Auto Text Summarization and Categorization

Richa Tanna[1], Akanksha More[2], Ankita Thakur[3], Prof. D. J. Dhangar[4]

[1,2,3] Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India

Abstract - *Automatic text summarization is undergoing wide research and gaining importance as the availability of online information is increasing. Automatic text summarization is to compress the larger original text into shorter text called as summary. Abstraction and Extraction are the two main methods to carry out text summarization. Our approach uses extractive method. Summarization by extraction involves identifying important features and extracting sentences based on their scores. 30 documents from news based URLs are used as input. After pre-processing the input document, five features are used to calculate their score for each sentence. In this paper fuzzy logic method is proposed for improvement in the extraction of summary sentences.*

Key Words: Categorization, Feature matrix, Fuzzy Logic, Deep Learning Algorithm, Text Summarization.

1. INTRODUCTION

What is a Summary?
Summary is basically the shorter version of the larger documents and it is just showing the most important part of the documents or giving us the brief idea of what the document has to convey.

Goal of Auto Text Summarization:

With the super growth in the quantity and the complex versions of the documents on the internet on the every single topic, it has become very important to provide more developed mechanism for the user to find the exact and particular important information that a document want to convey. Text summarization has become an important and handy tool for the users by helping users to interpret the large volumes of text data that is available in the documents.

Summarization Methods:

There are two main important methods of summarization. Firstly, there is abstractive summarization and other one being the extractive summarization. Here our approach uses extractive method. Summarization is carried out by extracting the features from each sentences and based on the higher score of feature matrix, we generate the summary.

1.1 Purpose

Business associates, academic researchers, students, and almost every person go through many documents every day to keep him/her updated, and a large amount of their time is spent just to figure out what exactly the documents are about and what they want to convey. By extracting important sentences and creating summaries, it is possible to quickly assess whether or not a document is worth reading. Auto text summarization is most useful for students and authors. Imagine that you are able to automatically generate an abstract based for your research paper or for any chapter in a book in a clear and easy way that is faithful to the original source material. With the growing amount of data in the world, interest in the field of auto summary generation has been widely increasing as it is reducing the manual effort of a person working on large amount of documents.

1.2 Scope

Auto summarization involves reducing a text file into a passage or paragraph that can convey the important points of the text file. The searching of important information or to understand what the large text file wants to convey is very difficult job as well as it is very time consuming for the users thus to automatically extract the important information or summary of the text file comes in as a handy tool. This auto generated summary helps the users to reduce time instead of reading the whole text file and it provides useful information quickly from the large documents. In today's world to extract information from the World Wide Web is very easy, but extracted information is a huge text file. With the rapid growth of the World Wide Web, information overload is becoming a problem for an increasing large number of people. Auto summarization can be a very important solution to reduce the information overload problem on the web.

2. SYSTEM ANALYSIS

2.1 Existing System

Yan Liu et al have proposed a document summarization framework via deep learning model, which has demonstrated distinguished extraction ability in document summarization. The framework consists of concepts extraction, summary generation and reconstruction validation. A query-oriented extraction technique has been concentrated information distributed in multiple documents to hidden units layer by layer. Then, the whole deep architecture was fine-turned by minimizing the information loss in reconstruction validation part. According to the concepts extracted from deep architecture, dynamic programming was used to seek most informative set of sentences as the summary. Experiments on three benchmark dataset demonstrate the effectiveness of the framework and algorithm.

2.2 Disadvantages of Existing System

The existing system is not user friendly and it is difficult to compute summaries using the existing methods and summaries suffer from inconsistencies and lack of balance results in lengthy summary. Detailed Information is not present. Existing methods does not give quick overview. It provides authors view not the user friendly view. Existing systems cannot summarize multiple documents of the same type. Existing systems cannot handle different languages.

3. PROPOSED SYSTEM

Proposed System consists of:

1) Preprocessing: Initially, the input to the proposed approach is a set of document from DUC 2002.Dataset that has to be summarized. The document utilized for text summarization is organized by a set of pre-processing steps like, sentence segmentation, stop words removal and stemming.

2) Segmentation: It is performed by identifying the delimiter commonly denoted by "." called as full stop. This step is used to separate the sentences in the document. It is mainly useful for the user to understand each individual sentence which is there in the document.

3) Stop Words Removal: Stop words are removed mainly to reduce the insignificant and noisy words. These are predefined words such as a, an, in, by, etc., are called stop words which are filtered out before the pre-processing phase from the documents.

