

A Statistical Approach for Document Summarization

Vishal Patil, Mahalakshmy Krishnamoorthy, Parag Oke, Prof. M. Kiruthika
Department of Computer Engineering
Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai, Maharashtra, India

Abstract

This paper discusses about our work on statistical approach for document summarization. It is an application which takes a user query from the user which may be a keyword like cat, python and then gives summary for that keyword from the two or more web documents retrieved for that keyword. Summarization is composed of various phases like searching for top results for user query, gathering documents, extraction of data, analysis, generation of summary, presentation of summary to user. Performance of summarization will be based on some parameters like retention ratio, length of summary. In this paper section 1 discusses some introductory concepts, section 2 the need for document summarization, section 3 describes the existing systems, section 4 gives the design of our summarizer, section 5 discusses how we have implemented the summarizer based on the design, section 6 gives the various applications of the summarizer and section 7 gives the conclusion and how the summarizer can be improved further

Keywords— *Summarization, NLP, Python, regular expression, sentence generation.*

I. INTRODUCTION

Summarization means giving a brief statement of the main points of some information. It is always better to have summary of something rather than a long description about something. The best example of summary is the trailers of movies. Trailer shows the theme of movie, actors, villains, best dialogues etc. about movie. By watching the trailer itself we got enough idea about the movie.

Natural language processing (NLP) is a very efficient tool to deal with the documents. To make use of NLP concepts, python provides Natural Language Tool Kit (NLTK) library which consist of various

tools, methods, datasets available freely with their documentation for various platforms.[1]

A .Classifications of text summarization

Text summarization systems can be categorized as:

1. Extractive
2. Abstractive

In extractive summarization approaches [2], the goal is identifying most important concepts in the input document, and giving related sentences found in the document as an output. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. The summary created using these sentences may not be coherent, but gives idea about the content of the input document.

In abstractive summarization approaches [2], first the system understands the texts and then it creates summaries with its own words. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. Another way of categorization of the text summarization systems is based on the approaches used in the summarization algorithms. There are different algorithms which are based on supervised or unsupervised techniques. Supervised techniques use data sets that are labelled by human annotators. Unsupervised approaches do not use annotated data, but they use linguistic and statistical information that are obtained from the document itself.

Also, text can be summarized with the help of single or multiple documents. Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic.

B. Some of approaches for text summarization

1. Statistical Approaches

The most well-known summarization approaches that use statistics are based on concept relevance and Bayesian classifier. This approach uses word frequency, uppercase words, sentence length, keywords, position in complete text, and phrase structure.

2. Text Connectivity Based Approaches

It deals with problems of referencing to the already mentioned parts of a document. Methods that use this approach are lexical chains and Rhetorical Structure Theory (RST). Lexical chain consist of extracting semantic relations of words (synonym, antonym) using dictionaries and WordNet. Using semantic relations lexical chains are constructed and used for extracting important sentences in a document. RST organizes text units into a tree like structure. Then this structure is used for summarization purposes.

3. Graph Based Approaches

The nodes in graph based summarization approaches represent the sentences, and the edges represent the similarity among the sentences. The similarity values are calculated using the overlapping words or phrases. The sentences with highest similarity to the other sentences are chosen as a part of the resulting summary. TextRank and Cluster LexRank are two methods that use graph based approach for document summarization.

4. Algebraic approach

Algebraic methods such as Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), and Semi-discrete Matrix Decomposition (SDD) are used for document summarization. Among these algorithms most well-known one is LSA, which is based on singular value decomposition (SVD). In this algorithm similarity among sentences and similarity among words are extracted.

5. Non-Extractive Summarization Methods

Abstractive summarization methods try to fully understand the given documents, even non-explicitly mentioned topics, and generate new sentences for the summary. This approach is very similar to the way of human summarization. There are approaches that create summaries in a non-extractive manner, using information extraction, ontological information, information fusion and compression.

C. Performance/Evaluation Measures

It determines the quality of summary with respect length, coherence, structure, content. Two parameters are important in text summarization, the Compression Ratio, i.e. how much shorter the summary is than the original, and the Retention Ratio, i.e. how much of the central information is retained. A broad division into evaluation techniques would be Intrinsic and Extrinsic evaluation.

1. Intrinsic Evaluation

Intrinsic evaluation criteria are those relating to a system's objective. This is often done by comparison to some gold standard, which can be made by a reference summarization system or, more often than not, is man-made using informants. Intrinsic evaluation mainly focuses on the coherence and informativeness of summaries.

2. Extrinsic Evaluation

Extrinsic methods evaluate how summaries are good enough to accomplish the purpose of some other specific task, e.g. filtering in information retrieval or report generation. Extrinsic evaluation measures the efficiency and acceptability of the generated summaries in some task.

