MULTI-DOCUMENT SUMMARIZATION USING SENTENCE CLUSTERING

[1]Akash Kamble, [2]Vivek Chaudhary, [3]Adesh Dhakane, [4]Abhijit Ghorpade

[1]Student, [2]Student, [3]Student, [4]Student

[1]Department of Computer Engineering,

[1]Sinhgad Academy of Engineering, Kondhwa(Bk.), Pune, India

Abstract: Multi-document summarization is used for generating the summary of the documents which will provide the central idea about the documents in short. As large amount of information is available on internet, there is a need of system that can provide informative summary in a short time period. In this paper, we have proposed a system generating the summary of multiple documents related to a particular topic as well as it will show summary of particular document. In our proposed system, we use tokenization and stopword removal method as a preprocessing step of latent dirichlet allocation (LDA) model. Our approach also uses sentence clustering and term frequency algorithm for generating short text summary.

Keywords – Summary, Stop-word removal, Clustering, Term frequency

1. INTRODUCTION

As large number of documents are available on the internet, it seems difficult to get the required information related to a particular topic. In order to solve this problem, we have proposed a multiple documents summarization using sentence clustering. Using this summarization system, user gets the short informative summary related to a particular topic from multiple documents. Instead of reading number of documents for particular topic, user can make use of this system to get topic related information. It will save the reading time of user and user will get the lot of topic related information in short time period. This system generates the summary of multiple documents of .txt and .pdf file formats. Also user gets the summary for multiple topics if all that topics present in a particular document. System also provides the documents related to a searched topic among the collection of documents before generating the summary. For generating summary, we use following steps:-

- Firstly we use tokenization and stop-word removal method as a preprocessing step of LDA algorithm. It is able to remove the useless data such as prepositions, connectives, etc. to remove the irrelevant information and extract topic from documents.
- Considering topic related search by the user, we have used term frequency algorithm to count multiple occurrence of words in a document and then accordingly calculate the relevant score of each document for arranging them as per their relevant score.
- By using sentence level clustering, clusters of sentences are formed from given documents. It makes the clusters of sentences which are relatively similar based on a particular word.

2. OBJECTIVES

1. To summarize multiple documents into a short informative summary related to a particular topic and save the time of users.
2. To generate summary of the multiple documents that can be easily readable and understandable by the user.
3. To generate the topic related summary of documents supporting more than one file formats.
4. To provide the documents related to particular topic to the user.

3. RELATED WORK

Wei Li(2010) was proposed a summarization system in which BSU semantic link network is used to generate the summaries. The approach used in this paper was very effective for extraction of information and providing good summaries. Baotian Hu(2015) had written a paper on a text summarization in which summary of Chinese text is generated using recurrent neural network. Jason Weston(2015) was proposed a model for the summarization of sentences and provides a short summary.