4) Stemming: Stemming is process of bringing the word to its base or root form for example using words singular form instead of using the plural. It basically removes the prefix and suffix of the concerned word to get the base form. There are many more number of algorithms, which are called as stemmers used to perform the stemming process.

5) Training Phase: On behalf of the training phase, the proposed approach defines five features sets. The feature sets are listed as follows,

(i) Title Similarity Feature: The ratio of the number of words in the sentence that occur in title to the total number of words in the title helps to calculate the score of a sentence for this feature and it is calculated by the formula given below

$$Title\ Feature\ (f1) = \frac{S \cap t}{t}$$

(ii) POS Tagger Feature: To calculate the positional score of sentence, the proposed approach considers the following conditions. If the sentence given is in the starting of the sentence or the last in the sentence of the paragraph then the feature value $f2$ is assigned as 1. Else if the sentence is in the middle of the paragraph then the feature value of f 2 is assigned as 0.

(iii) Term Weight Feature: The Term Frequency of a word will be given by TF (f, d) where f is the frequency of the given word and d is text present the document. The Total Term Weight is calculated by Term Frequency and IDF for a document .Here IDF denotes the inverse document frequency which just implies that the term is common or rare across all documents.

$$IDF(t, D) = \log\left(\frac{D}{d \in D : t \in d}\right)$$

(iv)Concept Feature: The concept feature from the text document is retrieved using the mutual information and windowing process. In windowing process a virtual window of size „k‟ is moved over document from left to right. Here we have to find out the co-occurrence of words in same window and it can be calculated by following formula,

$$f_4 \Rightarrow MI(w_i, w_j) = \log 2 \frac{P(w_i, w_j)}{P(w_i) \times P(w_j)}$$

(v) POS Tagger Feature: Part of speech tagging is the process of categorizing the words of text on the basis of part of speech category such as noun, verbs, adverb, adjectives, they belong to. Algorithms such as hidden Markova models, using dynamic programming are used to perform this task. The POS Tags on each document is feature five of the given documents.

6) Association of deep learning with fuzzy logic: The sentence matrix S= $(s_1, s_2,.....s_n)$ which is the feature vector set having element as s_i which is set contains the all the five features extracted for the sentence s_i. Here this set of feature vectors S will be given as input to deep architecture of RBM as visible layer. Some random values is selected as bias 3_ where i = 1, 2 since a RBM can have at least two hidden layer. The whole process can be given by following equation: S= $(s_1,s_2,....,s_n)$.where s_i= $(f_1,f_2,....f_n)$, i<=n where n is the number of sentences in the document.

7) Feature Matrix: Here sentence matrix where S=$(s_1,s_2,....s_n)$ where s_i= $(f_1,f_2,....f_n)$, i<=n is the feature vector. The five features are the main attributes of the proposed text summarization algorithm. The whole documents under consideration are subjected for the feature extraction and a set of features are extracted accordingly. Now based on the collected features a feature matrix is formed by mapping the features values. The feature matrix is constructed according to the sentences extracted from the multiple documents. In addition to the five features, an additional attribute also associated with the feature matrix. The addition feature associated with the feature matrix is the class labels for each sentence. The Figure below represents the feature matrix of the set of documents under consideration.

3.1 Characteristics of Proposed System

The proposed system is User Friendly. It easily generates summary. There is no problem with the size of the document. It boosts the summary generation time.