3. ROUGE (Recall-Oriented Understudy for Gist Evaluation)

This is another measure apart from above two. N-gram co-occurrence measure is another content based evaluation method. Given multiple human judged, ideal summaries, maximum number of n-gram co-occurring between extracted and ideal summaries is calculated. The value is then divided by the total number of n-grams in ideal summaries. ROUGE-n score is a recall based score. ROUGE-n calculation is given, where RSS is the reference summary set and C is the candidate summary and n is the length of the n-gram.

II. NEED

Internet is huge source of information. Any information from internet can be got at one click. Consider a case where a user wants to search some information about bullet. User might get search results including bullet as bike, bullet as weapon, bullet as train etc. Suppose user want to search for bike, and then user selects web sites for results bullet bike, now different web sites shows different information with them, common, uncommon. User may need to go through all these information; it will be a very tedious job some times. Due to the great amount of

information available on Web, it is not possible to analyse every text on a topic.

Hence, there is a need of producing coherent summaries to this information have become essential. User will be satisfied if all these information would be available at one place and in a pleasant format that would help user to make it more understandable. In text summarization, the intention is to express the contents of document in condensed form suitable to user.

III. EXISTING SYSTEMS

Document summarizers are implemented along with the help of search engines or use their own methods to obtain the documents. The existing systems include TextRank, LexRank MEAD, WebInEssence[3]. All the systems have their main goal as to obtain relevant data and summarize them as per the users query. They all differ in the main aspects of the features provided and the architecture used.

IV. PROPOSED SYSTEM

We would like to propose a system in which user would provide a topic/question of search and then a summary would be presented in different user friendly formats by summarizing web page content of top N results found from different search engines. The main idea behind our system is to summarize clusters of related Web pages to provide more contextual and summary information to help users get results more quickly. For example, a historical topic search would result in table consisting of dates at which some event happened and the names of people highlighted. Further, the summary result can be customized based on user's criteria like length of summary, retention ratio, keyword match etc. The summarized contents should be clear to understand, should represent text within sections into a meaningful paragraphs.

V. DESIGN

Our algorithm works on basic concepts of summarization like TF (Term frequency), IDF (Inverse document frequency), semantic similarity etc. The steps of our algorithm are as follows:

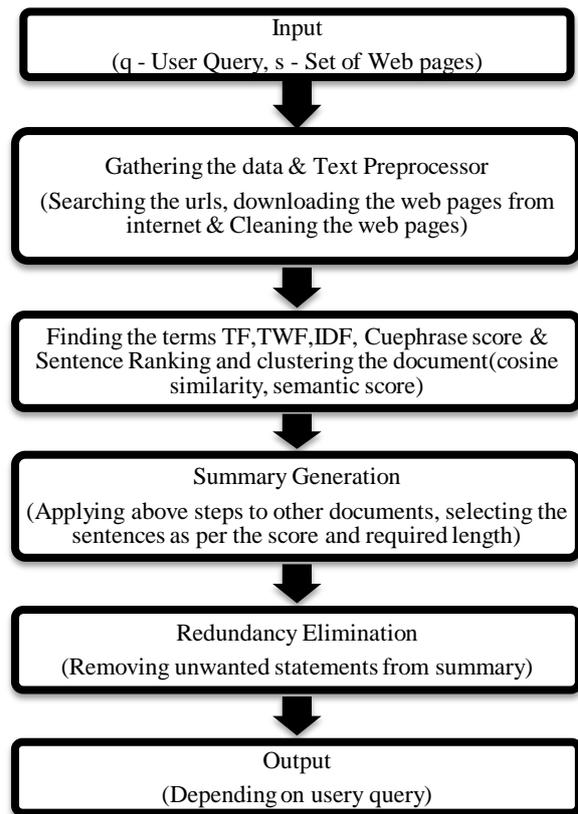


Fig. 1 Flowchart for Multi-Document Summarizer

The Fig. 1 shows the flow chart of the algorithm which is used by our system. Every phase performs step by step functionalities which represents the modules of this system.

1. Input

q=user query consists of keywords, phrases
s=set of web pages for q from different SE(top 10 from every SE (search engines))

2. Gathering the data&Text Pre-processor

With the help of keyword provided by user, the URLs for that keyword are searched on Internet and some web pages are downloaded.

To search the web pages Google API is used which returns the list of URLs for the given query.

The web pages contains the html tags which are unnecessary for the summarization hence to extract the contents from the web page, we need to remove all the tags[1]. For this we can use regular expressions and NLTK

3. Calculations

Generating the summary requires various terms to be calculated like TF, TWF, IDF, W, cue phrases score.[2] [4]



TF, TWF, IDF terms represents occurrence of word in the sentence, document. W is the multiplication of TF and IDF of all the words.

