Bangla Informative Chatbot using TF-IDF and Cosine Similarity on Newspaper data

Shagoto Rahman , ID:170210

Abstract

Chatbot is a basic application that will reply on our messages just like humans. In accord with the query chatbot finds the most particular answer. In recent times, there are lots of research are going on regarding chatbots. Many applications offer humanlike chatting services where both proper hardcoding and learning procedures are used.In this report I have created a chatbot on Bengali language which is quite informative in nature as it is based on Bangla newspaper articles.The application involves different queries from users and find the perfect contemporary for that query using some methodologies like tf-idf, cosine similarity.

Methodology

We have used two methods.They are discussed below.

Dataset

For dataset, I have used the articles from famous Bengali newspaper “The Daily Ittefaq”.Almost 54K documents have been retrieved from the website of the newspaper for the years 2019 and 2020.Almost 9,00,000 sentences are present in the corpus.

A. Preprocessing

Preprocessing is very important in terms of raw data is concerned and especially in Bengali language where different punctuation marks are there.

a. Punctuation Removal

First of all, I have ensured the proper cleaning of each document by cleaning the punctuation marks. Again many Unicode were not detected, so they were needed to be cleaned as well.

b. Duplicate Sentence Removal

There were many duplicate sentences in the corpus as well. So I have made sure that they unique sentences will only be there clearing the duplicate ones.

c. Stop words Removal

There are many stop words in Bengali languages like "ও" , "এবং" ,"আর" etc.These stopwords are identified by computing frequency based unigrams.The most frequent words were the stop words so many stopwords were identified from the frequency based unigrams and from Bengali language point of view.

d .Stemming

There are several words in different formations like শহর, শহরে,শহরের etc. They basically denote the same meaning শহর but with the inclusion of terms like 's', 'es' it shows the different form. To remove redundancy, the conversion of several formations into an actual word is also needed.

e. Tokenization

Word tokenization is done here for each document to compute the tf-idfs of each word of each document which is described in the coming sections.

B. Sentence ranking with cosine similarity with tfidfs

As I have the tokens for each document, So the tf-idf of each word of each document will be calculated.

a.TF-IDF

TF-IDF is a measurement that symbolizes how important a word is to a document in a association of documents. Two metrics are multiplied in this regard. They are the term frequency and the inverse document frequency. Term frequency is how many times a word appears in a document divided by the length of the document in terms of word count and the inverse document frequency is the number of total documents divided by the number of documents that the word has appeared in.

$$TF*IDF= \frac{No\ of\ perticualr\ term\ in\ that\ doccument}{Total\ no\ nof\ term\ in\ that\ doccument} * \frac{Total\ No\ of\ doccument}{No\ of\ doccument\ where\ the\ perticular\ word\ appears}$$

…….(1)

Fig: TF-IDF

b. Cosine Similarity

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

…….(2)

C. Sentence Ranking

Tf-idf and cosine similarity are used in this approach.The method follow the flowchart below:

