DELHI TECHNOLOGICAL UNIVERSITY

Auto Text Summarization

Automated approach to summarize plain text data

Team:

S. GANESAN	SUMIT DUGAR	ROHIT LARIA	SUYASH GUPTA
DSE/2K9/050	DSE/2K9/058	DSE/2K9/049	DSE/2K9/050

Extractive summarization works by selecting a subset of sentences from the original text. Our proposed work aims at finding the important sentences using statistical properties like frequency of word, occurrence of important information in the form of numerical data, proper noun, keyword and sentence similarity factor.

CERTIFICATE

We **Sumit Dugar, Rohit Laria, S. Ganesan, Suyash Gupta** hereby solemnly affirm that the project report entitled **TEXT SUMMARIZATION**, being submitted by us in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Software Engineering, to the DELHI TECHNOLOGICAL UNIVERSITY (Formerly Delhi College of Engineering), is a record of bonafide work carried out by me under the guidance of Dr. **Rajni Jindal**. The work reported in this report in full or in part has not been submitted to any University or Institute for the award of any degree or diploma.

Place: ……..

Date: ……..

Rajni Jindal
(Assistant Professor)

Dr. Daya Gupta
(**Head of the Department**)

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to Dr. Rajni Jindal. Their understanding, encouragement and personal guidance have provided a good foundation for the dissertation.

It is our pleasant duty to thank all our seniors, friends, co-interns and well-wishers for their instant and efficient co-operation.

Finally, we are extremely indebted to our parents for providing the constant encouragement, financial and moral support to enable us to come up to this level in our lives.

S. GANESAN
SUMIT DUGAR
ROHIT LARIA
SUYASH GUPTA

Flow of Content

1. Abstract of the Project
2. Introduction to Data Mining
3. Concept of Text Summarization
4. System Requirements
5. Software Requirements
6. Brief Algorithm Description
7. Interface Description
8. Testing Report
9. Limitation of the Software
10. Conclusion
11. References
12. Team Members Description

Abstract

Summaries are an important tool for familiarizing oneself with a subject area. Text summaries are essential when forming an opinion on if reading a document in whole is necessary for our further knowledge aquiring or not. In other words, summaries save time in our daily work. To write a summary of a text is a non-trivial process where one, on one hand has to extract the most central information from the original text, and on the other has to consider the reader of the text and her previous knowledge and possible special interests. Today there are numerous documents, papers, reports and articles available in digital form, but most of them lack summaries. The information in them is often too abundant for it to be possible to manually search, sift and choose which knowledge one should aquire. This information must instead be automatically filtred and extracted in order to avoid drowning in it.

Automatic Text Summarization is a technique where a computer summarizes a text. A text is given to the computer and the computer returns a shorter less redundant extract of the original text. So far automatic textsummarization has not yet reached the quality possible with manual summarization, where a human interprets the text and writes a completely new shorter text with new lexical and syntactic choices. However, automatic text summarization is untiring, consistent and always available.

Evaluating summaries and automatic text summarization systems is not a straightforward process. What exactly makes a summary beneficial is an elusive property. Generally speaking there are at least two properties of the summary that must be measured when evaluating summaries and summarization systems - the Compression Ratio, i.e. how much shorter the summary is than the original, and the Retension Ratio, i.e. how much of the central information is retained. This can for example be accomplished by comparison with existing summaries for the given text. One must also evaluate the qualitative properties of the summaries, for example how coherent and readable the textis. This is usually done by using a panel of human judges. Furthermore, one can also perform task-based evaluations where one tries to discern to what degree the resulting summaries are beneficent for the completion of a specific task.

Thus, the technique has been developed for many years and in recent years, with the increased use of the Internet, there have been an awakening interest for summarization techniques. Today the situation is quite the opposite from the situation in the sixties. Today storage is cheap and seemingly limitless.

Digitally stored information is available in abundance and in a myriad of forms to an extent as to making it near impossible to manually search, sift and choose which information one should incorporate. This information must instead be filtered and extracted in order to avoiding drowning in it.

INTRODUCTION TO DATA MINING

Overview

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The Foundations of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases can be larger in both depth and breadth:

- **More columns**. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

- **More rows**. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."2 Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability)."3

The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

How Data Mining Works

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship's captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure.