This step has more importance as it deals with ranking the sentences that appears in the summary.

The inner product of sentences is calculated by considering the sentences as group to check the relevancy between the sentences. This is required to calculate the cosine of two sentences. [2]

$$\text{Cosine} [(S_i, S_j)] = \frac{\text{inner_product}}{\text{sqrt}(\text{norm1} * \text{norm2})}$$

where, S_i and S_j are two sentences, norm1 and norm 2 is the square of W for all the words in the sentences.

The semantic score is calculated from the W.

After finding the scores, clustering is performed on document based on the scores. Based on sentences in cluster and their score, overall cluster score is calculated. [5]

4. Summary generation

All above steps are repeated for the other documents and based on the length provided by the user, the sentences having highest score are chosen for the summary.

5.Redundancy elimination

Generating summary from various documents may add some irrelevant, redundant sentences that may get high score. So identifying such sentences and removing them is essential.

6. Output

Finally the fine summary is presented to the user. It contains some suggestions with respect to the given query, highlighting the query in the summary.

VI. IMPLEMENTATION

The above modules discussed in design were implemented.

Fetching documents from web

To get the document from the web we need the list of URLs from where we can download the documents. The keywords given by the end user can be used to get the URLs from the web by using the API provided by the Google. <http://ajax.googleapis.com/ajax/services/search/web?v=1.0&q=query&start=0&rsz=large> here the

query is the keyword provided by the user

Example:

If the query is 'Microsoft windows' then result is {"responseData":"http://windows.microsoft.com/","url":"http://windows.microsoft.com/","visibleUrl":"windows.microsoft.com"}. This result is to be parsed to get the URLs.

• Single Document Summarizer:

Finding the TF, IDF, W terms

TF-IDF is term frequency-inverse document frequency, is a numerical statistic which shows how important a word is to a document in a collection of texts. Term frequency is the number of times a term occurs in a document and Inverse document frequency is a measure that diminishes the weight of terms that occur very frequently in the collection and increases the weight of terms that occur rarely. The product of TF and IDF is stored in W vector which is used in finding semantic score. The TF and IDF are considered as dictionary data structure of python; this gives efficient way to find the frequency. If word exists in TF dictionary then increment its frequency else add the word in the dictionary, same can be done for IDF.

Cue scores are calculated based on cue phrases found in the sentence, e.g. 'is a', 'defined as', 'Thus', 'to conclude' etc. Also the score is normalized by dividing it with sentence length.

Finding similarity and semantic score:

The similarity between two sentences can be computed by considering two statements as two nodes of graph and the similarity between them is the link. Finding the intersection and inner product of two sentences gives the cosine similarity. The semantic score can be calculated by using the W vector and the formula as follows:

$$\begin{aligned} \text{Score}(S_i, S_j) = \exp(2 & \\ & * \cos(\text{TF} \\ & - \text{IDF vectors of both sentences}) \\ & + \left(\frac{\text{no. of pronouns in } S_j}{\text{length of } S_j} - 0.2 \right) \\ & * \log \left(\frac{1.0}{\text{sentence_sep} + 1.0} \right) \\ & + \text{sigmoid}(-N + 15) \\ & * \log \left(\frac{1.0}{\text{sentence_sep} + 1.0} \right) \end{aligned}$$

Where, S_i and S_j are two consecutive sentences.

Clustering the document

Similar sentences are grouped together to form a cluster. For every cluster, query overlap is calculated to find cluster that is closest to query. Normalization for clusters is carried out by dividing the cluster score by total number of sentences present. Maximum scoring clusters are chosen depending on the size of summary needed. Using the clusters the cluster score and sentences score is calculated by the formula:

$$cluster_score = \frac{2 * \left(x - \left(\frac{a}{2}\right)\right) * 2}{\left(\frac{a}{2}\right) * 2 + 1}$$

Where x is position of sentence in the cluster and a is size of cluster. Sorting the clusters by their score gives the summary.

- **Multi Document Summarizer**

In this module, summary is generated from two documents. These two documents are summarized individually by single document summarization. These summaries are then stored in a document separated by a separator. Thus parsing these two summaries becomes easy rather than reading two files at a time. Initial steps for calculating the TF, IDF, W vector are same. Calculating similarity and semantic

score have some changes. The two sentences for which similarity is to be calculated are now not the consecutive sentences, but they are sentences from two documents. So, here we are considering sentence as statement from starting and sentence after separator.

