terms

Calculate simple Euclidean distance between documents using number of term and number of tokens only consider first and seconds documents

Doc/words	objects	memory	frequency	distance	metric	domain
Doc1	7/55	5/55	8/55	5/55	0/55	0/55
Doc2	0/60	4/60	6/60	0/60	6/60	8/60

$$d(doc1, doc2) = \sqrt{(7/55 - 0)^2 + (5/55 - 4/60)^2 + (8/55 - 6/60)^2 + (5/55 - 0)^2 + (0 - 6/60)^2 + (0 - 8/60)^2}$$

$$d(doc1,doc2) = \sqrt{(0.1272)^2 + (0.02423)^2 + (0.0454)^2 + (0.0909)^2 + (0.1)^2 + (0.1333)^2}$$

$$d(doc1,doc2) = \sqrt{0.001617 + 0.00058 + 0.002061 + 0.008262 + 0.01 + 0.01776}$$

$$d(doc1,doc2) = \sqrt{0.04022}$$

$$d(doc1,doc2) = 0.2005$$

$$d(doc1,doc2) = 0.2005 * 100$$

$$d(doc1,doc2) = 20.05$$

COMPARISON

Table 2 Distances between documents using terms only

Doc/Terms	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
D1	1	13.37	12.32	12.80	7.54	10.09	8.71	10.53	7.93	11.95
D2	13.37	1	10.34	10.63	12.40	7.14	11.74	10.29	11.66	6.78
D3	12.32	10.34	1	6.40	9.32	11.18	7.74	14.62	10.24	7.93
D4	12.80	10.63	6.40	1	7.14	9.89	9.48	13.22	12.68	11.26
D5	7.54	12.40	9.32	7.14	1	10.14	5.91	13.22	7.87	10.86
D6	10.09	7.14	11.18	9.89	10.14	1	12.32	10.72	12.12	10.63
D7	8.71	11.74	7.74	9.48	5.91	12.32	1	11.26	5.74	4.79
D8	10.53	10.29	14.62	13.22	13.22	10.72	11.26	1	10.29	12.24
D9	7.93	11.66	10.24	12.68	7.87	12.12	5.74	10.29	1	7.61
D10	11.95	6.78	7.93	11.26	10.86	10.63	4.79	12.24	7.61	1

Table 3 Distance between documents using terms and tokens

Doc/Terms	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
D1	1	16.3	17.3	13.8	4.5	13.0	6.73	4.23	4.21	14.12
D2	16.3	1	15.3	17.6	21.4	8.12	21.7	18.2	21.6	7.27
D3	17.3	15.3	1	7.41	7.41	7.41	7.41	7.41	7.41	7.41
D4	13.8	17.6	7.41	1	0.124	0.124	0.124	0.124	0.124	0.124
D5	4.5	21.4	12.3	12.4	1	0.091	0.091	0.091	0.091	0.091
D6	13	8.12	15.1	16.8	9.11	1	0.223	0.223	0.223	0.223
D7	6.73	21.7	8.72	15.8	8.52	22.3	1	0.091	0.091	0.091
D8	4.23	18.2	22.6	18.2	19.2	13.7	9.21	1	0.081	0.081
D9	4.21	21.6	15.2	20.6	12.7	17.1	8.7	8.11	1	10.7
D10	14.12	7.27	13.9	18.2	17.8	9.61	2.27	9.22	10.7	1

Table 5.6 Distance of doumnt1 with other documents

Documents	Terms	Terms and Tokens
(D1,D2)	13.37	16.3
(D1,D3)	12.32	17.3
(D1,D4)	12.8	13.8

