ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 7, Issue 8 , August 2020

Document Summarization

Twisha Shah, Abhya Tripathi, Mansi Jha, Sonali Bodekar

Dept. of Computer Science, Usha Mittal Institute Of Technology, Mumbai, India
Dept. of Computer Science, Usha Mittal Institute Of Technology, Mumbai, India
Dept. of Computer Science, Usha Mittal Institute Of Technology, Mumbai, India
Dept. of Computer Science, Usha Mittal Institute Of Technology, Mumbai, India

ABSTRACT: Document Summarization is a very challenging task in text mining. Summarizing a large document in concise short sentences which is a subgroup of the initial text is called as extractive summarization. There are various applications of text summarization, but here the CNN News articles are summarized to its key sentences. In this paper, Topic Modeling Algorithm the Latent Dirichlet Allocation is used to generate extractive text summarization. It is used in capturing important topics from the text and later using distribution weighting mechanism sentences are fetched from the text. The model performs well on the data and fetches the summary for the news article. This helps in saving time to read long texts or documents.

KEYWORDS: Document Summarization; Topic Modeling; Latent Dirichlet Allocation; Extractive Summarization; Abstractive Summarization

I. INTRODUCTION

A summary is a brief form of a large text. It conveys important information about the document or text. There is a tremendous volume of information available on the internet, and it is extremely difficult to get the relevant meaningful information quickly. For searching a particular piece of information we need to go through various documents or plenty of information from the internet. A human being has difficulty in summarizing large documents. To cater to these two problems, automatic document/text summarizing has become necessary nowadays. Rather than reading the whole document, one can identify whether the document is of any advantage by reading the summary. [1]

Document summarization is a means of deriving significant and relevant data from the document and to make a piece of comprehensive and meaningful information. In this project, an extractive summarization of large documents is carried out using Topic Modeling based on the papers [2] and [3]. The document is segmented in a list of sentences and applied to the Latent Dirichlet Allocation (LDA) algorithm to extract main topics. Then using the frequency of words of those topics in sentences, key sentences are extracted having highest distribution to summarize the text. The report is structured below in following sections. The Literature Review in Section II which discusses the work of various authors towards document summarization and LDA. The Section III specifies the actual methodology implemented using LDA model and includes data processing. Empirical results in text modeling and document summarization are discussed in the segment IV. Finally, Section V bestows the conclusion and the future scope

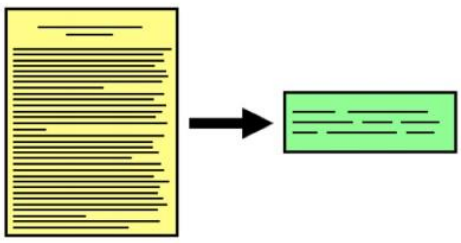


Fig. 1: Document Summarization

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 7, Issue 8 , August 2020

Fig. 2: Types of Document Summarization towards the project. II.

II. LITERATURE REVIEW

In [3], the authors show an experiment to automatic summarize large novel documents through extracting relevant sentences from text documents and then recombining them to form summary. The approach used in this paper is topic modeling on large documents. Firstly, topic modeling algorithm, LDA is applied to extract important topic words from the document. Then associated candidate sentences are extracted. Then, candidate sentence importance evaluation function is used to achieve topic diversity. Finally, with the help of heuristic summarization algorithm, the candidate sentences are used to summarize the document. Further, external resources like SemCor [4] and synonym thesaurus are used for smoothing the summarized text. The evaluation criteria considering Topic diversity and high compression ratio achievement were experimented. The major conclusion to this experiment was that it did achieved topic diversity using the topic modeling LDA approach to summarize the large document with a compression ratio of 0.1-0.2 percent. However, there were problems of text summary redundancy which was needed to be tackled. Also, the extraction of semantic entities were to be further analyzed. In [5], Kazantseva and Szpakowicz describe the approach to summarize short stories. The paper aims at getting user relevant information of the story and not revealing the actual plot of the story. It focuses on main entities of the story documents. An average compression rate of 94 percent was achieved using machine learning tools and rules defined to summarize. However, the quality of the summary was not evaluated. The readability and coherence of the summary was not analyzed. It also failed to summarize large documents. In [6], the authors describe a generative model LDA for a collection of distinct data from a text document. Each topic is tokenized as an approximation of words from the text called latent topics. The fundamental thought is that documents are described over as these latent topics. It is based on the bag-of-words model where the order of words is not considered important [7]. The authors stated the advantage of exchangeable that is conditionally independent and identically distributed through LDA over LSI (Latent Semantic Analysis) and PLSI (Probabilistic LSI). Through LDA a document could be associated with multiple topics. It very wells illustrates how LDA could be extended to be used for topic distribution conditioned over paragraphs and sentences providing a powerful tool for text mining. It also helps in dimensionality reduction over LSI. It assumes a document as a group of words and uses these words to form topics and combine them and provide distribution over a document or set of documents. So these topics could be shared between multiple documents thus, enhancing exchangeability. In [8], Nagwani et al. have summarized multiple documents which are related to one another which helps users to further analyze these documents. As there are multiple documents, it becomes difficult for the summarizer to summarize. Therefore, semantic similar terms are found in these documents and used for summarization.