4. SYSTEM ARCHITECTURE

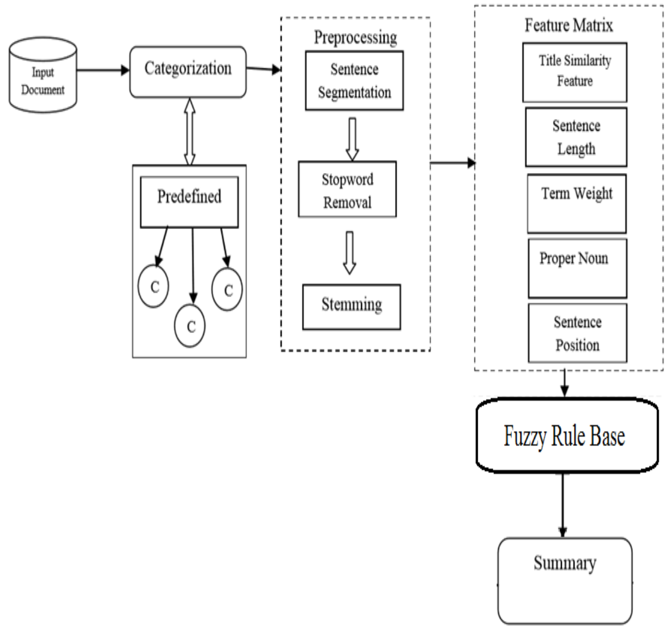


Fig -1: Architecture

This method of summarization take in to account all the characteristics of any text such as length of the sentence, similarity of sentences with the title, similarity of the sentences with the key words, etc. and then it is given as an input to the fuzzy classifier system. After the fuzzy classifier system, all the rules that are needed for summary generation are entered, in the knowledge base of the system. After that a value from zero to two is obtained for each sentence. The obtained scores of each sentence in the output determine the degree of the importance of the sentence during the final summary generation. The input membership function for each feature is divided into three membership functions which are composed of insignificant values. The important sentences are then extracted using IF-THEN rules according to the feature criteria.

