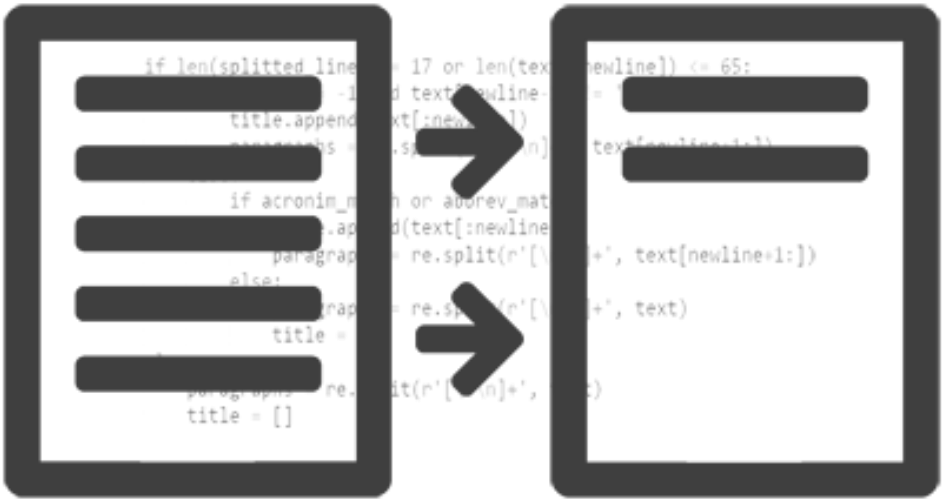
**Natural Language Processing (UML602) Project Report**

**(B.E 3rd Year May 2018)**



# Text Summarization


**Submitted by:**

Abhi Mahajan – 101683033

Group: COE-1


**Submitted to:**

Ms. Prabhleen Juneja


THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

# Index

# TEXT SUMMARIZATION


## Abstract:

In this new era where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.


## Introduction:

Before going to the Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is, it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language. There are two different groups of text summarization: indicative and informative. Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. On the other hand, the informative summarization systems gives concise information of the main text. The length of informative summary is 20 to 30 percent of the main text.

## Applications:

1. Summaries of Email Threads
2. Action Items from A Meeting
3. Simplifying Text by Compressing Sentences

## Types of text summarization:

1. Extractive text summarization
2. Abstractive text summarization

## Abstractive and Extractive Summarization:

There are two main approaches to the task of summarization—extraction and abstraction. Extraction involves concatenating extracts taken from the corpus into a summary, whereas abstraction involves generating novel sentences from information extracted from the corpus. It has been observed that in the context of multi-document summarization of news articles, extraction may be inappropriate because it may produce summaries which are overly verbose or biased towards some sources. However, there has been little work identifying specific factors which might affect the performance of each strategy in summarizing evaluative documents containing opinions and preferences, such as customer reviews or blogs. This chapter aims to address this gap by exploring one dimension along which the effectiveness of the two paradigms could vary; namely, the contra-versatility of the opinions contained in the corpus. We make the following contributions. Firstly, we define a measure of contra-versatility of opinions in the corpus based on information entropy. Secondly, we run a user study to test the hypothesis that a controversial corpus has greater need of abstractive methods and consequently of NLG techniques. Intuitively, extracting sentences from multiple users whose opinions are diverse and wide-ranging may not reflect the overall opinion, whereas it may be adequate content-wise if opinions are roughly the same across users. As a secondary contribution, we propose a method for structuring text when summarizing controversial corpora. This method is 2 used in our study for generating abstractive summaries. The results of the user study support our hypothesis by showing that a NLG summarizer outperforms an extractive summarizer to a larger extent when the contra-versatility is high.

