Indian Institute of Technology, Kanpur
Department of Computer Science and Engineering

CS657A: Information Retrieval
Project Report

Query Based Summarization of Ramayana

Under the guidance of
Prof. Arnab Bhattacharya
Academic Year 2021 - 2022

Group 17

Abhishek Dnyaneshwar Revskar (21111003)
(abhishekdr21@iitk.ac.in)

Akash Gajanan Panzade (21111006)
(akashp21@iitk.ac.in)

Deepak Kumar (21111023)
(dkumar21@iitk.ac.in)

Mayuresh Diwakar Shandilya (21111041)
(mayuresh21@iitk.ac.in)

Prajwal Pradiprao Thakare (21111047)
(prajwalt21@iitk.ac.in)

Acknowledgement

We would like to extend our sincere gratitude to our Project Guide, Prof. Arnab Bhattacharya for his advice, guidance, patience and timely help during the project. The freedom he gave us, to work with anything at any time, encouraged us to try out new things and helped to work more efficiently. It is a great honor for us to have been a part of his course. The successful completion of the work has been only possible due to his excellent guidance, meticulous observation and critical analysis.

Abhishek Dnyaneshwar Revskar (21111003)
Akash Gajanan Panzade (21111006)
Deepak Kumar (21111023)
Mayuresh Diwakar Shandilya (21111041)
Prajwal Pradiprao Thakare (21111047)

Contents

1 Abstract

Summarization is a technique that is commonly used by several websites and applications to create a short news feed and article summaries. It is very essential to give only important text that conveys the meaning of the long articles. User prefers to read a short summary to get the insight of the entire article rather than going through it completely. The main aim of our project is to generate a summary of an event in Ramayana based on the query given by the user. We have a dataset which contains the topics in Ramayana which are divided into different documents. Each document is atomic and self contained in nature. We retrieve only the most relevant document to the given query and then extract the important sentences from it by using some similarity measures between the query and the sentences in this document to create the summary.

2 Introduction

Search-engines have now become increasingly popular, and have become the common entry-ways for people into the world wide web. With the ever-increasing amount of information available on the internet, and the excessively busy schedules of people, individuals do not have time to read the complete document(s) that is retrieved for their query. What they need is a concise and precise summarisation of the information available in that document. The relevance of this document to the query is judged based on this summary only. Even a extremely relevant document may be marked as irrelevant due to poor summary generated. This leads to the growing importance of document summarization in modern information retrieval systems.

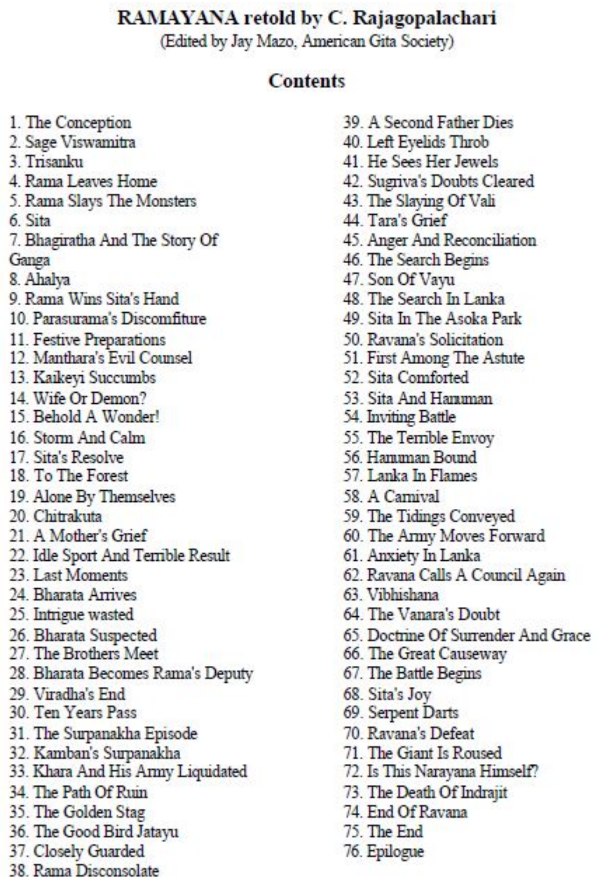
Summarization in general comes in two flavours, namely Extractive and Abstractive. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express. Such a summary might include verbal innovations.

Traditional information retrieval systems return a ranked list of whole documents as the answer to a query. However, in many cases, not every part of an entire document is relevant to the query. Thus, it is desirable to retrieve only relevant passages, as opposed to whole documents, which in effect helps to filter out irrelevant information in a long relevant document.