(D1,D5)	7.54	4.5
(D1,D6)	10.09	13
(D1,D7)	8.71	6.73
(D1,D8)	10.53	4.23
(D1,D9)	7.93	4.21
(D1,D10)	11.95	14.12

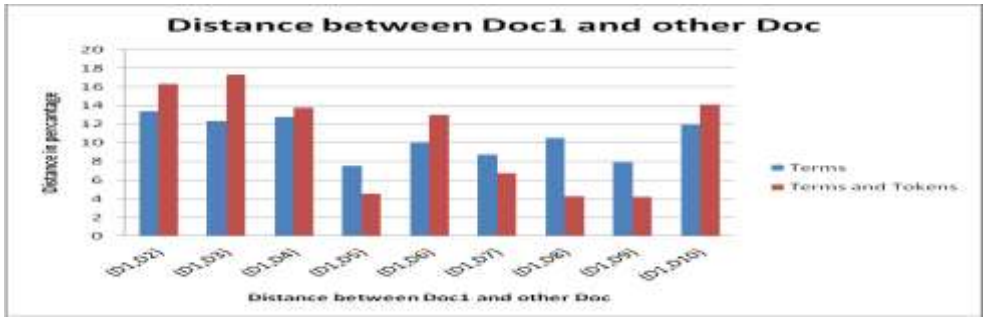

Figure 2 Distance of doumnt1 with other document

CONCLUSION

Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. Distance between different texts documents can be calculated using Euclidean distance with terms only and Euclidean distance with terms and tokens. So discover which of the distance gives more correct output is difficult. Euclidean distance with terms only is more correct as compare to Euclidean distance with terms and token. In the proposed work we compare these two approaches to find the correct distance between two text documents. By the experimental analysis we calculate the distance between 10 documents.

REFERENCE

1. James Thomas, John McNaught and Sophia Ananiadoub Applications of text mining within systematic reviews Received 2 September 2010, Revised 24 January 2011, Accepted 28 January 2011 Published online 11 April 2011 in Wiley Online Library(wileyonlinelibrary.com) DOI: 10.1002/jrsm.27.
2. Su Gon Cho and Seoung Bum Kim Identification of Research Patterns and Trends through Text Mining International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012 Manuscript received March 20, 2012; revised April 12, 2012.The authors are with the School of Industrial Management Engineering, Korea University, Seoul, Korea (e-mail: sbkim1@korea.ac.kr.)
3. K. L. Sumathy, M. Chidambaram, Text Mining: Concepts, Applications, Tools and Issues An Overview International Journal of Computer Applications (0975 – 8887) Volume 80 No.4, October 2013.
4. Sonali Vijay Gaikwad Archana Chaugule Text Mining Methods and Techniques International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014.
5. K. Thilagavathi, V. Shanmuga Priya, A Survey on Text Mining Techniques International Journal of Research in Computer Applications and Robotics www.ijrcar.com Vol.2 Issue.10, Pg.: 41-50 October 2014.
6. E. Alan Calvillo, Alejandro Padilla1,Jaime Muñoz1 Searching Research Papers Using Clustering and Text Mining All content following this page was uploaded by Julio Cesar Ponce on 12 November 2015.
7. Nadir Zanini and Vikas Dhawan Research Division Text Mining: An introduction to theory and some applications. A Cambridge Assessment publication. www.cambridgeassessment.org.uk/research-matters/ UCLES 2015.
8. Ramzan Talib, Muhammad Kashif Hanify, Shaeela Ayeshaz, and Fakeeha Fatima Text Mining: Techniques, Applications and Issues (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.
9. Abhishek Kaushik and Sudhanshu Naithani A Comprehensive Study of Text Mining Approach IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.2, February 2016.
10. R. Janani, Dr. S.Vijayarani Text Mining Research: A Survey International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization)Vol. 4, Issue 4, April 2016.
11. Binling Nie and Shouqian Sun Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research Institute of Industrial Design, College of Computer Science, Zhejiang University, Hangzhou 310027, China; ssq@zju.edu.cn * Correspondence: nbl1221@zju.edu.cn; Tel.: +86-131-0770-2260 Academic Editor: Yang Kuang Received: 23 December 2016; Accepted: 11 April 2017; Published: 15 April 2017.
12. D. Jasmine Guna Sundari A Study of Various Text Mining Techniques International Journal of Advanced Networking & Applications (IJANA) Volume: 08, Issue: 05 Pages: 82-85 (2017) Special Issu Snezhana Sulova Using text mining to classify research papers Using text mining to classify research papers July 2017 with 3,109 Reads DOI: 0.5593/SGEM2017/21/S07.083Conference: 17th International Multidisciplinary Scientific Geo Conference SGEM 2017, Volume: 17
13. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan Using Text Mining Techniques for Extracting Information from Research Articles © Springer International Publishing AG 2018 K. Shaalan et al. (eds.), Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740, https://doi.org/10.1007/978-3-319-67056-0_18.
14. Latinka Todoranova, Bonimir Penchev, Radka Nacheva Using Text Mining To Classify Research Papers 17 th International Multidisciplinary Scientific Geo Conference SGEM 2017.
15. Morgan Kaufmann Data Mining: Concepts and Techniques Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791.