The approach used for summarization of several documents is MapReduce framework. At first, MapReduce framework is used to perform clustering on big datasets (more than one documents). Then probabilistic LDA algorithm is used to find topics in these clustered documents of similar text. Then a WordNet API is used to compute semantic identical words for a given topic and then sentences are extracted for these terms from the documents. Further, redundant sentences are eliminated and finally the document is summarized. The major conclusion to this experiment is the MapReduce framework is tested up to four nodes and compression rate, computation time is gauged for multiple documents and has performed well. Also, results using semantic similarity provided better performance. In [9], Latent Dirichlet Allocation Topic Modeling algorithm is explained and how it is probabilistic in nature. Topic Modeling helps to find particular themes in a document, how these themes are related and how they change over time. There is no labeled data required for topic modeling. Latent Dirichlet Allocation is a statistical model which portrays a document as a list of topics. In this model, documents are a dissemination of topics which is a spread of words of the document. It assumes to remove all stop words like but, a, the, is, etc. and also assumes that the sequence of words in the document does not matter. So it uses the bag of words as the corpus input and discovers topics from the collection of documents. It computes the conditional probabilities or posterior probabilities of hidden structure that is topics. It can be concluded that the topic model is unsupervised machine learning algorithm and could be used for summarizing and understanding

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 7, Issue 8 , August 2020

our growing amount of information. In [10], users' Twitter messages are analyzed and classified for various applications like breaking news detection, recommendation systems, sentiment analysis, and others. The messages are of short length, so applying topic modeling does not give a good performance. There are experiments conducted using Author-topic models and standard LDA for microblogging environments of classifying user and messages and concluded that LDA gives better performance. There are various aspects compared between these models like the topics generated and their coherence, their quality and so on. Various messages of a single user are aggregated and classified or analyzed. It seems that aggregating messages gives superior performance.

III. METHODOLOGY

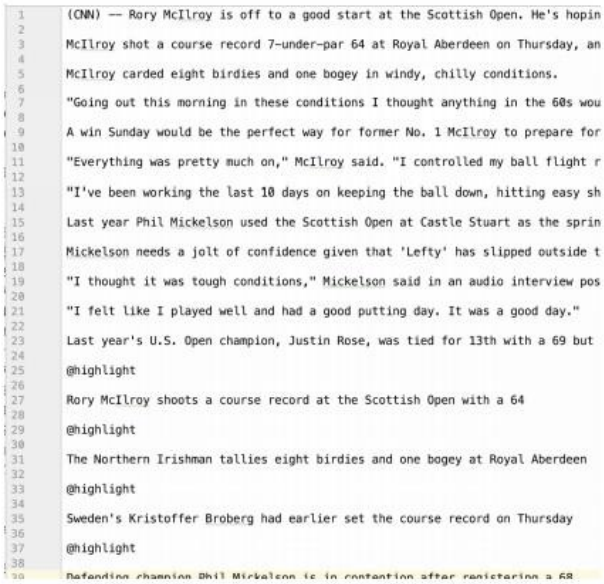
In this project, we have performed extractive summarization on large documents. The main idea of extractive text summarization is to generate a set of n sentences for a document of m sentences where Sn is a subset of Sm [2]. The generated n sentence summary text should contain the relevant information from the underlying document.

A. Reason for using Topic Modeling

The main motive to use topic modeling for summarization is to view a text as a probability of topic words. There is an important metric of topic diversity evaluation required when summarizing a large text. It provides an inherent relation of topics in the documents which will greatly improve the quality of summarization.