An Architecture for Data Mining

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

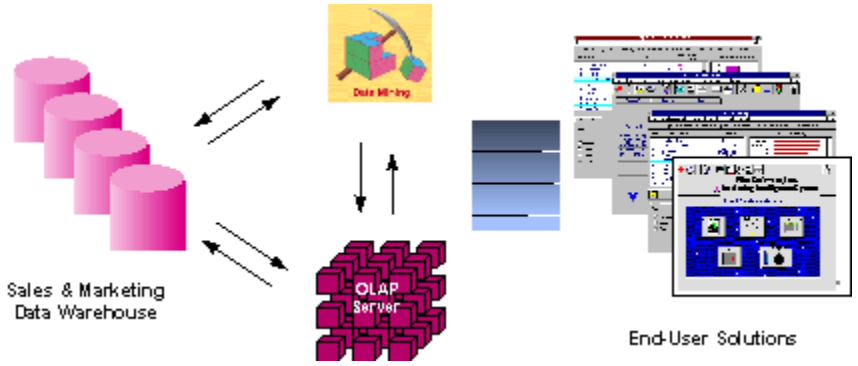


Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused

business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

Conclusion

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

Concept of Text Summarization

Overview

With the coming of the information revolution, electronic documents are becoming a principle media of business and academic information. Thousands and thousands of electronic documents are produced and made available on the internet each day. In order to fully utilizing these on-line documents effectively, it is crucial to be able to extract the giz of these documents. Having a Text Summarization system would thus be immensely useful in serving this need.

In order to generate a summary, we have to identify the most important pieces of information from the document, omitting irrelevant information and minimizing details, and assemble them into a compact coherent report. This however, is easier said than done as it involves some of the main problems of natural language processing. To produce a domain-independent system would require work in natural language understanding, semantic representation, discourse models, world knowledge, and natural language generation. Successes in domain-independent systems are few and limited to identifying key passages and sentences of the document. More successful systems have been produced for limited domain applications such as report generations for weather, financial and medical databases.

Information Extraction

Information extraction systems analyze unrestricted text in order to extract information specific to a particular domain. It does not attempt to understand all of the text in all input documents, but analyzes portions of documents that contain relevant information. The relevancy of the information is determined by pre-defined domain guidelines which must specify, as accurately as possible, exactly what types of information the system is expected to find. One method to think of information extraction in terms of database construction. Here unstructured text documents is converted into classfied database entries which are then used to fill a "template". A summary report can then be generated using pieces of canned text from the "template".

Examples of Existing Information Extraction Systems

A number of Information Extraction Summarization Systems have been developed in specific fields. Even though most of these systems are not currently used in the internet, the potential is great and implementation of such systems in the internet is relatively simple.

Systems have been designed to monitor technical articles in the field of microelectronic chips fabrication. These systems could be easily modified to target some other fields by simply modifying the domain guidelines and extended to use on the internet where many articles could be found.

Systems have been designed to analyse and summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results and therapeutic treatments. These systems could be used to help health care providers with quality assurance studies. Central databases accessible from different medical centres could be setting up to facilitate transfer of patients as well as for emergencies when the medical records are required immediately. These systems could again be extended to analyse some other databases, eg. financial statement databases.

Case-Based Approach

Here the input document is matched against a corpus of relevant and irrelevant texts. Instead of having an explicit set of domain guidelines from a user, the system simply exploits a "training corpus" of representative texts that a user or domain expert has manually classified as either relevant or irrelevant. These predefined representative texts are matched with the document, using statistical techniques to determine the texts that are relevant to the domain of the document. Basically texts that contains only general information are unlikely to be highly correlated with the domain because similar cases will be found in irrelevant as well as relevant texts in the training corpus.Texts that are too specific are also unlikely because there will be very few matching cases. Thus using this statistical technique, only representative texts that contain ley domain-specific information will be extracted. These matched relevant texts could then be used to generate a summary of the text.

Basically, the Case-based approach can be think of as an extension of the basic information extraction system. The problem with the information extraction system is that it retains virtually all the information that is relevant to the domain without any discrimination between important information and details and general information. By including a statistical test in the Case-based approach, we are able to get an idea of the importance and relevancy of the informations being extracted for the summary.

Domain Independent Approaches

As being pointed out previously, domain-independent text summarization is much more difficult than a domain-specific task. A good summary should include the most relevant information, but omit details and irrelevant information. However, different piece of information will be relevant to different people, depending on their individual interests and needs, ie the domain. Thus not many successful, fully operational systems have been developed. In the following, we will look at some of the most common approaches used as well as some experimental systems produced.

Document Abridgement

Here a summary is produced by deleting irrelevant texts from the document, retaining only the key passages and sentences of the document. Basically, a typical system consist of two sections, the *Reader* and the *Extractor*. The Reader basically reads in the input text and converts it into internal representations, taking into account the word occurrences and calculating the word weights. The Extractor then determines the particular sentences to be included in the summary by analyzing the word weightings and sentence weightings and then generating the summary from the internal representations.