VII. APPLICATIONS

Other than presenting information in summarized form there can be various applications in which this system can be used.

The applications of a multi document summarizer are:

- It can be used as a news portal and can help to present articles from different sources.
- Corporate emails or emails in general can be organized by subjects with relevant and most important information.
- It can help to obtain precise information which is represented as charts or graphs along with related text.
- It can be used to generate medical reports for patients.
- Aggregating social media data.

VIII. RESULTS

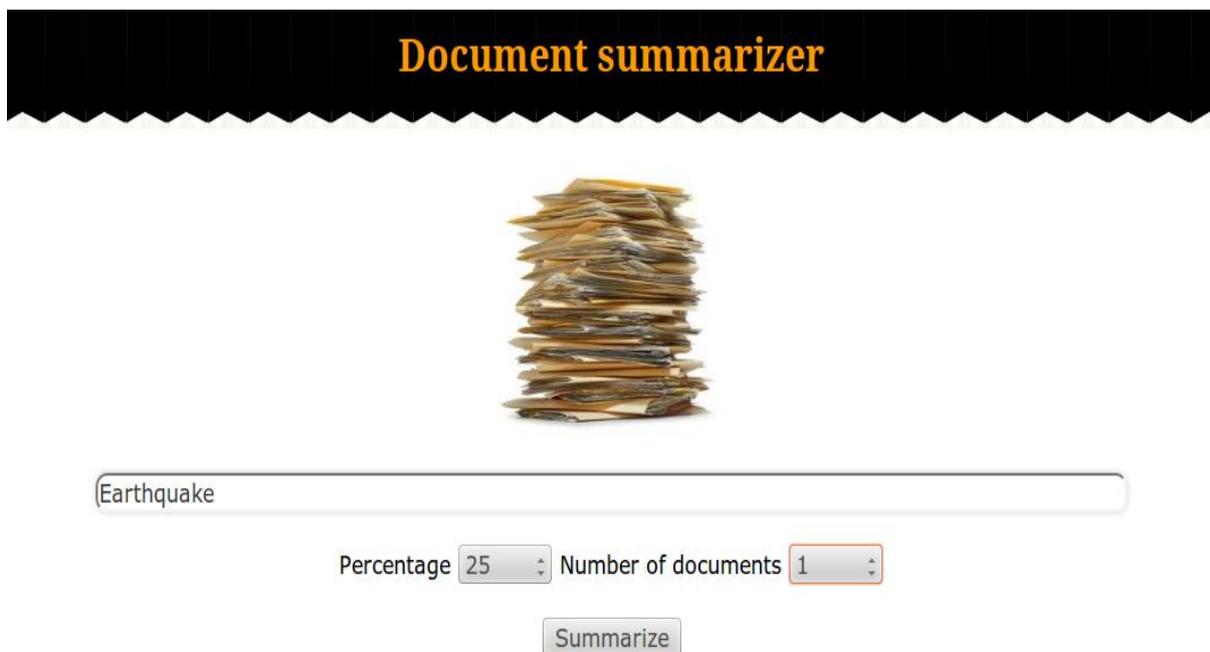


Fig 1 Single document summarization - Input



Summary on earthquake

Summary	Document 1	Statistics
---------	------------	------------

Source = <http://en.wikipedia.org/wiki/Earthquake>

Earthquake is the result of a sudden release of energy in the Earth's crust that creates seismic waves. The seismicity, seismism or seismic activity of an area refers to the frequency, type and size of earthquakes experienced over a period of time. Earthquakes are measured using observations from seismometers. The moment magnitude is the most common scale on which earthquakes larger than approximately 5 are reported for the entire globe. The more numerous earthquakes smaller than magnitude 5 reported by national seismological observatories are measured mostly on the local magnitude scale, also referred to as the Richter scale. These two scales are numerically similar over their range of validity. Magnitude 3 or lower earthquakes are mostly almost imperceptible or weak and magnitude 7 and over potentially cause serious damage over larger areas, depending on their depth. The largest earthquakes in historic times have been of magnitude slightly over 9, although there is no limit to the possible magnitude. The most recent large earthquake of magnitude 9.0 or larger was a 9.0 magnitude earthquake in Japan in 2011 (as of October 2012), and it was the largest Japanese earthquake since records began. Intensity of shaking is measured on the modified Mercalli scale. The shallower an earthquake, the more damage to structures it causes, all else being equal. [1] At the Earth's surface, earthquakes manifest themselves by shaking and sometimes displacement of the ground. When the epicenter of a large earthquake is located offshore, the seabed may be displaced sufficiently to cause a tsunami. Earthquakes can also trigger landslides, and occasionally volcanic activity. In its most general sense, the word earthquake is used to describe any seismic event whether natural or caused by humans that generates seismic waves. Earthquakes are caused mostly by rupture of geological faults, but also by other events such as volcanic activity, landslides, mine blasts, and nuclear tests. An earthquake's point of initial rupture is called its focus or hypocenter. The epicenter is the point at ground level directly above the hypocenter. Tectonic earthquakes occur anywhere in the earth where there is sufficient stored elastic strain energy to drive fracture propagation along a fault plane. The sides of a fault move past each other smoothly and aseismically only if there are no irregularities or asperities along the fault surface that increase the