B. Data

In this experiment, we have utilized the DeepMind QA Dataset which contain CNN news articles that are collected and prepared by Hermann et al. Details of the Data-set are as follows. Data in form of CNN News stories was collected from the CNN News stories data-set[1]. It contains 92579 different news stories in form of text documents. Every story contain content pursued by numerous sentences as highlights which are utilized as reference synopses for assessment reason. Example of a news story document.



```
(CNN) -- Rory McIlroy is off to a good start at the Scottish Open. He's hopin

McIlroy shot a course record 7-under-par 64 at Royal Aberdeen on Thursday, an

McIlroy carded eight birdies and one bogey in windy, chilly conditions.

"Going out this morning in these conditions I thought anything in the 60s wou

A win Sunday would be the perfect way for former No. 1 McIlroy to prepare for

"Everything was pretty much on," McIlroy said. "I controlled my ball flight r

"I've been working the last 10 days on keeping the ball down, hitting easy sh

Last year Phil Mickelson used the Scottish Open at Castle Stuart as the sprin

Mickelson needs a jolt of confidence given that 'Lefty' has slipped outside t

"I thought it was tough conditions," Mickelson said in an audio interview pos

"I felt like I played well and had a good putting day. It was a good day."

Last year's U.S. Open champion, Justin Rose, was tied for 13th with a 69 but

@highlight

Rory McIlroy shoots a course record at the Scottish Open with a 64

@highlight

The Northern Irishman tallies eight birdies and one bogey at Royal Aberdeen

@highlight

Sweden's Kristoffer Broberg had earlier set the course record on Thursday

@highlight

Defending champion Phil Mickelson is in contention after registering a 68
```

Fig. 3: Example story text

Each story document is stored as an individual text. Therefore, all story documents are stored as a list of texts. Also, every story is separated as the actual story text and highlights. Highlights could then be utilized as a reference summary for the story.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 7, Issue 8 , August 2020

C. Proposed Approach

- Text Preprocessing: The are various text preprocessing steps taken which are explained below in detail. Clean the news articles, remove stop words, data tokenization, stemming the words and so on. This makes the data more concentrated.
  - Data Normalization - The articles are cleaned by re- moving punctuations, numbers, removing extra spaces and all words changed to lower case. These tasks normalizes the data before processing.
  - Tokenization - In an automatic summary generation, we need to fetch candidate sentences for generating summary of the text. Therefore, we need to segment the text into sentences. We have utilized NLTK library [14] to part sentences by perceiving different sentence terminations. These sentences are further segmented into words using the space character of English dictionary as the splitting index.
  - Removing Stopwords - Commonly used words that could be ignored are called stopwords. For example, a, the, as, etc. are common words. This could reduce our vocabulary and time to process them. NLTK library[14] provides a list of English common words that could be removed from the text.
  - Word stemming - It is the way toward changing Each story document is stored as an individual text. Therefore, all story documents are stored as a list of texts. Also, every story is separated as the actual story text and highlights. Highlights could then be utilized as a reference summary for the story.
  - (v) Lemmatization - It reduces words to its lemma or linguistic context which may have a different meaning. After based on the part of speech. For example, the word removing the highlights from the stories and saving them as better could be converted to good. [12].
- Feature Extraction: The topic modeling algorithm Latent Dirichlet Allocation is used for fetching important topics from the text [9]

The ideal number of topics is determined through the coherence score of the model for a specific text. Coherence score of the LDA Model tells how good a topic model is. With an increasing number of topics, coherence score increases and flatten outs but it ends up having redundancy in topics. Therefore, we have picked up the model that gave the highest coherence score before leveling out. Coherence score is calculated using Gensim CoherenceModel library available in Python. However, internally it applies below formula.

$$CoherenceScore = \sum_{i<j} score(w_i, w_j) \quad (1)$$

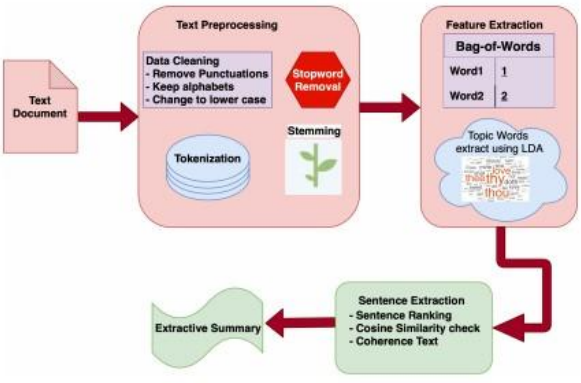


Fig. 4: Workflow

$$Score(wi, wj) = log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (2)$$

where p(w) is the probability of word w_i in the document and $p(w_i,w_j)$ is the co-occurrence of words w_i and w_j in the document.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 7, Issue 8 , August 2020

• Sentence Extraction: Once the topic words are fetched using the LDA algorithm, a sentence distribution is created which gives the score of a sentence for each topic. The score is the length of intersection of topic words and the sentence words. The number of similar words between each sentence and the list of topic words is the score for the sentence. Then the sentences arranged in descending order of this score. Now, before summarizing there is a need to check the redundancy of these sentences. Using cosine similarity, the sentences are further ranked as per their similarity score. Cosine similarity is basically a dot product of two vectors which outputs how similar the two vectors are. The top k sentences from the sentence lists are then picked up with the highest scores for summarization. In this project, the news articles have highlights which are used as reference summaries. The k value is determined through the number of sentences present in highlights.