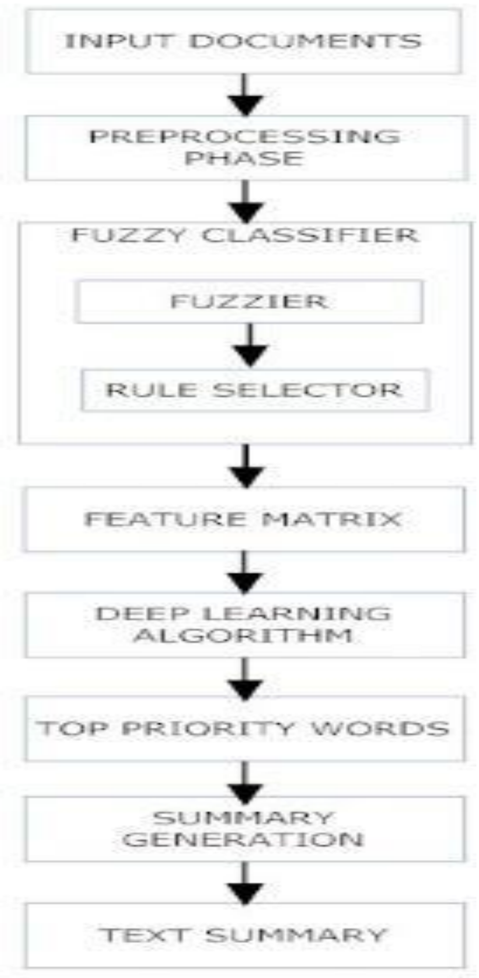


Fig -2: Overall block diagram of text summarization

5. SOFTWARE ENVIRONMENT

5.1 Front End

Windows XP, Windows 7, 8

Visual Studio 2010

Windows Operating System

5.2 Back End

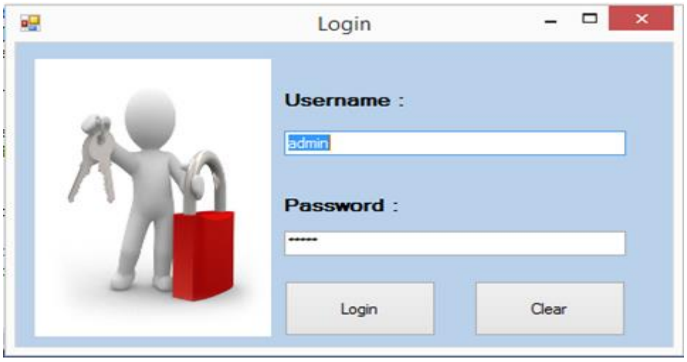
Windows XP, Windows 7,8

Visual Studio 2010

MS SQL Server 2008

Windows Operating System

6. RESULT



Username : admin

Password : *****

Fig -3: Login

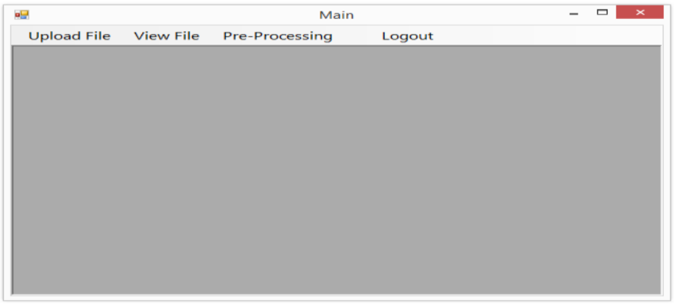
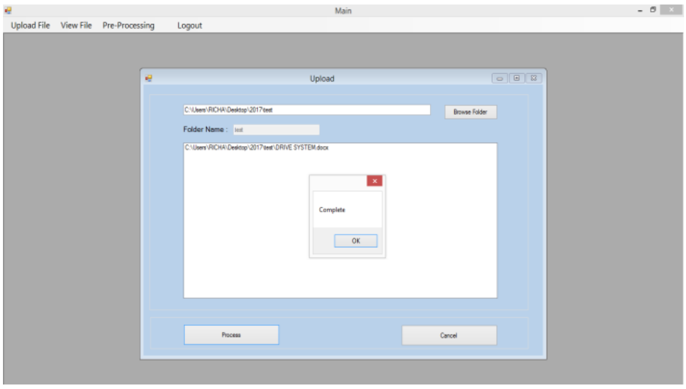


Fig -4: Main GUI Screen



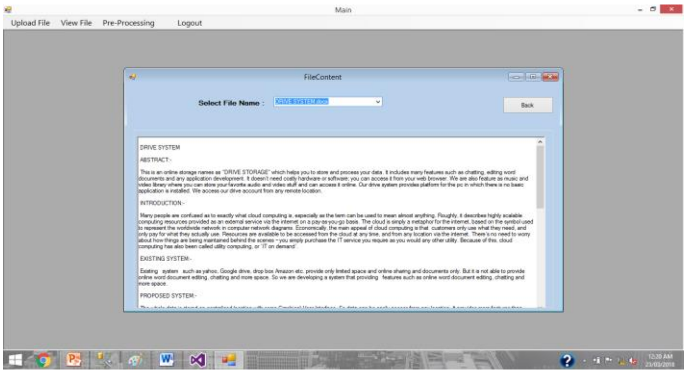
C:\Users\RICHA\Desktop\2017\test

Folder Name : test

C:\Users\RICHA\Desktop\2017\test\DRIVE SYSTEM.docx

Complete

Fig -5: Uploading a file



Select File Name : DRIVE SYSTEM.docx

DRIVE SYSTEM

ABSTRACT:-

This is an online storage names as "DRIVE STORAGE" which helps you to store and process your data. It includes many features such as chatting, editing word documents and any application development. It doesn't need costly hardware or software; you can access it from your web browser. We are also feature as music and video library where you can store your favorite audio and video stuff and can access it online. Our drive system provides platform for the pc in which there is no basic application is installed. We access our drive account from any remote location.

INTRODUCTION:-

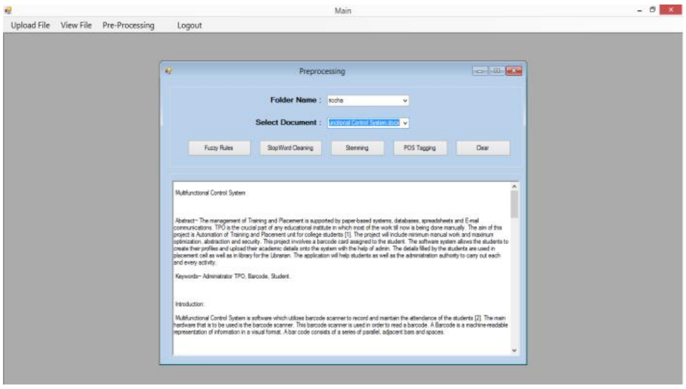
Many people are confused as to exactly what cloud computing is, especially as the term can be used to mean almost anything. Roughly, it describes highly scalable computing resources provided as an external service via the internet on a pay-as-you-go basis. The cloud is simply a metaphor for the internet, based on the symbol used to represent the worldwide network in computer network diagrams. Economically, the main appeal of cloud computing is that customers only use what they need, and only pay for what they actually use. Resources are available to be accessed from the cloud at any time, and from any location via the internet. There's no need to worry about how things are being maintained behind the scenes – you simply purchase the IT service you require as you would any other utility. Because of this, cloud computing has also been called utility computing, or 'IT on demand'.