Fig 2 Original Document

Summary on earthquake

Summary	Document 1	Statistics
---------	------------	------------

Earthquake is the result of a sudden release of energy in the Earth's crust that creates seismic waves. Earthquakes are measured using observations from seismometers. Earthquakes can also trigger landslides, and occasionally volcanic activity. Earthquakes are caused mostly by rupture of geological faults, but also by other events such as volcanic activity, landslides, mine blasts, and nuclear tests.

Normal and reverse faulting are examples of dip-slip, where the displacement along the fault is in the direction of dip and movement on them involves a vertical component. Normal faults occur mainly in areas where the crust is being extended such as a divergent boundary.

Reverse faults occur in areas where the crust is being shortened such as at a convergent boundary. Strike-slip faults are steep structures where the two sides of the fault slip horizontally past each other; transform boundaries are a particular type of strike-slip fault.

Many earthquakes are caused by movement on faults that have components of both dip-slip and strike-slip; this is known as oblique slip. Reverse faults, particularly those along convergent plate boundaries are associated with the most powerful earthquakes, including almost all of those of magnitude 8 or more.

Strike-slip faults, particularly continental transforms can produce major earthquakes up to about magnitude 8. Earthquakes associated with normal faults are generally less than magnitude 7.

Earthquakes are not only categorized by their magnitude but also by the place where they occur. The world is divided into 754 Flinn-Engdahl regions (F-E regions), which are based on political and geographical boundaries as well as seismic activity.

More active zones are divided into smaller F-E regions whereas less active zones belong to larger F-E regions. 1755 copper engraving depicting Lisbon in ruins and in flames after the 1755 Lisbon earthquake, which killed an estimated 60,000 people.

Shaking and ground rupture are the main effects created by earthquakes, principally resulting in more or less severe damage to buildings and other rigid structures. The severity of the local effects depends on the complex combination of the earthquake magnitude, the distance from the epicenter, and the local geological and geomorphological conditions, which may amplify or reduce wave propagation. The ground-shaking is measured by ground acceleration.

Specific local geological, geomorphological, and geostructural features can induce high levels of shaking on the ground surface even from low-intensity earthquakes. This effect is called site or local amplification.