We are performing extractive summarization of the Ramayana based on the query given by the user. All the existing techniques for abstractive summarization have used daily news, blogs, Wikipedia articles etc. as the pre-training dataset hence they cannot produce a good summary if we give this Ramayana dataset to produce abstractive summarization. Google's PEGASUS a is model which is pre-trained on a large corpus of documents to produce abstractive summaries. We can fine tune this model to produce the summaries for the Ramayana dataset by giving 1000 examples of the text and their corresponding summaries. But this task is very daunting and requires a lot of time and efforts hence we decided to produce extractive summaries which is a relatively easy task.

3 Dataset

We have used a book written on Ramayana in English by C. Rajagopalachari. It contains the topics in Ramayana divided into different chapters. We have taken these chapters and created different documents from them. It gave us 75 documents representing 75 different topics from the book.



RAMAYANA retold by C. Rajagopalachari
(Edited by Jay Mazo, American Gita Society)

Contents

4 Methodology

To generate the summaries, we have used 4 different methods :

1. GLoVe Word Embeddings.
2. FastText Word Embeddings and Paraphrasing.
3. Sentence-BERT Sentence Embeddings.
4. Clustering of Sentences.

4.1 GLoVe Word Embeddings

In this method, we have used the GLoVe word embeddings having 100 dimensions to encode the sentences and the query. The detailed process of generating summaries is as follows:

1. Get the query from the user and retrieve the most relevant document from the corpus. We are using the Okapi BM25 document retrieval model for this task.
2. After getting this document, we are calculating the similarity between the query vector and all the sentences in this document. We are using two similarity measures here viz. Cosine similarity and the Euclidean distance. To get the vectors for the query and the sentences, we are finding the word embeddings for all the words in the sentences and the query and then taking their average to get the final vector.
3. We are finding the top 10 most similar sentences with the query and then applying the text rank algorithm on these sentences.
4. We are selecting the top 5 sentences after applying this algorithm and then arranging them in the order they appear in the document.

Some of the points in the above process are the same for some of the other methods that we have used.

4.2 FastText Word Embeddings and Paraphrasing

In this method we are using fasttext to encode the query and the sentences. All the steps are same in this method as in the above method except that we paraphrase the sentences that we get in the final step.

We are using Google's PEGASUS model to paraphrase the sentences. The main motive to apply paraphrasing is to go beyond the extractive summarization and getting the results somewhat close to those of abstractive summarization.

4.3 Sentence-BERT Sentence Embeddings

In this model we are using sentence BERT model to encode the sentences and the query. All the steps are same as in the first method. We are using this model since it captures the context of the words to find the vector for the given sentence.

BERT is trained using a denoising objective (masked language modeling), where it aims to reconstruct a noisy version of a sentence back into its original version. The concept is similar to autoencoders. Older systems like Word2vec and Glove had poorer performance because their word embeddings didn't dynamically change based on the context of the surrounding vector. In other words, they were fixed.

4.4 Clustering of Sentences

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

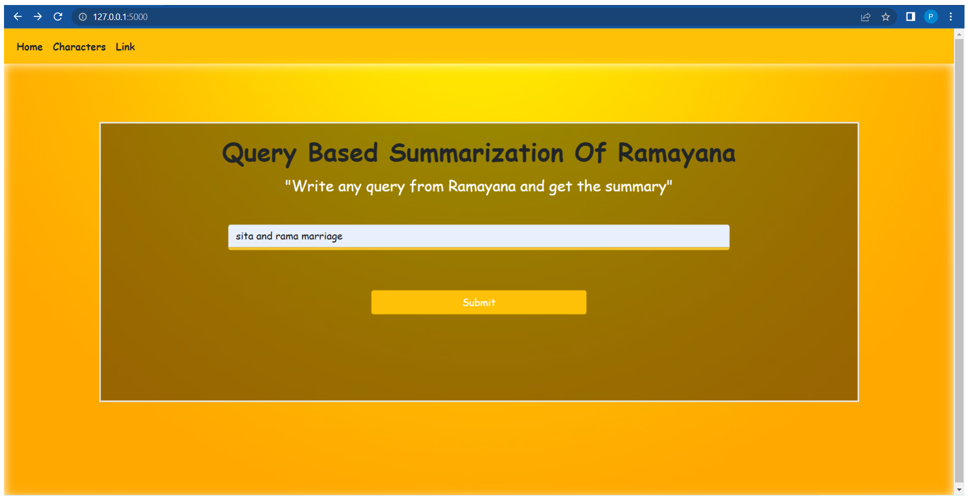
After getting the most relevant document for the given query, we have used the clustering technique to cluster the similar sentences from the retrieved document into one cluster. We have then selected the cluster that is closest to the query vector. All the sentences that are present in this cluster will act as the summary for the given query. We have used the GLoVe word embeddings to find the query vector and the sentence vector. We have not used the page rank algorithm in this method.