An example of an experimental system using these methods is the Automatic News Extraction System (ANES) developed by Lisa Rau. This system aims to produce summaries of news from many different sources, had achieved relatively good results in spite of the fact that it is limited by the constraint that it is publication-independent. If developed, this function would prove to be extremely useful for categorising and locating informations on the internet by providing summaries to all wide varieties of documents available on the net.

Another example of a domain-independent text summarization available on the net is the NetSumm web page summarization tool which is able to highlight the key points in articles, as well as abridge documents.

Conclusion

Even though Natural Language Processing is still at its very early stage of development, its potential of being widely used in the internet is already emerging. Though many of the current systems are still in their experimental stages, results have been promising. With the recent explosive growth of the internet, much more extensive usage of NLP would be expected in the near future. Multimedia databases and digital libraries might well overtake their present counterparts as the main media for information sharing and communications. NLP might well be "the subject" of the future.

Software Description

SCORING ALGORITHM:

Our program essentially works on the following logics:

Word Score

PRIMARY WORD SCORE

1. **Stop Words:** These are some insignificant words that are so commonly used in the English language that no text can be created without them. They therefore provide no real idea about the textual theme, and have therefore, been neglected while scoring sentences.

Eg. I, a, an, of, am, the, et cetera.

2. **Cue Words:** These are words usually used in concluding sentences of a text, making sentences containing them crucial for any given summary. Cue Words provide closure to a given matter, and have therefore, been given prime importance while scoring sentences.

Eg. Thus, hence, summary, conclusion, et cetera.

3. **Basic Dictionary Words:** 850 words of the English language have been defined as the most frequently used words that add meaning to a sentence. These words form the backbone of our algorithm, and have been vital in the creation of a sensible summary. We have hence, given these words moderate importance while scoring sentences.

4.**Proper Nouns, Numbers and abbreviations:** Proper Nouns in most cases form the central theme of a given text. Albeit, the identification of proper nouns without the use of linguistic methods was difficult, we have been successful in identifying them in most cases. Proper Nouns provide semantics to the summary, and have therefore been given high importance while scoring sentences.

FINAL WORD SCORE

6. **Word Frequency:** Once basic scores have been allotted to words, their final score is calculated on the basis of their frequency of occurrence in the document. Words in the text which are repeated more frequently than others contain a more profound impression of the context, and have therefore been given a higher importance.

Sentence Score:

1. **Primary Score:** Using the above methods, a final word score is calculated, and the sum of word scores of a sentence gives a sentence score. This gives long sentence a clear advantage over their smaller counterparts, which might not necessarily be of lesser importance.

2. **Final Score:** By multiplying the score so obtained by the ratio "average length / current length" the above drawback can be nullified to a large extent, and a final sentence score is obtained.

Optimizations

1. List of basic words, cue words etc are delivered in the form of a text file along with the software. Contents of these text files are stored in the memory at runtime using 'Maps' (in C++).

2. The entered text has been stored into two types of Data Structures:

a. Linked List

b. Maps

3. Special consideration has been given to words written in quotes because many times these words convey important facts.

4. Generally first sentence from a text presents main points about the content of the entire text. So words appearing in the initial sentence are considered more important.

5. A special list of titles such as Mr, Mrs, Sir,Capt etc is also considered while extracting words from a sentence.

6.A special file called 'log.txt' is created for synchronization between C++ and VB during the runtime.

Important Features

1.GUI is implemented, which allows the user to directly input text or select a file from his system, and get output displayed on the screen itself.

2. Output is written in a new file called '<original file name>_summary.txt', it's default location is same as the input file. User can also decide the destination of the output file.

3. If a user inputs the text by himself a file called 'article.txt' is created in the 'summary' directory. Now this file is treated as the input file.

4. Logic is implemeted using C++ and GUI is implemented in Visual basic.

5. An intaller file is available for easy distribution of the software.

LIMITATIONS:

1. For data with matter on multiple topics, generating a balanced summary is not possible.

2. Pronouns create another problem, as the parent noun cannot be identified.

3. The software does not provide the user freedom to mention the percentage of summary.

IMPROVEMENTS:

1. Some more scoring techniques, based upon the placement of sentences could have been implemented.

2. Reading from Word documents and HTML files could have been implemented.

3. Natural Language processing could have been implemented.

ALGORITHM DESCRIPTION

1. Open the input file, if it has some content then go to next step otherwise a warning is poped.

2. List of basic words , cue words and title words are loaded into the momory using maps.

3. Extract the first sentence from the given text file and create a map of the words present in this sentence.

4. Extract every sentence from the file and store it in a linked list.

5. For every extracted sentence extract words from them, filter the extracted words and pass them through various word scoring techniques. Firstly a primary socre is computed and in the end a final score for that word is calculated.