Fig 3 Summary

Summary on earthquake

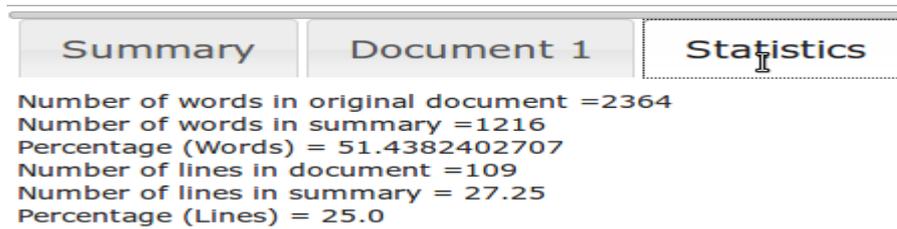


Fig 4 Statistics



Fig 5 Multi - Document Summarization: Test case 1 - Input

Summary on sachin tendulkar

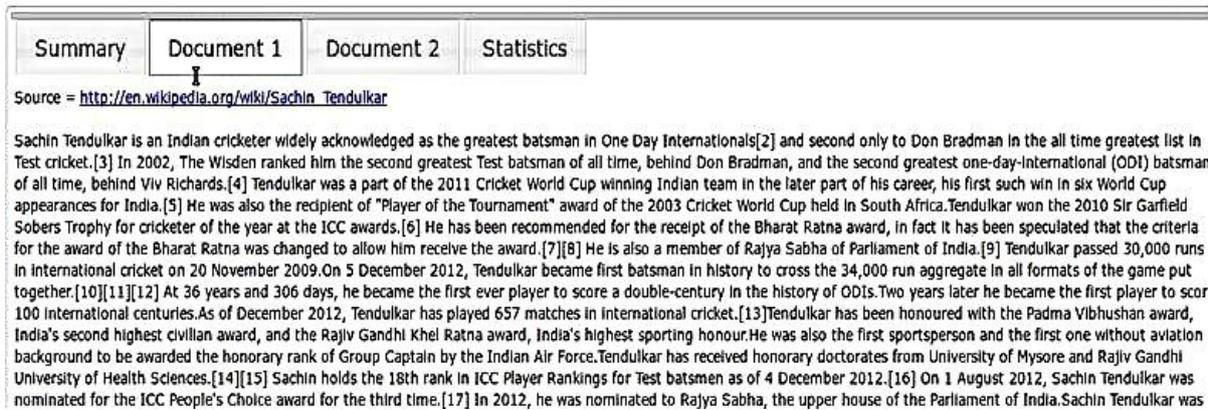


Fig 6 Original Document 1

Summary on sachin tendulkar



Fig 7 Original Document 2

Summary on sachin tendulkar

Summary	Document 1	Document 2	Statistics
---------	------------	------------	------------

Sachin Tendr is an Indian cricketer widely acknowledged as greatest batsman in One Day Internationals and second only to Don Bradman in all time greatest list in Test cric 002, The Wisden ranked him second greatest Test batsman of all time, behind Don Bradman, and second greatest one-day-international (ODI) batsman of all time, behind Richards.[4] Tendr was a part of 2011 Cricket World Cup winning Indian team in later part of his career, his first such win in six World Cup appearances for India.[5] He was recipient of "Player of Tournament" award of 2003 Cricket World Cup held in South Africa. Tendulkar won the 2010 Sir Garfield Sobers Trophy for cricketer of the year at th awards. He has been recommended for the receipt of the Bharat Ratna award, in fact it has been speculated that the criteria for the award of the Bharat Ratna was changed him receive the award.[7] [e iso a member of Rajya Sabha of Parliament of India.[9] Tendulkar scored 30,000 runs in international cricket on 20 November 2009. On 5 December 2012, Tendulkar became first batsman in history to cross the 34,000 run aggregate in all formats of the game put together.[11] At ears and 306 days, he l the first ever player to score a double-century in the history of ODIs. As of December 2012, Tendulkar has played 657 matches in international cricket. Tendulkar has been honoured with the Padma Vibhushan award, India's second highest civilian award, and the Rajiv Gandhi Khel Ratna award, India's highest sporting honour. Tendulkar has received honorary doctorates from University of Mysore and Rajiv Gandhi University of Health Sciences.[15] hin holds the 18th rank in ICC Player Ran Test batsmen as of 4 December 2012.[16] On gust 2012, hin Tendulkar nominated for the ICC People's Choice award for the third time.[17] In 2012, he nominated to Rajy the upper house of the Parliament of India.

Tendulkar's record as captain						
	Matches	Won	Lost	Drawn	Tied	No result
Test ^[80]	25	4	9	12	0	-
ODI ^[81]	73	23	43	-	2	6

Tendulkar's results in international matches ^[110]						
	Matches	Won	Lost	Drawn	Tied	No result

Fig 8 Summary

Summary on sachin tendulkar

Summary	Document 1	Document 2	Statistics
---------	------------	------------	------------

Number of words in original document = 8232
 Number of words in summary = 1501
 Percentage (Words) = 18.2337220603
 Number of lines in document = 303
 Number of lines in summary = 16
 Percentage (Lines) = 5.28052805281

Fig 9 Statistics:

Document summarizer



keyboard

Percentage Number of documents

Fig 10 Multi - Document Summarization: Test case 2 - Input

Summary on keyboard

Summary	Document 1	Document 2	Document 3	Statistics
---------	------------	------------	------------	------------

Source = http://en.wikipedia.org/wiki/Computer_keyboard

keyboard is a typewriter-style device, which uses an arrangement of buttons or keys, to act as mechanical levers or electronic switches. Following the decline of paper tape, interaction via teleprinter-style keyboards became the main input device for computers. A keyboard typically has characters engraved or printed on its keys, and the press of a key typically corresponds to a single written symbol. However, to produce some symbols requires pressing and holding several keys simultaneously. Most keyboard keys produce letters, numbers or signs (characters), other keys or simultaneous key presses can produce actions or computer commands. As an alternative input device, such as the mouse, touchscreen, pen devices, character recognition and voice recognition, the keyboard remains the most common and versatile device used for direct (human) input into computers. In normal usage, the keyboard is used to type text and numbers into a word processor, text editor, or other programs. In a modern computer, the interpretation of key presses is generally left to the software. A computer keyboard distinguishes each physical key from the others and reports all key presses to the controlling software. Keyboards are also used for computer gaming, either with regular keyboards or by using keyboards with special keys which can expedite frequently used keystroke combinations. A keyboard is also used to give commands to the operating system of a computer, such as Windows.