EXISTING SYSTEM:-

Existing system such as yahoo, Google drive, drop box Amazon etc. provide only limited space and online sharing and documents only. But it is not able to provide online word document editing, chatting and more space. So we are developing a system that providing features such as online word document editing, chatting and more space.

PROPOSED SYSTEM:-

Fig -6: View an uploaded file



Folder Name : sccha

Select Document : nctional Control System.docx

Multifunctional Control System

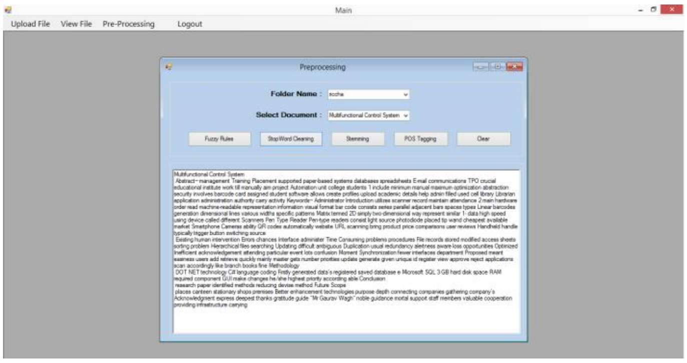
Abstract– The management of Training and Placement is supported by paper-based systems, databases, spreadsheets and E-mail communications. TPO is the crucial part of any educational institute in which most of the work till now is being done manually. The aim of this project is Automation of Training and Placement unit for college students [1]. The project will include minimum manual work and maximum optimization, abstraction and security. This project involves a barcode card assigned to the student. The software system allows the students to create their profiles and upload their academic details onto the system with the help of admin. The details filled by the students are used in placement cell as well as in library for the Librarian. The application will help students as well as the administration authority to carry out each and every activity.

Keywords– Administrator TPO, Barcode, Student.

Introduction:

Multifunctional Control System is software which utilizes barcode scanner to record and maintain the attendance of the students [2]. The main hardware that is to be used is the barcode scanner. This barcode scanner is used in order to read a barcode. A Barcode is a machine-readable representation of information in a visual format. A bar code consists of a series of parallel, adjacent bars and spaces.

Fig -7: Pre Processing

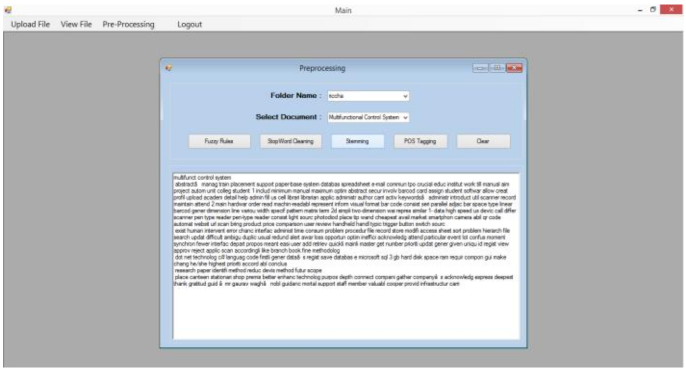


Folder Name : sccha

Select Document : Multifunctional Control System

Multifunctional Control System
Abstract– management Training Placement supported paper-based systems databases spreadsheets E-mail communications TPO crucial educational institute work till manually aim project Automation unit college students 1 include minimum manual maximum optimization abstraction security involves barcode card assigned student software allows create profiles upload academic details help admin filled used cell library Librarian application administration authority carry activity Keywords– Administrator Introduction utilizes scanner record maintain attendance 2 main hardware order read machine-readable representation information visual format bar code consists series parallel adjacent bars spaces types Linear barcodes generation dimensional lines various widths specific patterns Matrix termed 2D simply two-dimensional way represent similar 1- data high speed using device called different Scanners Pen Type Reader Pen-type readers consist light source photodiode placed tip wand cheapest available market Smartphone Cameras ability QR codes automatically website URL scanning bring product price comparisons user reviews Handheld handle typically trigger button switching source
Existing human intervention Errors chances Interface administer Time Consuming problems procedures File records stored modified access sheets sorting problem Hierarchical files searching Updating difficult ambiguous Duplication usual redundancy alertness aware loss opportunities Optimized Inefficient acknowledgement attending particular event lots confusion Moment Synchronization fewer interfaces department Proposed meant easiness users add retrieve quickly mainly master gets number priorities update generate given unique id register view approve reject applications scan accordingly like branch books fine Methodology
DOT NET technology C# language coding Firstly generated data's registered saved database e Microsoft SQL 3 GB hard disk space RAM required component GUI make changes he/she highest priority according able Conclusion
research paper identified methods reducing devise method Future Scope
places canteen stationary shops premises Better enhancement technologies purpose depth connecting companies gathering company's Acknowledgment express deepest thanks gratitude guide "Mr Gaurav Wagh" noble guidance mortal support staff members valuable cooperation providing infrastructure carrying