6. Before going for the next sentence compute the total score of this sentence. This is done by using sentence scoring techinques.

7. Sort the sentences in decending order based on their final scores.

8. Select 50% sentences from the above list and sort them in ascending order based on their sentence number.

9.The final list of sentences is the summary of the given text.

Future Prospects:

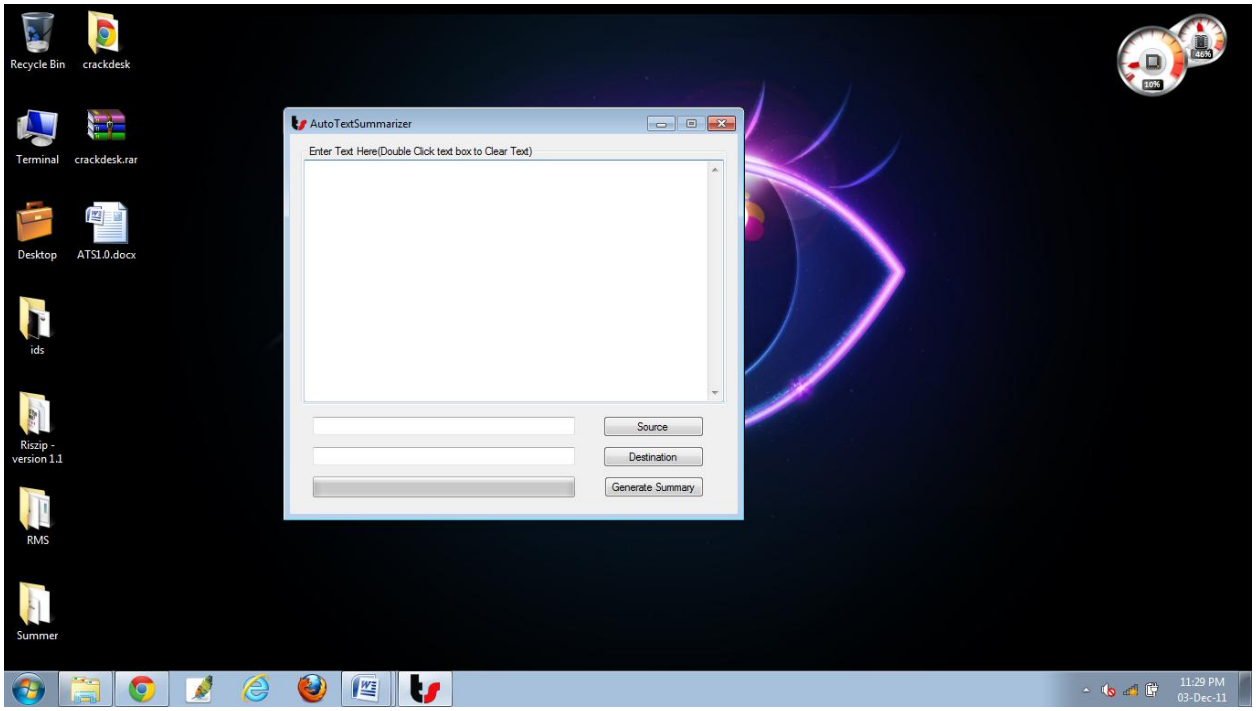
The possibilities in this project are endless. With the development of Natural Language Processing (NLP), the following don't remain mere thoughts...

a. Generating newspaper headlines, given the article.