## Features for extractive text summarization:

Most of the current automated text summarization systems use extraction method to produce a Summary. Sentence extraction techniques are commonly used to produce extraction summaries. One of the methods to obtain suitable sentences is to assign some numerical measure of a sentence for the summary called sentence scoring and then select the best sentences to form document summary based on the compression rate. In the extraction method, compression rate is an important factor used to define the ratio between the length of the summary and the source text. As the compression rate increases, the summary will be larger, and more insignificant content is contained. While the compression rate decreases the summary to be short, more information is lost. In fact, when the compression rate is 5-30%, the quality of summary is acceptable.

## Extractive summarization methods:

1. Term Frequency-Inverse Document Frequency (TF-IDF) method:
2. Cluster based method:
3. Graph theoretic approach:
4. Machine Learning approach:
5. Text summarization with neural networks
6. Automatic text summarization based on fuzzy logic

## Term Frequency-Inverse Document Frequency (TF-IDF) method:

It is a numerical statistic which reflects how important a word is in a given document. The TF-IDF value increases proportionally to the number of times a word appears in the document. This method mainly works in the weighted term-frequency and inverse sentence frequency paradigm .where sentence-frequency is the number of sentences in the document that contain that term. These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary. Summarization is query-specific. The hypothesis assumed by this approach is that if there are ‘‘more specific words’’ in a given sentence, then the sentence is relatively more important. The target words are usually nouns. This method performs a comparison between the term frequency (tf) in a document -in this case each sentence is treated as a document and the document frequency (df).

## Algorithm:

### Term Frequency-Inverse Document Frequency

1. Get URL from user input
2. Web crawl to extract the natural language from the URL html (by paragraphs <p>).
3. Execute the summarize class algorithm (implemented using NLTK) on the extracted sentences stemmed to their root stems.
4. The algorithm rank sentences according to the frequency of the words they contain, and the top sentences are selected for the final summary.
5. Return the highest ranked sentences as a final summary.

## Libraries Used:

1. **Web crawling - Beautiful Soup** is a Python library for pulling data out of HTML and XML files. It works to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.
2. **Text summarization - Nltk** (Natural Language Toolkit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

# Results:

## File to be summarized

### Chapter 1. Mr. Sherlock Holmes

Mr. Sherlock Holmes, who was usually very late in the mornings, save upon those not infrequent occasions when he was up all night, was seated at the breakfast table. I stood upon the hearth-rug and picked up the stick which our visitor had left behind him the night before. It was a fine, thick piece of wood, bulbous-headed, of the sort which is known as a "Penang lawyer." Just under the head was a broad silver band nearly an inch across. "To James Mortimer, M.R.C.S., from his friends of the C.C.H.," was engraved upon it, with the date "1884." It was just such a stick as the old-fashioned family practitioner used to carry—dignified, solid, and reassuring.

"Well, Watson, what do you make of it?"

Holmes was sitting with his back to me, and I had given him no sign of my occupation.

"How did you know what I was doing? I believe you have eyes in the back of your head."

"I have, at least, a well-polished, silver-plated coffee-pot in front of me," said he. "But, tell me, Watson, what do you make of our visitor's stick? Since we have been so unfortunate as to miss him and have no notion of his errand, this accidental souvenir becomes of importance. Let me hear you reconstruct the man by an examination of it."

"I think," said I, following as far as I could the methods of my companion, "that Dr. Mortimer is a successful, elderly medical man, well-esteemed since those who know him give him this mark of their appreciation."

"Good!" said Holmes. "Excellent!"

"I think also that the probability is in favour of his being a country practitioner who does a great deal of his visiting on foot."

"Why so?"

"Because this stick, though originally a very handsome one has been so knocked about that I can hardly imagine a town practitioner carrying it. The thick-iron ferrule is worn down, so it is evident that he has done a great amount of walking with it."

"Perfectly sound!" said Holmes.

"And then again, there is the 'friends of the C.C.H.' I should guess that to be the Something Hunt, the local hunt to whose members he has possibly given some surgical assistance, and which has made him a small presentation in return."

"Really, Watson, you excel yourself," said Holmes, pushing back his chair and lighting a cigarette. "I am bound to say that in all the accounts which you have been so good as to give of my own small achievements you have habitually underrated your own abilities. It may be that you are not yourself luminous, but you are a conductor of light. Some people without possessing genius have a remarkable power of stimulating it. I confess, my dear fellow, that I am very much in your debt."