Fig -8: Stop Word Clearing

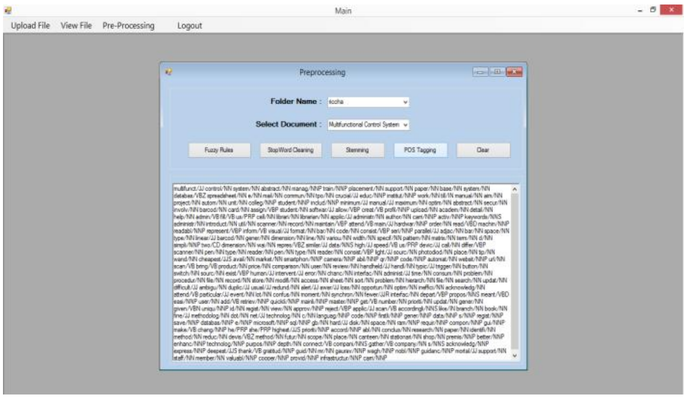


Folder Name : sccha

Select Document : Multifunctional Control System

multifunct control system
abstractâ manag train placement support paper-base system databas spreadsheet e-mail commun tpo crucial educ institut work till manual aim project autom unit colleg student 1 includ minimum manual maximum optim abstract secur involv barcod card assign student softwar allow creat profil upload academ detail help admin fill us cell librari librarian applic administr author carri activ keywordsâ administr introduct util scanner record maintain attend 2 main hardwar order read machin-readabl represent inform visual format bar code consist seri parallel adjac bar space type linear barcod gener dimension line variou width specif pattern matrix term 2d simpli two-dimension wai repres similar 1- data high speed us devic call differ scanner pen type reader pen-type reader consist light sourc photodiod place tip wand cheapest avail market smartphon camera abl qr code automat websit url scan bring product price comparison user review handheld handl typic trigger button switch sourc
exist human intervent error chanc interfac administ time consum problem procedur file record store modifi access sheet sort problem hierarch file search updat difficult ambigu duplic usual redund alert awar loss opportun optim ineffici acknowledg attend particular event lot confus moment synchron fewer interfac depart propos meant easi user add retriev quickli mainli master get number prioriti updat gener given uniqu id regist view approv reject applic scan accordingli like branch book fine methodolog
dot net technolog c# languag code firstli gener dataâ s regist save databas e microsoft sql 3 gb hard disk space ram requir compon gui make chang he/she highest prioriti accord abl conclus
research paper identifi method reduc devis method futur scope
place canteen stationari shop premis better enhanc technolog purpos depth connect compani gather companyâ s acknowledg express deepest thank gratitud guid â mr gaurav waghâ nobl guidanc mortal support staff member valuabl cooper provid infrastructur carri