4. SYSTEM ARCHITECTURE

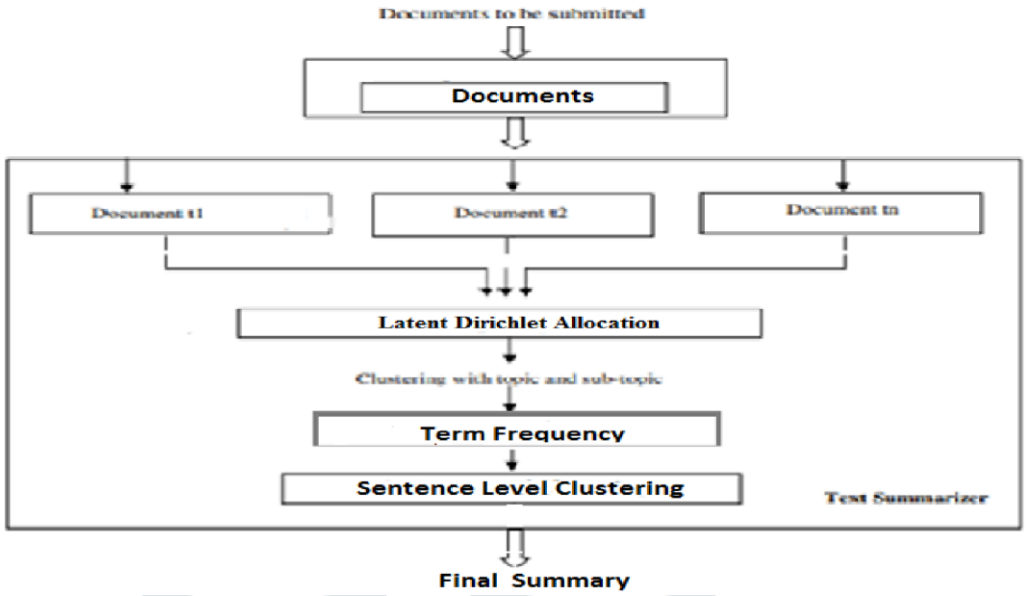


Figure 4.1

5. SYSTEM ALGORITHM

5.1 Latent Dirichlet Allocation(LDA)

• Latent dirichlet allocation is one of the widely used topic modeling technique. Topic modeling is used to discover the abstract topics which are in a set of documents. In data preprocessing steps of LDA, we use tokenization and stop word removal method for the keyword extraction.

1. Tokenization :- It splits the text data into sentences and generated sentences into words. Also it converts uppercase letters to lowercase and remove punctuations.
2. Stop-word removal method :- The process of removing stop-words helps to save the time and reduces the computation. For extraction of keyword, we remove the stop-words by comparing them with the words in stop-words list given

5.2 .Term Frequency

- Term frequency is used to count multiple occurrences of words which are occurred in each document.
- Most important words related to document are taken into account and the count particular term depends on number of presence of words in a specific document.
- Based on this count, documents containing a particular topic are arranged in a descending order of their relevant score as follows :-

  Relevant score:- No of matched words ⁄ Total no of words in a document

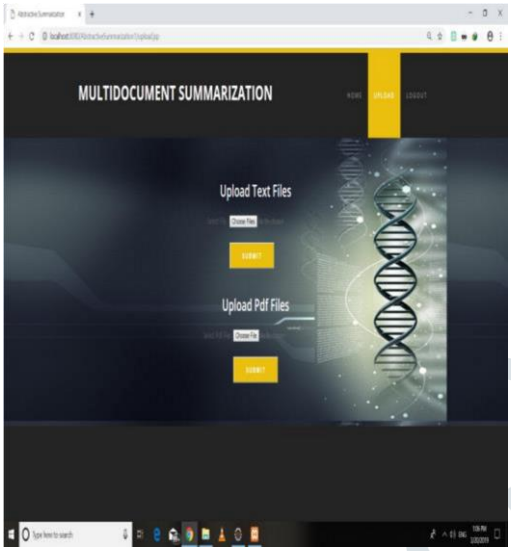
5.3 Sentence Level Clustering

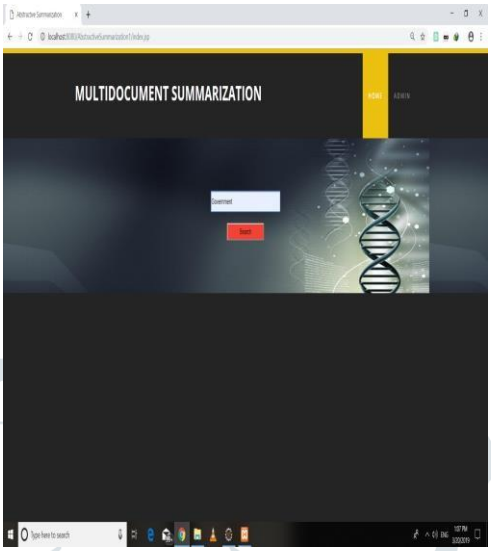
- In this algorithm, representative documents and sentences including topics are grouped into multiple clusters using sentence level clustering.
- Sentence clustering avoids the redundancy of sentences. It is domain and language independent.
- Documents containing a given topic are only considered for the summarization. Then it makes cluster of sentences based on a given topic and generates the topic related summary.
- This summary is informative and readable which will save overall time of user