He had never said as much before, and I must admit that his words gave me keen pleasure, for I had often been piqued by his indifference to my admiration and to the attempts which I had made to give publicity to his methods. I was proud, too, to think that I had so far mastered his system as to apply it in a way which earned his approval. He now took the stick from my hands and examined it for a few minutes with his naked eyes. Then with an expression of interest he laid down his cigarette, and carrying the cane to the window, he looked over it again with a convex lens.

"Interesting, though elementary," said he as he returned to his favourite corner of the settee. "There are certainly one or two indications upon the stick. It gives us the basis for several deductions."

"Has anything escaped me?" I asked with some self-importance. "I trust that there is nothing of consequence which I have overlooked?"

"I am afraid, my dear Watson, that most of your conclusions were erroneous. When I said that you stimulated me I meant, to be frank, that in noting your fallacies I was occasionally guided towards the truth. Not that you are entirely wrong in this instance. The man is certainly a country practitioner. And he walks a good deal."

"Then I was right."

"To that extent."

"But that was all."

"No, no, my dear Watson, not all—by no means all. I would suggest, for example, that a presentation to a doctor is more likely to come from a hospital than from a hunt, and that when the initials 'C.C.' are placed before that hospital the words 'Charing Cross' very naturally suggest themselves."

"You may be right."

"The probability lies in that direction. And if we take this as a working hypothesis we have a fresh basis from which to start our construction of this unknown visitor."

"Well, then, supposing that 'C.C.H.' does stand for 'Charing Cross Hospital,' what further inferences may we draw?"

"Do none suggest themselves? You know my methods. Apply them!"

"I can only think of the obvious conclusion that the man has practised in town before going to the country."

"I think that we might venture a little farther than this. Look at it in this light. On what occasion would it be most probable that such a presentation would be made? When would his friends unite to give him a pledge of their good will? Obviously at the moment when Dr. Mortimer withdrew from the service of the hospital in order to start a practice for himself. We know there has been a presentation. We believe there has been a change from a town hospital to a country practice. Is it, then, stretching our inference too far to say that the presentation was on the occasion of the change?"

"It certainly seems probable."

"Now, you will observe that he could not have been on the staff of the hospital, since only a man well-established in a London practice could hold such a position, and such a one would not drift into the country. What was he, then? If he was in the hospital and yet not on the staff he could only have been a house-surgeon or a house-physician—little more than a senior student. And he left five years ago—the date is on the stick. So your grave, middle-aged family practitioner vanishes into thin air, my dear Watson, and there emerges a young fellow under thirty, amiable, unambitious, absent-minded, and the possessor of a favourite dog, which I should describe roughly as being larger than a terrier and smaller than a mastiff."

I laughed incredulously as Sherlock Holmes leaned back in his settee and blew little wavering rings of smoke up to the ceiling.

"As to the latter part, I have no means of checking you," said I, "but at least it is not difficult to find out a few particulars about the man's age and professional career." From my small medical shelf I took down the Medical Directory and turned up the name. There were several Mortimers, but only one who could be our visitor. I read his record aloud.

*"Mortimer, James, M.R.C.S., 1882, Grimpen, Dartmoor, Devon.*
*House-surgeon, from 1882 to 1884, at Charing Cross Hospital.*
*Winner of the Jackson prize for Comparative Pathology,*
*with essay entitled 'Is Disease a Reversion?' Corresponding*
*member of the Swedish Pathological Society. Author of*
*'Some Freaks of Atavism' (Lancet 1882). 'Do We Progress?'*
*(Journal of Psychology, March, 1883). Medical Officer*
*for the parishes of Grimpen, Thorsley, and High Barrow."*

"No mention of that local hunt, Watson," said Holmes with a mischievous smile, "but a country doctor, as you very astutely observed. I think that I am fairly justified in my inferences. As to the adjectives, I said, if I remember right, amiable, unambitious, and absent-minded. It is my experience that it is only an amiable man in this world who receives testimonials, only an unambitious one who abandons a London career for the country, and only an absent-minded one who leaves his stick and not his visiting-card after waiting an hour in your room."