D. Results and Discussions

We have produced the summary for each document. Assessment of a summary is a troublesome assignment in light of the fact that there is no perfect summary for a document and the meaning of a decent summary is an open ended question to a huge degree.

The two most significant elements based on which the generated summary is to be evaluated is the quality of the summary and how much it's likeness with the reference summary.

The Table 1 shows few example summaries of documents and their Recall scores. It was observed that the generated summary is similar to actual summary. Also, it is an extract of actual document so the sentence length is larger. For some cases, the generated summary is quite different from the actual summary but the quality is good. On an average, the score is 0.5 if we compared with the actual summary. We have generated summaries of various lengths and compared with the actual summary. The Table 2 gives the compression ratio, readability grade and recall score for a summary with different length. We can infer that if we increase the length of the summary, the readability grade decreases (the summary becomes easier to read), the recall score increases (the similarity between reference summary and system generated summary increases). Flesch-Kincaid Readability Tests provided the English readability grade of a text[11]. It measures how fluent or readable is the language of a text. The Flesch-Kincaid Grade tells the number of years a person needs to study to understand the text. It represents the grade as US grade level. The Table 3 gives the various grade levels and its difficulty levels. We have aimed of grade level 8-9.

Compression Ratio is given by the length of generated summary to the length of original text. It is a measure which tells how much shorter the summary is with respect to reference summary. There is a trade-off between the compression ratio and the quality of the summary. We can observe that in the event that we endeavor to get a decent quality of summary, at that point the compression ratio increments and in the event that we decline the compression ratio, at that point quality is at a hurl. There is an improvement required to achieve good compression ratio with good quality summary

IV. FUTURE WORK

Automatic document summarization is a challenging task in the current world where trends are towards biomedical, emails, blogs, internet, education and so on. There is a huge influx of information and increasing day by day. Automatically summarizing these information is of great importance and a need. Document Summarization has turned into a significant research in Natural Language Processing (NLP) and Big Data arenas. The extractive summarization using topic modeling LDA algorithm successfully generates a summary of important sentences from the original document. It also provides good level of topic diversity. Later on, we might want to investigate progressively target works and improve the summary generation further and utilize diverse topic modeling techniques. Likewise, we mean to assess our way to deal with various dialects. There is a future scope of generating abstractive summaries which are more human like summaries and will require heavy machine learning tools for semantic language generation.

REFERENCES

[1] R. Hafeez, S. Khan, M. A. Abbas, and F. Maqbool. Topic based summarization of multiple documents using semantic analysis and clustering. In 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT IoT (HONET-ICT), pages 70–74, Oct 2018.

[2] S. Babar, M. Tech-Cse, and R. . Text summarization:an overview. 10 2013.

[3] Z. Wu, L. Lei, G. Li, E. Chen, H. Huang, G. Xu, and C. Zheng. A topic modeling based approach to novel document automatic summarization. Expert Systems with Applications, 84:12–23, 05 2017.

[4] G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. A semantic concordance. In Proceedings of the Workshop on Human Language Technology, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 7, Issue 8 , August 2020

[5] A. Kazantseva and S. Szpakowicz. Summarizing short stories. Comput. Linguist., 36(1):71–109, Mar. 2010.
[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, Mar. 2003.
[7] J. Liu. Image retrieval based on bag-of-words model. 04 2013.
[8] N. Nagwani. Summarizing large text collection using topic modeling and clustering based on mapreduce framework. Journal of Big Data, 2,
[9] D. M. Blei. Probabilistic topic models. Commun. ACM, 55(4):77–84, Apr. 2012.
[10] D. M. Blei. Probabilistic topic models. Commun. ACM, 55(4):77–84, Apr. 2012.
[11] D. M. Blei. Probabilistic topic models. Commun. ACM, 55(4):77–84, Apr. 2012.
[12] M. Honnibal and I. Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 7, Issue 8 , August 2020