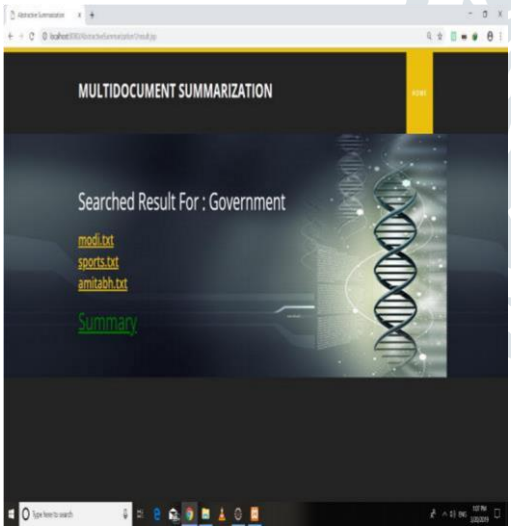
Figure 5.3.1

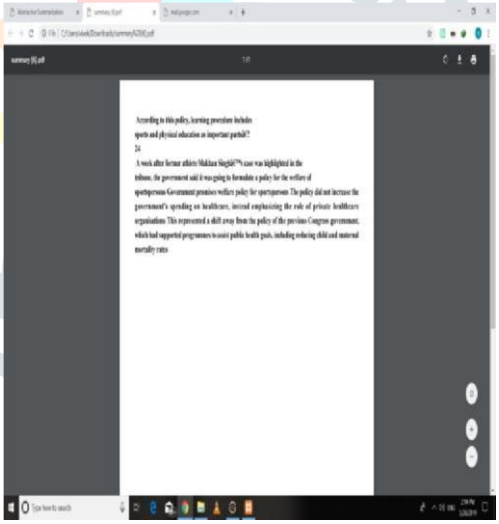
6. RESULT ANALYSIS

In our summarization system, it generates the summary of .txt and .pdf files. So first of all, we upload .txt and .pdf files and submit to the system. Then the users enter the topic of their interest and click on "Search" button. After searching for particular topic by the system, it displays only that documents which include a topic searched by user and if topic related information does not found in uploaded documents, then system displays "No Results" to users. Finally when users click on "Summary" button, system generates informative summary related to the topic given by user.









7. CONCLUSION AND FUTURE WORK

Multi-document summarization using sentence clustering creates a short summary of multiple documents. Generated summary is easy to read and understand than existing systems. It provides important and informative sentences in the summary and removes repetitive sentences for saving user's time. Our system generates the summary which contains the central ideas of given documents and provides the better approach that how the multiple documents of file formats such as .txt, .pdf can be summarized into short one related to a particular topic. This summary will provide useful information in a short time period.

In future work, we plan to make system which can provide the summary of web pages and multiple file formats such as .exe, .html, .ps, etc. We will integrate and implement more algorithms related to various summarization approaches with our system to make our system that can provide the summary of multiple images and videos.

8. REFERENCES

[1] G. Carenini and J. C. K. Cheung, "Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality," in Proceedings of the Fifth International Natural Language Generation Conference. Association for Computational Linguistics, 2008, pp. 33– 41.

[2] T. Mikolov, M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur, "Recurrent neural network based language model." in Interspeech, vol. 2, 2010, p. 3. DOI: https://dx.doi.org/10.26808/rs.ca.i8v6.07 International Journal of Computer Application (2250-1797) Issue 8 Volume 6, Nov.- Dec. 2018 55

[3] K. Filippova, "Multi-sentence compression: finding shortest paths in word graphs," in Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010, pp. 322–330.

[4] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out: Proceedings of the ACL-04 workshop. Barcelona, Spain, 2004.

[5] S. Banerjee, P. Mitra, and K. Sugiyama, "Multi-document abstractive summarization using ilp based multi-sentence compression," in Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press, 2015, pp. 1208–1214.

[6] W. Li, "Abstractive multi-document summarization with semantic information extraction," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1908–1913.

[7] B. Hu, Q. Chen, and F. Zhu, "Lcsts: A large scale chinese short text summarization dataset," arXiv preprint arXiv:1506.05865, 2015.

[8] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," arXiv preprint arXiv:1509.00685, 2015.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[10] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to sequence learning," arXiv preprint arXiv:1603.06393, 2016.

[11] S. Bird, "Nltk: the natural language toolkit," in Proceedings of the COLING/ACL on

Interactive presentation sessions. Association for Computational Linguistics, 2006, pp. 69– 72.

[12] K. Toutanova, D. Klein, C. Manning et al., "Stanford core nlp," The Stanford Natural Language Processing Group. Available: http://nlp. stanford. edu/software/corenlp. shtml. Accessed, 2013.