"And the dog?"

"Has been in the habit of carrying this stick behind his master. Being a heavy stick the dog has held it tightly by the middle, and the marks of his teeth are very plainly visible. The dog's jaw, as shown in the space between these marks, is too broad in my opinion for a terrier and not broad enough for a mastiff. It may have been—yes, by Jove, it is a curly-haired spaniel."

He had risen and paced the room as he spoke. Now he halted in the recess of the window. There was such a ring of conviction in his voice that I glanced up in surprise.

"My dear fellow, how can you possibly be so sure of that?"

"For the very simple reason that I see the dog himself on our very door-step, and there is the ring of its owner. Don't move, I beg you, Watson. He is a professional brother of yours, and your presence may be of assistance to me. Now is the dramatic moment of fate, Watson, when you hear a step upon the stair which is walking into your life, and you know not whether for good or ill. What does Dr. James Mortimer, the man of science, ask of Sherlock Holmes, the specialist in crime? Come in!"

The appearance of our visitor was a surprise to me, since I had expected a typical country practitioner. He was a very tall, thin man, with a long nose like a beak, which jutted out between two keen, gray eyes, set closely together and sparkling brightly from behind a pair of gold-rimmed glasses. He was clad in a professional but rather slovenly fashion, for his frock-coat was dingy and his trousers frayed. Though young, his long back was already bowed, and he walked with a forward thrust of his head and a general air of peering benevolence. As he entered his eyes fell upon the stick in Holmes's hand, and he ran towards it with an exclamation of joy. "I am so very glad," said he. "I was not sure whether I had left it here or in the Shipping Office. I would not lose that stick for the world."

"A presentation, I see," said Holmes.

"Yes, sir."

"From Charing Cross Hospital?"

"From one or two friends there on the occasion of my marriage."

"Dear, dear, that's bad!" said Holmes, shaking his head.

Dr. Mortimer blinked through his glasses in mild astonishment. "Why was it bad?"

"Only that you have disarranged our little deductions. Your marriage, you say?"

"Yes, sir. I married, and so left the hospital, and with it all hopes of a consulting practice. It was necessary to make a home of my own."

"Come, come, we are not so far wrong, after all," said Holmes. "And now, Dr. James Mortimer—"

"Mister, sir, Mister—a humble M.R.C.S."

"And a man of precise mind, evidently."

"A dabbler in science, Mr. Holmes, a picker up of shells on the shores of the great unknown ocean. I presume that it is Mr. Sherlock Holmes whom I am addressing and not—"

"No, this is my friend Dr. Watson."

"Glad to meet you, sir. I have heard your name mentioned in connection with that of your friend. You interest me very much, Mr. Holmes. I had hardly expected so dolichocephalic a skull or such well-marked supra-orbital development. Would you have any objection to my running my finger along your parietal fissure? A cast of your skull, sir, until the original is available, would be an ornament to any anthropological museum. It is not my intention to be fulsome, but I confess that I covet your skull."

Sherlock Holmes waved our strange visitor into a chair. "You are an enthusiast in your line of thought, I perceive, sir, as I am in mine," said he. "I observe from your forefinger that you make your own cigarettes. Have no hesitation in lighting one."

The man drew out paper and tobacco and twirled the one up in the other with surprising dexterity. He had long, quivering fingers as agile and restless as the antennae of an insect.

Holmes was silent, but his little darting glances showed me the interest which he took in our curious companion. "I presume, sir," said he at last, "that it was not merely for the purpose of examining my skull that you have done me the honour to call here last night and again today?"

"No, sir, no; though I am happy to have had the opportunity of doing that as well. I came to you, Mr. Holmes, because I recognized that I am myself an unpractical man and because I am suddenly confronted with a most serious and extraordinary problem. Recognizing, as I do, that you are the second highest expert in Europe—"

"Indeed, sir! May I inquire who has the honour to be the first?" asked Holmes with some asperity.

