Michael, J Comput Eng Inf Technol 2020, 9:3
DOI: 10.37532/jceit.2020.9(3).225

Journal of Computer Engineering & Information Technology

Short Communication

A SCITECHNOL JOURNAL

A Brief Note on Document Summarization

Abhya Tripathi*

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 project, 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 andfetchesthesummaryforthenewsarticle. Thishelpsinsaving timetoreadlongtextsordocuments.

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 documentissegmentedinalistofsentencesandappliedto 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 futurescope.

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 ofimportant 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.

Keywords: Document Summarization; Topic Modeling; Latent Dirichlet Allocation; Extractive Summarization; Abstractive Summarization.

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 hugedegree.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 evaluation will be done using the Recall-Oriented Understudy for Gisting Evaluation (Rogue) score which is used to check the similarity of the generated summary with the reference.

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. The 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 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 be-comes easier to read), the recall score increases (the similarity between reference summary and system generated summary.

Citation: *Tripathi A (2020) A Brief Note on Document Summarization. J Comput Eng Inf Technol 9:3.*

***Corresponding author:** Abhya Tripathi, Department of Computer Science Usha Mittal Institute of Technology, Mumbai, India, E-mail: sonali.bk86@gmail.com

Received: May 19, 2020 **Accepted:** July 24, 2020 **Published:** July 31, 2020

Author Affiliation Top

Department of Computer Science Usha Mittal Institute Of Technology,Mumbai, India