TABLE I
PERFORMANCE SCORE OF LDA MODEL

Doc Id	Actual Summary	Generated Summary	Recall Score
Story 1	Math geeks and others celebrate Pi Day every March 14. Pi, or roughly 3.14, is the ratio of circumference to diameter of a circle. The Pi Day holiday idea started at the Exploratorium museum in San Francisco. Albert Einstein was also born on March 14	March 14 is my favorite day to be a nerd. What's more, Albert Einstein was born on this day. Across the country, math geeks in museums, schools, private groups and elsewhere gather to celebrate the number pi, approximately 3.14. That's why March 14 – 3-14 – is Pi Day.	0.5
Story 2	MH370 families hold sit-in outside the Malaysian Embassy in Beijing. Relatives marched from their hotel after request to meet Malaysian ambassador failed. More than once in recent weeks Malaysian authorities have not shown up for talks with relatives. NEW: China appeals to protesters to express concerns in "legal and rational way"	The small gathering departed after delivering a letter of protest to the embassy demanding an explanation for how the prime minister of Malaysia has concluded that the plane ended its flight over the Indian Ocean. "The ambassador kept saying he would come but he never showed," an elderly man told CNN as he and scores of others walked quietly through Beijing's streets. On Friday a spokesperson for the Chinese foreign ministry, Qin Gang, called on the families protesting outside the embassy to "express their appeals in a legal and rational way" while pledging to give them further assistance. Families have organized committees, issued press releases, and printed t-shirts and hats with the slogan "Pray for MH370." But they have also been largely prevented from organizing public displays outside the windowless hotel conference room that serves as the families' improvised headquarters.	0.4
Story 3	Protesters hold up Japanese flags and chant slogans against China. Beijing says the Diaoyu Islands belong to China. In Japan, the islands are known as the Senkaku	Beijing and Tokyo have been clashing over the arrest of a Chinese fishing captain by Japan off the disputed islands. (CNN) – Anti-China protesters gathered Saturday in Tokyo and six other major cities in Japan to rally against what it calls an invasion of disputed islands that both claim are part of their territories. Protesters held up Japanese flags and chanted, "We will not allow Communist China to invade our territory." Tamogami called China a "thief" and vowed to protect the islands.	0.5
Story 4	Web sites of the House of Representatives are overwhelmed with e-mails. Administrators implement the "digital version of a traffic cop" to handle the overload. "This is unprecedented," says a House spokesman. Overload began Sunday as legislators said bailout agreement was posted online	WASHINGTON (CNN),– The servers hosting the Web sites of the House of Representatives and its members have been overwhelmed with millions of e-mails in the past few days, forcing administrators to implement the "digital version of a traffic cop" to handle the overload. As millions of people tried to look at the details of the bailout plan, the House.gov system became overwhelmed and many people saw notices on their computer screens saying "this page does not appear." Ventura said the House.gov Web site experienced a very high number of hits when the 9/11 commission released its final report on the September 11, 2001, terror attacks against the United States, but nothing like what the site has seen in the past few days. "We know it's in the millions," he said of the number of e-mails that lawmakers in the House have been receiving. Now, when House.gov or individual members' sites begin to get overloaded, a message will come up on the computer screen saying, in effect, "try back later," Ventura said.	0.5555
Story 5	Arsenal striker Eduardo is ruled out for two weeks due to a hamstring problem. The Croatian was hurt on Monday in his first game back after a year's absence. Eduardo scored twice against Cardiff on his return to action from a broken leg	(CNN) – Arsenal striker Eduardo has been ruled out for two weeks with a hamstring injury just days after returning from a broken leg. Eduardo sinks to his knees after opening the scoring on his Arsenal first team comeback on Monday. He picked up a hamstring injury two minutes before I took him off. The 25-year-old marked his return with two goals in the 4-0 success at the Emirates Stadium, but his latest setback is not described as serious	0.6216

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 7, Issue 8 , August 2020

TABLE II
EVALUATION OF DIFFERENT SUMMARY LENGTH

Summary Length	Recall Score	Flesch-Kincaid Readability Grade	Compression Ratio
200 Words	0.378	25	0.05
500 Words	0.405	14	0.13
Number of sentences same as reference	0.612	8	0.17

TABLE III
FLESCH-KINCAID READABILITY GRADE LEVEL TABLE

Grade Level	Reading Difficulty
US Grade 5	Very Easy
US Grade 6	Easy
US Grade 7	Fairly Easy
US Grade 8-9	Standard
US Grade 10-12	Fairly Difficult
US Grade 13-16	Difficult
US Grade 16 and above	Very Difficult