5 Results and Analysis

5.1 Results

We have created a User Interface in the form of a website which will accept the query from the user and then generate the summaries for all the 4 models that we have implemented.

The home page which accepts the query from the user is given below.

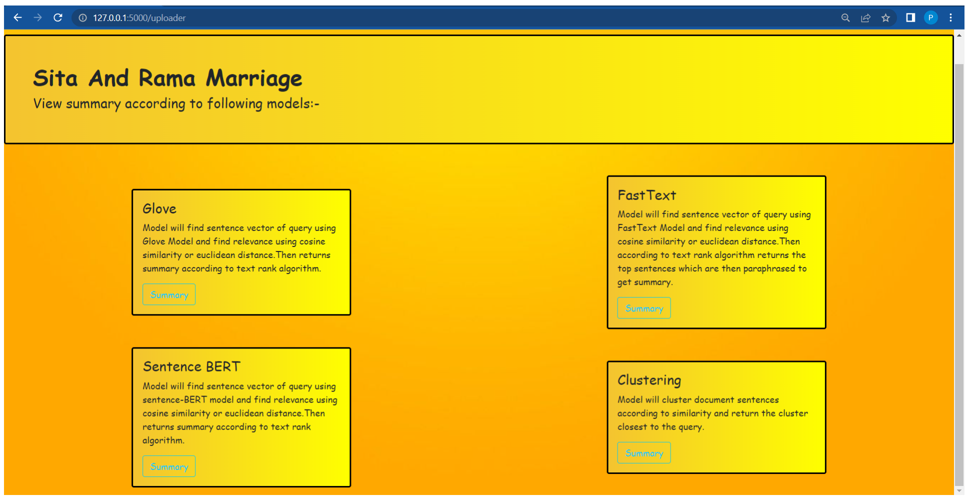


Query Based Summarization Of Ramayana

"Write any query from Ramayana and get the summary"

sita and rama marriage

We can see that the query entered by the user in the above page is "sita and rama marriage". After clicking on the "Submit" button below the query box, we get the following interface which contains all the 4 models with the description of how do they generate the summaries.



Sita And Rama Marriage

View summary according to following models:-

Glove

Model will find sentence vector of query using Glove Model and find relevance using cosine similarity or euclidean distance.Then returns summary according to text rank algorithm.

FastText

Model will find sentence vector of query using FastText Model and find relevance using cosine similarity or euclidean distance.Then according to text rank algorithm returns the top sentences which are then paraphrased to get summary.

Sentence BERT

Model will find sentence vector of query using sentence-BERT model and find relevance using cosine similarity or euclidean distance.Then returns summary according to text rank algorithm.

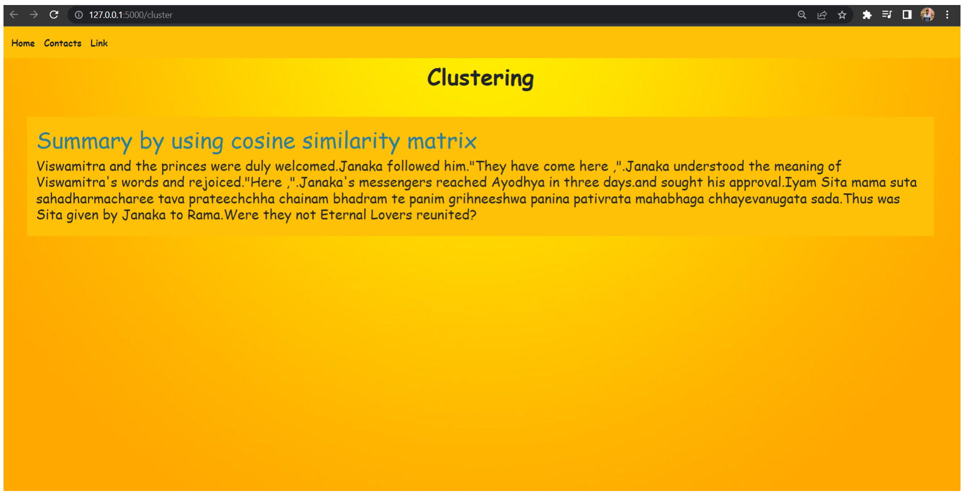
Clustering

Model will cluster document sentences according to similarity and return the cluster closest to the query.