Fig -9: Stemming

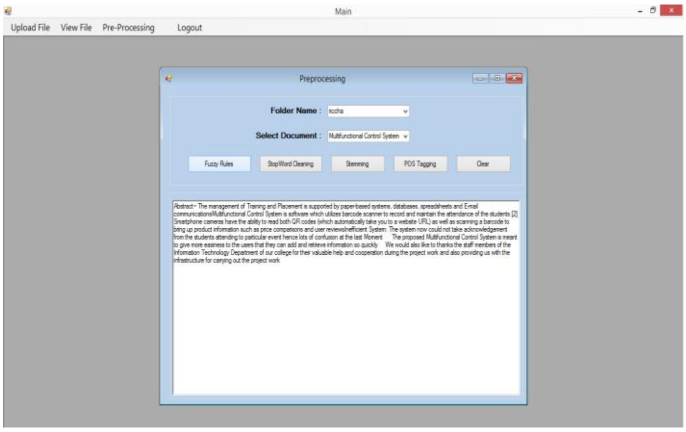


Folder Name : sccha

Select Document : Multifunctional Control System

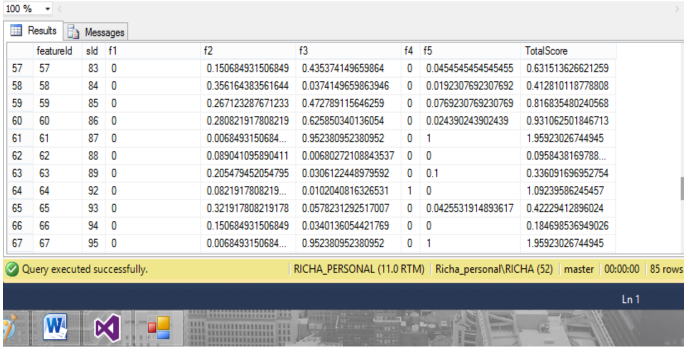
multifunct/JJ control/NN system/NN abstract/NN manag/NNP train/NNP placement/NN support/NN paper/NN base/NN system/NN databas/VBZ spreadsheet/NN e/NN mail/NN commun/NN tpo/NN crucial/JJ educ/NNP institut/NNP work/NN till/IN manual/NN aim/NN project/NN autom/NN unit/NN colleg/NNP student/NNP includ/NNP minimum/JJ manual/JJ maximum/NN optim/NN abstract/NN secur/NN involv/NN barcod/NN card/NN assign/VBP student/NN softwar/JJ allow/VBP creat/VB profil/NNP upload/NN academ/NN detail/NN help/NN admin/VB fil/VB us/PRP cell/NN librari/NN librarian/NN applic/JJ administr/NN author/NN carri/NNP activ/NNP keywords/NNS administr/NN introduct/NN util/NN scanner/NN record/NN maintain/VBP attend/VB main/JJ hardwar/NNP order/NN read/VBD machin/NNP readabl/NNP represent/VBP inform/VB visual/JJ format/NN bar/NN code/NN consist/VBP seri/NNP parallel/JJ adjac/NN bar/NN space/NN type/NN linear/JJ barcod/NN gener/NN dimension/NN line/NN variou/NN width/NN specif/NN pattern/NN matrix/NN term/NN d/NN simpli/NNP two/CD dimension/NN wai/NN repres/VBZ similar/JJ data/NNS high/JJ speed/VB us/PRP devic/JJ call/NN differ/VBP scanner/NN pen/NN type/NN reader/NN pen/NN type/NN reader/NN consist/VBP light/JJ sourc/NN photodiod/NN place/NN tip/NN wand/NN cheapest/JJS avail/NN market/NN smartphon/NNP camera/NNP abl/NNP qr/NNP code/NNP automat/NN websit/NNP url/NN scan/VB bring/VB product/NN price/NN comparison/NN user/NN review/NN handheld/JJ handl/NN typic/JJ trigger/NN button/NN switch/NN sourc/NN exist/VBP human/JJ intervent/JJ error/NN chanc/NN interfac/NN administ/JJ time/NN consum/NN problem/NN procedur/NN file/NN record/NN store/NN modifi/NN access/NN sheet/NN sort/NN problem/NN hierarch/NN file/NN search/NN updat/NN difficult/JJ ambigu/NN duplic/JJ usual/JJ redund/NN alert/JJ awar/JJ loss/NN opportun/NN optim/NN ineffici/NN acknowledg/NN attend/VB particular/JJ event/NN lot/NN confus/NN moment/NN synchron/NN fewer/JJR interfac/NN depart/VBP propos/NNS meant/VBD easi/NNP user/NN add/VB retriev/NNP quickli/NNP mainli/NNP master/NNP get/VB number/NN prioriti/NN updat/NN gener/NN given/VBN uniqu/NNP id/NN regist/NN view/NN approv/NNP reject/VBP applic/JJ scan/VB accordingli/NNS like/IN branch/NN book/NN fine/JJ methodolog/NN dot/NN net/JJ technolog/NN c/NN languag/NNP code/NNP firstli/NNP gener/NNP data/NNP s/NNP regist/NNP save/NNP databas/NNP e/NNP microsoft/NNP sql/NNP gb/NN hard/JJ disk/NN space/NN ram/NNP requir/NNP compon/NNP gui/NNP make/VB chang/NNP he/PRP she/PRP highest/JJS prioriti/NNP accord/NNP abl/NN conclus/NN research/NN paper/NN identifi/NN method/NN reduc/NN devis/VBZ method/NN futur/NN scope/NN place/NN canteen/NN stationari/NN shop/NN premis/NNP better/NNP enhanc/NNP technolog/NNP purpos/NNP depth/NN connect/VB compani/NNS gather/VB company/NN s/NNS acknowledg/NNP express/NNP deepest/JJS thank/VB gratitud/NNP guid/NN mr/NN gaurav/NNP wagh/NNP nobl/NNP guidanc/NNP mortal/JJ support/NN staff/NN member/NN valuabl/NNP cooper/NNP provid/NNP infrastructur/NNP carri/NNP