Fig 11 Original document 1

Summary on keyboard

Summary	Document 1	Document 2	Document 3	Statistics
---------	------------	------------	------------	------------

Source = http://en.wikipedia.org/wiki/enhanced_keyboard

keyboard is the Model M. Introduced in 1985 and manufactured by IBM, Lexmark and Unicomp, the vast majority of Model M keyboards feature a buckling spring design, many have fully swappable keycaps. The PC keyboard changed over the years, often at the launch of new IBM PC versions. Additional SysRq, i.e. additional function keys; 12 F keys in separate row along top, grouped F1-4, F5-8, and F9-12. Early models of Enhanced keyboard (notably those manufactured by Northgate) had a layout with function keys on the left side, arranged in two columns of six pairs. This layout was more efficient for touch typists but was superseded in the late 1980s by the F-keys along the top. Brazilian ABNT NBR 10346 variant 2 (alphanumeric portion) and 10347 (numeric portion). Additional Windows key (2) and Menu key to the right of the left control key, the other and the Menu key to the left of the right control key). The Windows keyboard was introduced for use with the Windows operating system. Most modern PCs, whether supplied with Windows or not, are now delivered with this layout. Brazilian ABNT NBR 10346 variant 2 (alphanumeric portion) and 10347 (numeric portion). Common additions to the standard layouts include additional power management keys, volume controls, media player controls, and miscellaneous short-cuts for e-mail client, web browser, etc. The PC keyboard with its various keys has a long history of evolution reaching back to teletypewriters. In addition to the standard keys, the TACO keyboard has accumulated several special keys over the years. Some of the additions have been inspired by the opportunity or requirements for increased productivity with general office application software, while other slightly more general keyboard additions have become the factory standards after being widely adopted.

Fig 12 Original document 2

Summary on keyboard

Summary	Document 1	Document 2	Document 3	Statistics
---------	------------	------------	------------	------------

Source = http://en.wikipedia.org/wiki/IBM_PC_keyboard

keyboard is a class of computer keyboard manufactured by IBM, Lexmark and Unicomp, starting in 1984. The many different variations of the keyboard have characteristics, with the vast majority having a buckling spring key design and many having fully swappable keycaps. Model M keyboards have been prized for their tactile and auditory feedback resulting from a keystroke. The Model M is also regarded as a timeless and durable piece of computer hardware. Many units manufactured since the mid 1980s are still in use today, while the computers and monitors of the day are obsolete. Unicomp, which no longer designs or manufactures keyboards, continues to sell Model M keyboards. Recently, the keyboards have made a comeback among writers and computer techs. [1] Unicomp has had difficulty in the past because they rarely break, and most retailers will not stock such an expensive keyboard. [1] The Model M was designed to be a more cost effective alternative to the Model F keyboards it replaced. Production for the original Model M began in 1985, and the keyboards were often bundled with new IBM computers in the 1980s. They were produced by IBM in their plants in Lexington, Greenock and Guadalajara. The most common Model M variant is the part number 1391401, which was used with PS/2. Until 1987, the keyboards featured a detachable AT cable; after that, they were bundled with a detachable PS/2 cable. Cables came in both 5- and 10-metre lengths. From about 1994 onwards, the majority of Model Ms were manufactured with non-detachable cables to cut down manufacturing costs, as well as to support the then-upcoming Windows 95 operating system. In March 1991, IBM divested a number of its hardware manufacturing operations, including keyboard manufacturing, to investment firm Clayton Dubilier, Inc. in a leveraged buyout to form Lexmark International, Inc. [2][3] The Model M keyboard continued to be produced by States and Mexico, and IBM in Scotland with IBM being Lexmark's major customer. [4] Many of the keyboards had IBM assembly part numbers 52G9658, 42H1292, and others. Because of pricing pressures, many of these Model M keyboards were manufactured with a new lower-cost keyboard design to improve the keyboard business. [5] Lighter weight plastic, integrated keyboard cable, and uniform print color on the keys were some of the changes made. In 1996, the keyboard business was sold to Lexmark.

Fig 14 Original document 3

Summary on keyboard

Summary	Document 1	Document 2	Document 3	Statistics
---------	------------	------------	------------	------------

keyboard is a typewriter-style device, which uses an arrangement of buttons or keys, to act as mechanical levers or electronic switches. Introduced in 1985 and manufactured by IBM, Lexmark and Unicom, the vast majority of Model M keyboards feature a buckling spring key design and many have fully swappable keycaps. Following the decline of punch cards and paper tape, interaction via teleprinter-style keyboards became the main input device for computers. The PC keyboard changed years, often at the launch of new IBM PC versions.