Each model above has a button "Summary" associated with it, which after clicking gives the user the summaries for the query that the user have given. All the models above except the "Clustering" give us 2 summaries for the same query, one for cosine similarity and the other for euclidean distance measure.

Below are the summaries generated by all the 4 models for the given query.

- Clustering

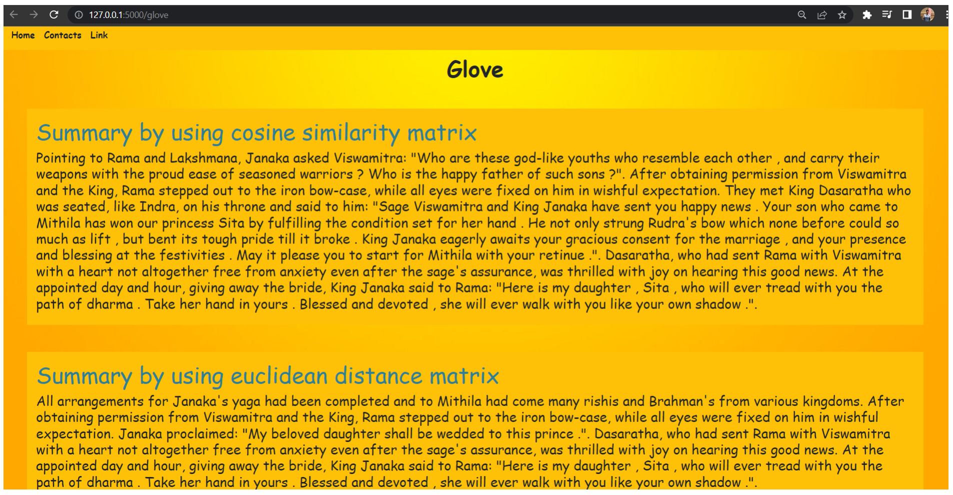


Clustering

Summary by using cosine similarity matrix

Viswamitra and the princes were duly welcomed.Janaka followed him."They have come here ,".Janaka understood the meaning of Viswamitra's words and rejoiced."Here ,".Janaka's messengers reached Ayodhya in three days.and sought his approval.Iyam Sita mama suta sahadharmacharee tava prateechchha chainam bhadram te panim grihneeshwa panina pativrata mahabhaga chhayevanugata sada.Thus was Sita given by Janaka to Rama.Were they not Eternal Lovers reunited?

- GLoVe



Glove

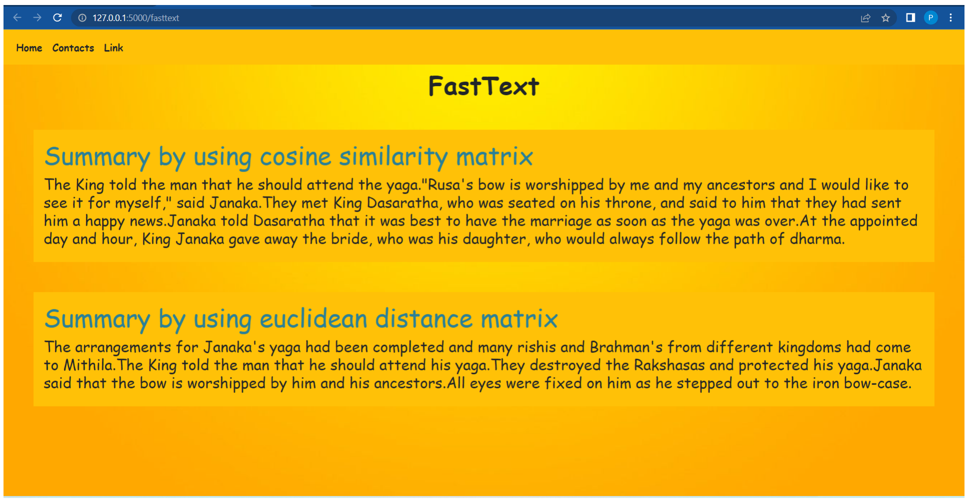
Summary by using cosine similarity matrix