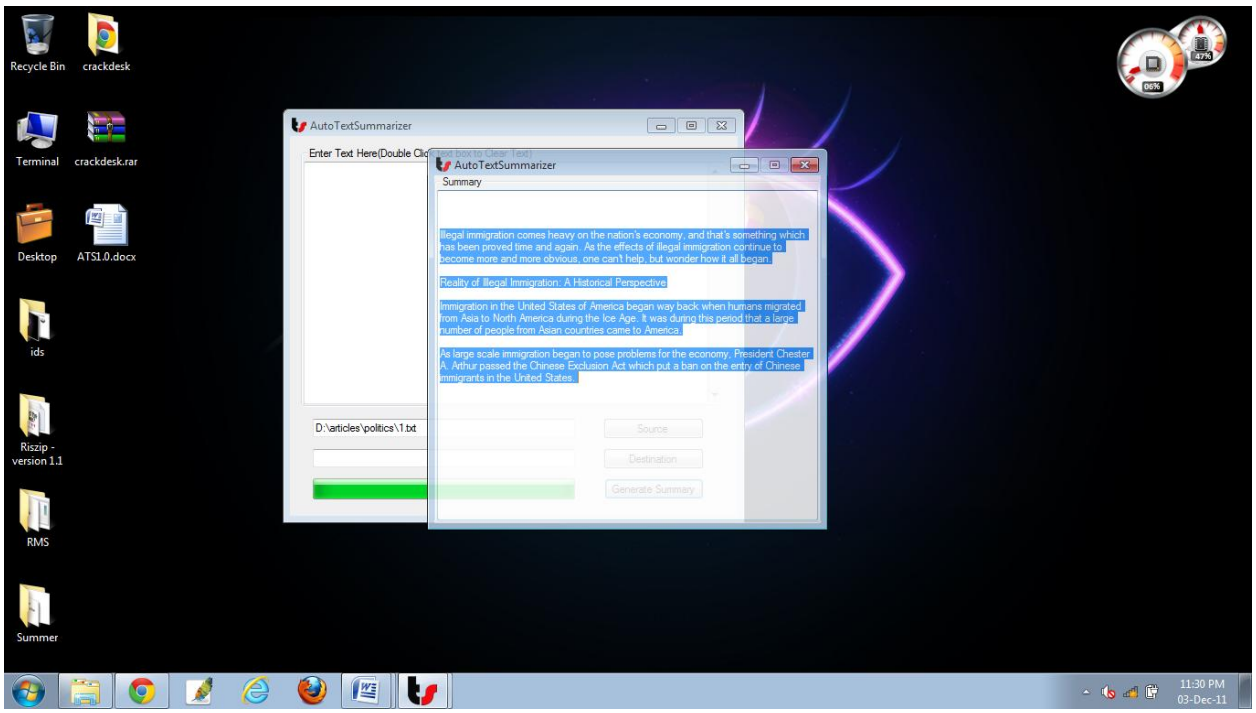
b. Filling up forms, given text containing the necessary data.

c. Creating a bio-data, from a textual detail if the person.

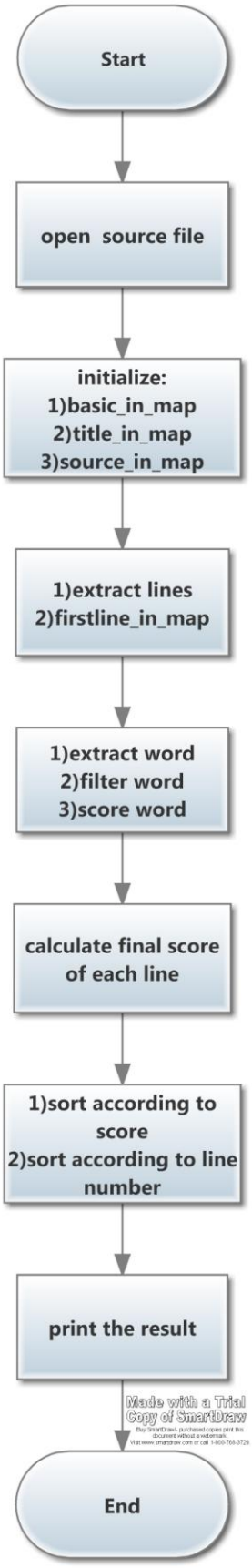
INTERFACE SNAPSHOTS



Initial interface



Interface after generating the summary


Start
open source file
initialize:
1)basic_in_map
2)title_in_map
3)source_in_map
1)extract lines
2)firstline_in_map
1)extract word
2)filter word
3)score word
calculate final score of each line
1)sort according to score
2)sort according to line number
print the result
Made with a Trial Copy of SmartDraw
End

References

1. Amy J.C Trappey, Charles V .Trappey, "An R&D Knowledge Management method of patent document summarization", Industrial Management & Data System, vol 108 pp -245-257,2008.
2. Edmundson, H. P(1969) “New method in automatic Extracting” journal of ACM 1969, 16(2): 264-285.
3. Erkan G., and Radev, D. R., “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization”, J. Artif. Intell. Res. (JAIR), 22, pp. 457-479, 2004.
4. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summarizing text documents: sentence selection and evaluation metrics. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’99), Berkeley, CA, USA, pp. 121–128. Hahn, U., Mani, I., 2000. The challenges of automatic summarization. IEEE-Computer 33 (11), 29–36.
5. Websites : Google , Wikipedia.