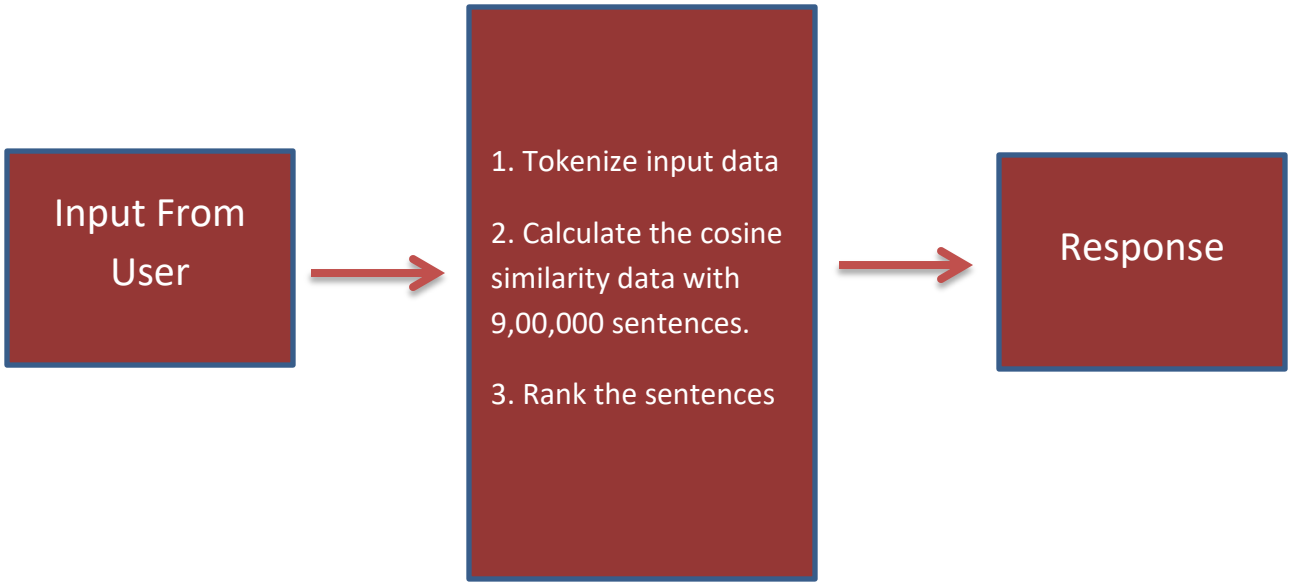

Fig:Flowchart of Sentence ranking method.

C. Article ranking with cosine similarity with tfidfs

In section B, we have already discussed tf-idf and cosine similarity. So the model follows by the flowchart below:

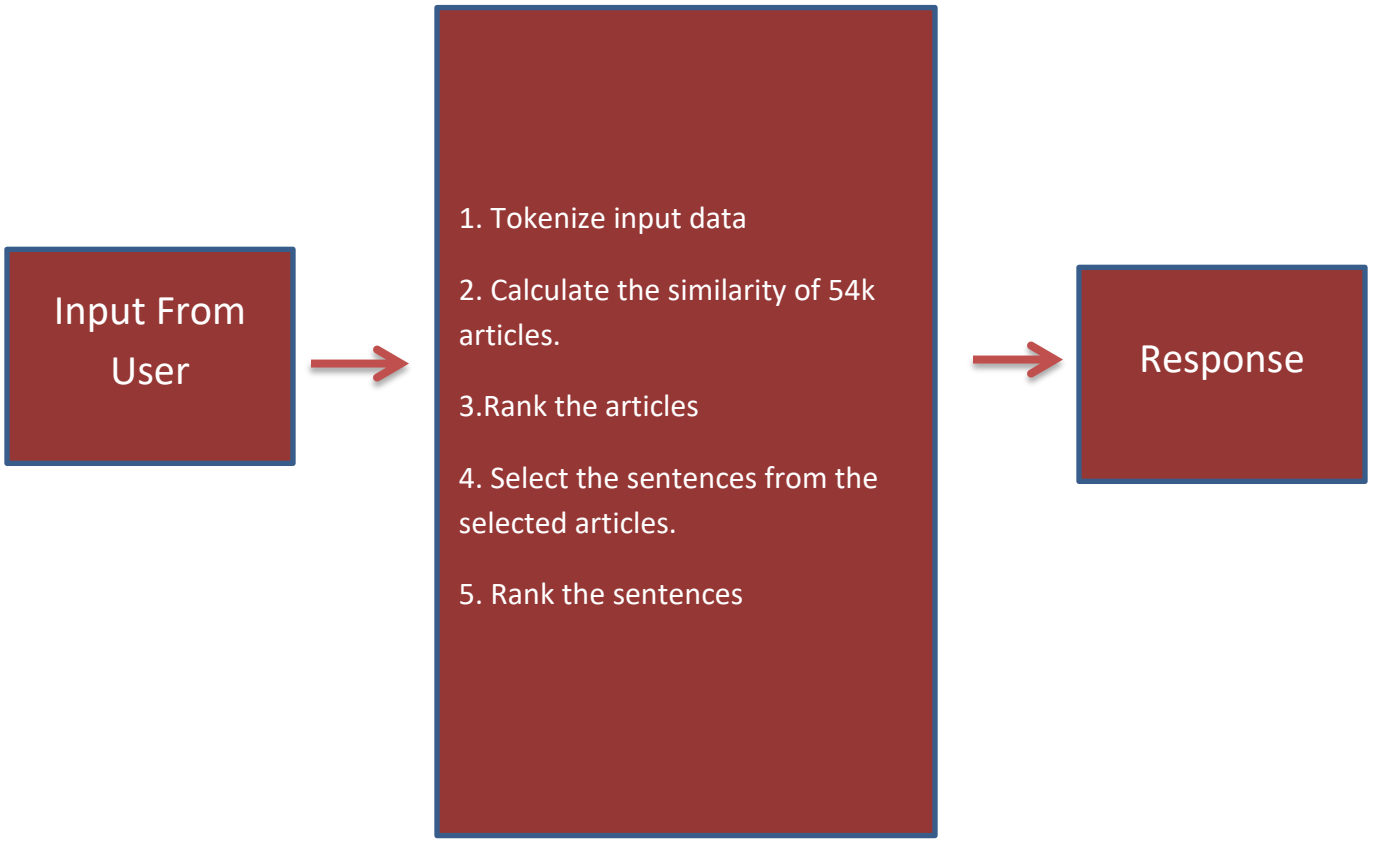

Fig:Flowchart of Article ranking method.

E. Experimental results and evaluation

Different questions have been queried for the methods. Experiment is done with a bunch of queries. In most cases, the method of article ranking with sentence was way better and satisfactory than the sentence ranking.

Method 1:

User Question	Response Selecting 1 sentence	Response Selecting 2 sentences
Bengali: করোনাভাইরাস প্রতিরোধে কি করছে খুলনা বিশ্ববিদ্যালয়? English: What is Khulna University doing to prevent coronavirus?	করোনাভাইরাস সংক্রমণ প্রতিরোধে ১৭ মার্চ থেকে দেশের সব শিক্ষা প্রতিষ্ঠান বন্ধ ঘোষণা করা হয়। English: All educational institutions in the country were declared closed from March 18 to prevent coronavirus infection.	করোনাভাইরাস সংক্রমণ প্রতিরোধে ১৭ মার্চ থেকে দেশের সব শিক্ষা প্রতিষ্ঠান বন্ধ ঘোষণা করা হয়। এসব এলাকার অনেকে মনে করছে, করোনাভাইরাস প্রতিরোধে এভাবে লকডাউন ঠিক আছে। English: All educational institutions in the country were declared closed from March 18 to prevent coronavirus infection. Many in these areas feel that coronavirus lockdown is okay.
Bengali: বাংলাদেশে প্রথম করোনা কবে শনাক্ত হয়? English:When was the first corona identified in Bangladesh?	Bengali: এ কারণে এই দুই দেশে প্রথম করোনা শনাক্ত দেখা গেছে। English: This is why the first corona has been identified in these two countries.	Bengali: এ কারণে এই দুই দেশে প্রথম করোনা শনাক্ত দেখা গেছে। করোনা শনাক্ত হয়েছে ২ হাজার ৮ জন। English: This is why the first corona has been identified in these two countries. Corona has identified 2,006 people

Fig: Experimental result of Chatbot with Sentence ranking method.

Method 2:

User Question	Response Selecting 1 sentence	Response Selecting 2 sentences
Bengali: করোনাভাইরাস প্রতিরোধে কি করছে খুলনা বিশ্ববিদ্যালয়? English: What is Khulna University doing to prevent coronavirus?	করোনাভাইরাস সংক্রমণ প্রতিরোধে হ্যান্ড স্যানিটাইজার উৎপাদন করছে খুলনা বিশ্ববিদ্যালয় । English: Khulna University is producing hand sanitizer to prevent coronavirus infection.	করোনাভাইরাস সংক্রমণ প্রতিরোধে হ্যান্ড স্যানিটাইজার উৎপাদন করছে খুলনা বিশ্ববিদ্যালয়। বুধবার (১৫ জুলাই) বিশ্ববিদ্যালয় এ স্যানিটাইজার তৈরির কাজ অনুষ্ঠিত হয় । English: Khulna University is producing hand sanitizer to prevent coronavirus infection. Sanitizer making work was held at the university on Wednesday (July 15).
Bengali: বাংলাদেশে প্রথম করোনা কবে শনাক্ত হয়? English:When was the first corona identified in Bangladesh?	Bengali: বাংলাদেশে ৮ মার্চ প্রথম তিন করোনা রোগী শনাক্ত হয় । English: The first three corona patients were identified on March 8 in Bangladesh	Bengali: বাংলাদেশে ৮ মার্চ প্রথম তিন করোনা রোগী শনাক্ত হয় । টাঙ্গাইলে প্রথম করোনা রোগী শনাক্ত হয় ৮ এপ্রিল । English: The first three corona patients were identified on March 8 in Bangladesh The first corona patient was identified in Tangail on 8 April

Fig: Experimental result of Chatbot Article ranking and sentence ranking method..

E. Bangla Chatbot application UI

Computer based software is implemented utilizing proposed method.

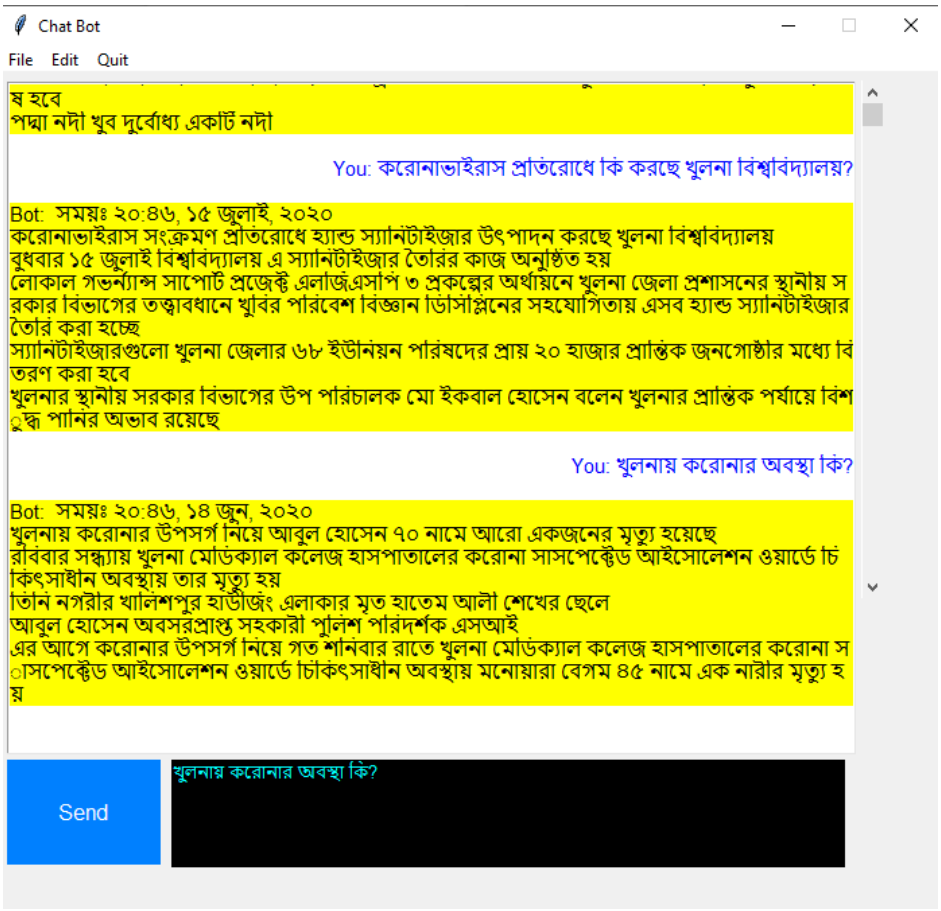

Fig:Bangla Informative Chatbot