Pointing to Rama and Lakshmana, Janaka asked Viswamitra: "Who are these god-like youths who resemble each other , and carry their weapons with the proud ease of seasoned warriors ? Who is the happy father of such sons ?". After obtaining permission from Viswamitra and the King, Rama stepped out to the iron bow-case, while all eyes were fixed on him in wishful expectation. They met King Dasaratha who was seated, like Indra, on his throne and said to him: "Sage Viswamitra and King Janaka have sent you happy news . Your son who came to Mithila has won our princess Sita by fulfilling the condition set for her hand . He not only strung Rudra's bow which none before could so much as lift , but bent its tough pride till it broke . King Janaka eagerly awaits your gracious consent for the marriage , and your presence and blessing at the festivities . May it please you to start for Mithila with your retinue .". Dasaratha, who had sent Rama with Viswamitra with a heart not altogether free from anxiety even after the sage's assurance, was thrilled with joy on hearing this good news. At the appointed day and hour, giving away the bride, King Janaka said to Rama: "Here is my daughter , Sita , who will ever tread with you the path of dharma . Take her hand in yours . Blessed and devoted , she will ever walk with you like your own shadow .".

Summary by using euclidean distance matrix

All arrangements for Janaka's yaga had been completed and to Mithila had come many rishis and Brahman's from various kingdoms. After obtaining permission from Viswamitra and the King, Rama stepped out to the iron bow-case, while all eyes were fixed on him in wishful expectation. Janaka proclaimed: "My beloved daughter shall be wedded to this prince .". Dasaratha, who had sent Rama with Viswamitra with a heart not altogether free from anxiety even after the sage's assurance, was thrilled with joy on hearing this good news. At the appointed day and hour, giving away the bride, King Janaka said to Rama: "Here is my daughter , Sita , who will ever tread with you the path of dharma . Take her hand in yours . Blessed and devoted , she will ever walk with you like your own shadow .".

- FastText



FastText

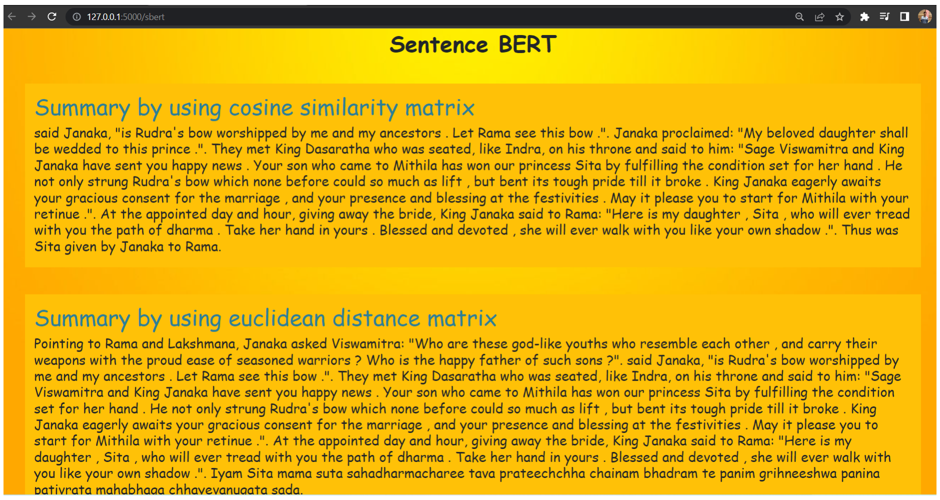
Summary by using cosine similarity matrix

The King told the man that he should attend the yaga."Rusa's bow is worshipped by me and my ancestors and I would like to see it for myself," said Janaka.They met King Dasaratha, who was seated on his throne, and said to him that they had sent him a happy news.Janaka told Dasaratha that it was best to have the marriage as soon as the yaga was over.At the appointed day and hour, King Janaka gave away the bride, who was his daughter, who would always follow the path of dharma.

Summary by using euclidean distance matrix

The arrangements for Janaka's yaga had been completed and many rishis and Brahman's from different kingdoms had come to Mithila.The King told the man that he should attend his yaga.They destroyed the Rakshasas and protected his yaga.Janaka said that the bow is worshipped by him and his ancestors.All eyes were fixed on him as he stepped out to the iron bow-case.