Fig -10: POS Tagger



Abstract-- The management of Training and Placement is supported by paper-based systems, databases, spreadsheets and E-mail communicationsMultifunctional Control System is software which utilizes barcode scanner to record and maintain the attendance of the students [2] Smartphone cameras have the ability to read both QR codes (which automatically take you to a website URL) as well as scanning a barcode to bring up product information such as price comparisons and user reviewsInefficient System: The system now could not take acknowledgement from the students attending to particular event hence lots of confusion at the last Moment The proposed Multifunctional Control System is meant to give more easiness to the users that they can add and retrieve information so quickly We would also like to thanks the staff members of the Information Technology Department of our college for their valuable help and cooperation during the project work and also providing us with the infrastructure for carrying out the project work

Fig -11: Summary Generated



	featureId	sId	f1	f2	f3	f4	f5	TotalScore
57	57	83	0	0.150684931506849	0.435374149659864	0	0.0454545454545455	0.631513626621259
58	58	84	0	0.356164383561644	0.0374149659863946	0	0.0192307692307692	0.412810118778808
59	59	85	0	0.267123287671233	0.472789115646259	0	0.0769230769230769	0.816835480240568
60	60	86	0	0.280821917808219	0.625850340136054	0	0.024390243902439	0.931062501846713
61	61	87	0	0.0068493150684...	0.952380952380952	0	1	1.95923026744945
62	62	88	0	0.089041095890411	0.00680272108843537	0	0	0.0958438169788...
63	63	89	0	0.205479452054795	0.0306122448979592	0	0.1	0.336091696952754
64	64	92	0	0.0821917808219...	0.0102040816326531	1	0	1.09239586245457
65	65	93	0	0.321917808219178	0.0578231292517007	0	0.0425531914893617	0.42229412896024
66	66	94	0	0.150684931506849	0.0340136054421769	0	0	0.184698536949026
67	67	95	0	0.0068493150684...	0.952380952380952	0	1	1.95923026744945

Query executed successfully. | RICHA_PERSONAL (11.0 RTM) | Richa_personal\RICHA (52) | master | 00:00:00 | 85 rows

Ln 1

Fig -12: Feature Matrix Calculation

7. CONCLUSIONS

In this paper we have implemented the auto text summarization which involves extraction of features of all sentences using fuzzy logic method. We have extracted five different features of all the sentences. Later using the sentence score we have generated the feature matrix. Then based on the scores of individual sentences and final feature matrix value we have generated the summary. We have tested the system with multiple numbers of documents. The results obtained show that the use of fuzzy logic in summary generation improves the quality of summary generated compared to other method of summary generation. We have applied our method of summary generation for single document which could be extended for multi-document summarization. In the input we used multiple documents; these documents can be further categorized like sports, politics, weather, music etc. hence the system will have the categorization of the text document feature which will help to know the text belongs to which category.

REFERENCES

[1] Allan Borra, Almira Mae Diola, Joan Tiffany T. Ong Lopez, Phoebus Ferdiel Torralba,Sherwin So, "Using Rhetorical Structure Theory in Automatic Text Summarization for Marcu-Authored Documents", In titaniaaddueduph, 2010.

[2] B.Arman kiani, M.R. Akbarzadeh, "Autmatic Text Summarization using: Hybrid Fuzzy GA-GP" IEEE International Conference on Fuzzy Systems,2006.

[3] C.Gordon and R.Debus, "Developing deep learning approaches and personal teaching efficacy within a preservice teacher education context," Web Intelligence and Intelligent Agent Technology, British Journal of Educational Psychology,Vol. 72, No. 4, 2002, PP. 483-511.

[4] F. kyoomarsi, H. khosravi, E. eslami and M. davoudi, "Extraction-based text summarization using fuzzy analysis," Iranian Journal of Fuzzy Systems, Vol. 7, No. 3, 2010, PP. 15-32.