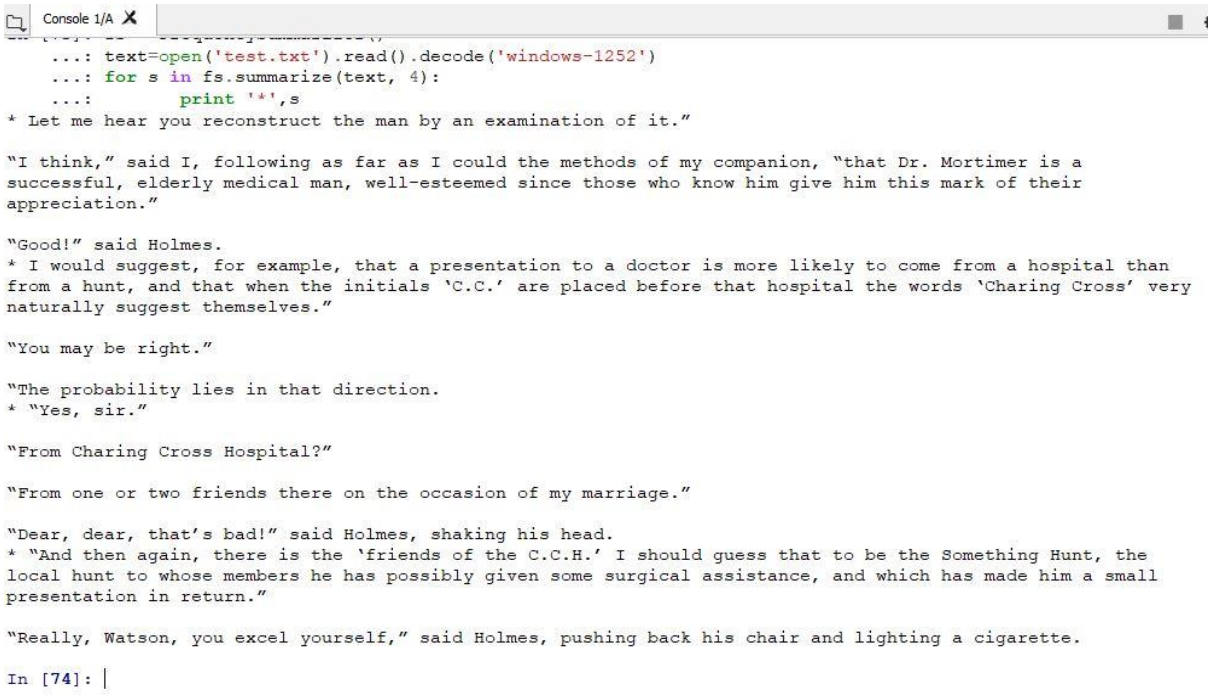
"To the man of precisely scientific mind the work of Monsieur Bertillon must always appeal strongly."

"Then had you not better consult him?"

"I said, sir, to the precisely scientific mind. But as a practical man of affairs it is acknowledged that you stand alone. I trust, sir, that I have not inadvertently—"

"Just a little," said Holmes. "I think, Dr. Mortimer, you would do wisely if without more ado you would kindly tell me plainly what the exact nature of the problem is in which you demand my assistance."

File summarized




# References:


1. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.: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 1999, pp. 121–128. ACM, New York (1999). http://doi.acm.org/10.1145/312624.312665
2. Graham, R., Knuth, D., Patashnik, O.: Concrete Mathematics: A Foundation for Computer Science. Addison-Wesley, Boston (1994)MATHGoogle Scholar
3. Graham, R., Knuth, D., Patashnik, O.: Concrete Mathematics: A Foundation for Computer Science. Addison-Wesley, Boston (1994)MATHGoogle Scholar
4. Hovy, E., Lin, C.-Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006) (2006)Google Scholar