- Sentence BERT



Sentence BERT

Summary by using cosine similarity matrix

said Janaka, "is Rudra's bow worshipped by me and my ancestors . Let Rama see this bow .". Janaka proclaimed: "My beloved daughter shall be wedded to this prince .". They met King Dasaratha who was seated, like Indra, on his throne and said to him: "Sage Viswamitra and King Janaka have sent you happy news . Your son who came to Mithila has won our princess Sita by fulfilling the condition set for her hand . He not only strung Rudra's bow which none before could so much as lift , but bent its tough pride till it broke . King Janaka eagerly awaits your gracious consent for the marriage , and your presence and blessing at the festivities . May it please you to start for Mithila with your retinue .". At the appointed day and hour, giving away the bride, King Janaka said to Rama: "Here is my daughter , Sita , who will ever tread with you the path of dharma . Take her hand in yours . Blessed and devoted , she will ever walk with you like your own shadow .". Thus was Sita given by Janaka to Rama.

Summary by using euclidean distance matrix

Pointing to Rama and Lakshmana, Janaka asked Viswamitra: "Who are these god-like youths who resemble each other , and carry their weapons with the proud ease of seasoned warriors ? Who is the happy father of such sons ?". said Janaka, "is Rudra's bow worshipped by me and my ancestors . Let Rama see this bow .". They met King Dasaratha who was seated, like Indra, on his throne and said to him: "Sage Viswamitra and King Janaka have sent you happy news . Your son who came to Mithila has won our princess Sita by fulfilling the condition set for her hand . He not only strung Rudra's bow which none before could so much as lift , but bent its tough pride till it broke . King Janaka eagerly awaits your gracious consent for the marriage , and your presence and blessing at the festivities . May it please you to start for Mithila with your retinue .". At the appointed day and hour, giving away the bride, King Janaka said to Rama: "Here is my daughter , Sita , who will ever tread with you the path of dharma . Take her hand in yours . Blessed and devoted , she will ever walk with you like your own shadow .". Iyam Sita mama suta sahadharmacharee tava prateechchha chainam bhadram te panim grihneeshwa panina pativrata mahabhaga chhayevanugata sada.

5.2 Analysis

As we are performing extractive summarization for some of the models, the quality of summary generated depends largely on the dataset used.The Clustering model is basically an unsupervised machine learning model which creates clusters based on the similarity of the sentences without considering their context and meaning. We select the cluster whose mean is closest to the query vector and all the sentences in this cluster act as the summary. The problem with this approach is that it selects the summary for the given query based on the means of the clusters and not the sentences themselves which was the case in the GLoVe model. Due to this, the sentences which are close to the query vector but are in a cluster whose mean is not the closest one to the query are ignored in generating the summary and all the sentences in the cluster whose mean is closest to the query are selected for the summary. This cluster might contain the sentences which are far away from query vector.

In GLoVe implementation, we are getting a summary which contains the sentences that conveys the meaning of the entire document pretty well. The sentences contained in the summary are a subset of the document retrieved for the query. As we know that the GLoVe model for word embeddings doesn't capture the context of the sentence, it sometimes can give a summary which might not be that much accurate for the given query.

The FastText model is implemented in the same manner as the GLoVe model, the only difference is that we are applying the paraphrasing on the generated results. Due to this, the final result that we get conveys the same meaning but in different words. Hence, we can generate a summary which is one step beyond what the extractive summarization might give but it is not fully abstractive. FastText model has an advantage over other word embedding models in the way that it gives the word embeddings based on the character n-gram embeddings and hence it can handle out-of-vocabulary words.

The Sentence BERT model gives the most relevant summary because it generates the sentence embeddings which captures the context of the words in the sentences. This feature is not there in the earlier word embedding models like GLoVe or FastText because their word embeddings didn't dynamically change based on the context of the surrounding vector i.e. they were fixed.

6 Future Work

There is a scope of performing abstractive summarization on this dataset by fine-tuning some pre-trained models like Google's PEGASUS - Large which requires around 1000 training examples to achieve state-of-the-art results. Abstractive summarization captures the context and generates the summaries after applying natural language generation techniques. The summaries thus generated can be very close to what a human might express.

7 References

1. https://en.m.wikipedia.org/wiki/Okapi_BM25
2. https://lifeintegrity.com/Ramayana.pdf
3. https://aclanthology.org/W04-3252.pdf
4. https://www.sbert.net/
5. https://nlp.stanford.edu/projects/glove/
6. https://fasttext.cc/docs/en/crawl-vectors.html
7. https://huggingface.co/tuner007/pegasus_paraphrase
8. https://nlp.stanford.edu/IR-book/information-retrieval-book.html