A keyboard typically has characters engraved or printed on the keys and each press of a key typically corresponds to a single written symbol. Early models of Enhanced I (notably those manufactured by Northgate Ltd.) maintained the layout with function keys on the left side, arranged in two columns of six pairs.

While most keyboard keys produce letters, numbers or signs (characters), other keys or simultaneous key presses can produce actions or computer commands. Additional key (2) and Menu key added (one Windows key to the right of the left control key, the other and the Menu key to the left of the right control key).

Despite the development of alternative input devices, such as the mouse, touchscreen, pen devices, character recognition and voice recognition, the keyboard remains a commonly used and most versatile device used for direct (human) input into computers. The Windows keyboard was introduced for use with the Windows 95 operating system. In normal usage, the keyboard is used to type text and numbers into a word processor, text editor or other programs. Most modern PCs, whether supplied with Windows or not, are now delivered with this layout.

Name	Keys	Description	Image
PC/XT	83	original left-hand side <u>function key</u> (F key) columns, F1 through F10; electronically incompatible with PC/AT keyboard types	
PC/AT	84	additional <SysRq>, i.e. System Request; numerical block clearly separated from main keyboard; added indicator LEDs for Caps/Scroll/Num lock	
		additional navigation and control keys; 12 F keys in separate row along top, grouped F1-4, F5-8, and F9-12. Early models of Enhanced keyboard (notably those manufactured by Northgate Ltd.) maintained the layout with function keys on the left side, arranged in two columns of six pairs. This layout was more efficient for touch typists but was superseded in the marketplace by that with F-keys along the top. There are different versions of the Enhanced keyboard layout:	

Fig 15 Summary

Summary on keyboard

Summary	Document 1	Document 2	Document 3	Statistics
---------	------------	------------	------------	------------

Number of words in original document = 6095
 Number of words in summary = 702
 Percentage (Words) = 11.5176374077
 Number of lines in document = 316
 Number of lines in summary = 16
 Percentage (Lines) = 5.06329113924

Fig 16 Statistics

IX. FUTURE SCOPE AND CONCLUSION

Our system can be improved further by:

1. Moving the system from extractive based summarization to abstractive approach. Abstractive approach would create summaries that humans can easily identify with using NLP, text generation etc.
2. After using the system it can be made to learn (supervised) through web page selection and personalized summarization.

Providing the information to user in a world full of data is a challenging job because of various ways to

express it and available sources of information. In this paper, we have stated a simple algorithm to implement the summarization of multiple documents. A system based on this algorithm that summarizes contents from multiple documents and presents them in a precise format has been developed and would help user to enhance his searching experience, and to gather information more effectively and understand it.

X. REFERENCES

- [1] Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit, By Steven

- Bird, Ewan Klein, Edward Loper, O'Reilly
Media June 2009.
- [2] Naresh Kumar Nagwani, Dr. Shrish
Verma, "A Frequent Term and Semantic
Similarity based Single Document Text
Summarization Algorithm", International
Journal of Computer Applications (0975 –
8887) Volume 17– No.2, March 2011.
- [3] Dragomir R. Rade and Weiguo Fan and
Zhu Zhang, "WebInEssence: A
Personalized Web-Based Multi-Document
Summarization and Recommendation
System"
- [4] Yongzheng Zhang, Nur Zincir-Heywood
and Evangelos Milios, "Term-based
Clustering and Summarization of Web
Page Collections"
- [5] Mr. Vikrant Gupta, Ms. Priya Chauhan,
Dr. Sohan Garg. Mrs. Anita Borude, Prof.
Shobha Krishnan "A Statistical Tool for
Multi-Document Summrization" ,
International Journal of Scientific and
Research publication, Volume 2, Issue 5,
may 2012 ISSN 2250-3153
- [6] Liu, H., and Singh, P.: ConceptNet — A
Practical Commonsense Reasoning Tool-
Kit, BT Technology Journal 22(4), volume
22, Kluwer Academic Publishers, 211–
226, 2004.

XI. BIOGRAPHY

Mrs. Kiruthika .M is currently working with Fr. C. Rodrigues Institute of Technology, Vashi, NaviMumbai as Associate Professor in Computer Engineering Department. Her total teachingexperience is 17 years. Her Research area is Data Mining,Webmining,Databases. She has doneB.E(Electronics and Communication Engineering) in 1992 from BharathidasanUniversity. Shehas completed M.E (CSE) in 1997 from NIT, Trichy . She has published 08 papers inInternational Journal,11 papers in International Conferences and 10 papers in NationalConferences.

Mrs. Kiruthika .M (Professor Author)