Bangla Article Summary using sentence ranking

Abstract

Automated Bangla article summary is basically finding out what type of data or sentences are most relevant that altogether defines the sole of the entire article.. Despite several comprehensive textual datasets are available for different languages, a few small datasets are curated on Bangla language. As a result, a few works address Bangla document summary problem, and due to the lack of enough datal. In this work, we created a large dataset of Bangla articles from **The Daliy Ittefaq**, which contains around 70k articles. This huge diverse dataset helps us to create a system using sentence ranking.

Methodology

In this method we have ranked the sentences by following procedures:

Dataset

For dataset, We 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.

Summary

#Total Article count ==========> 53571

#unique unigram ==========> 252073

#Total Unigram ==========>13326446

#unique bigram ==========> 3718409

#Total Bigram ==========> 13579033

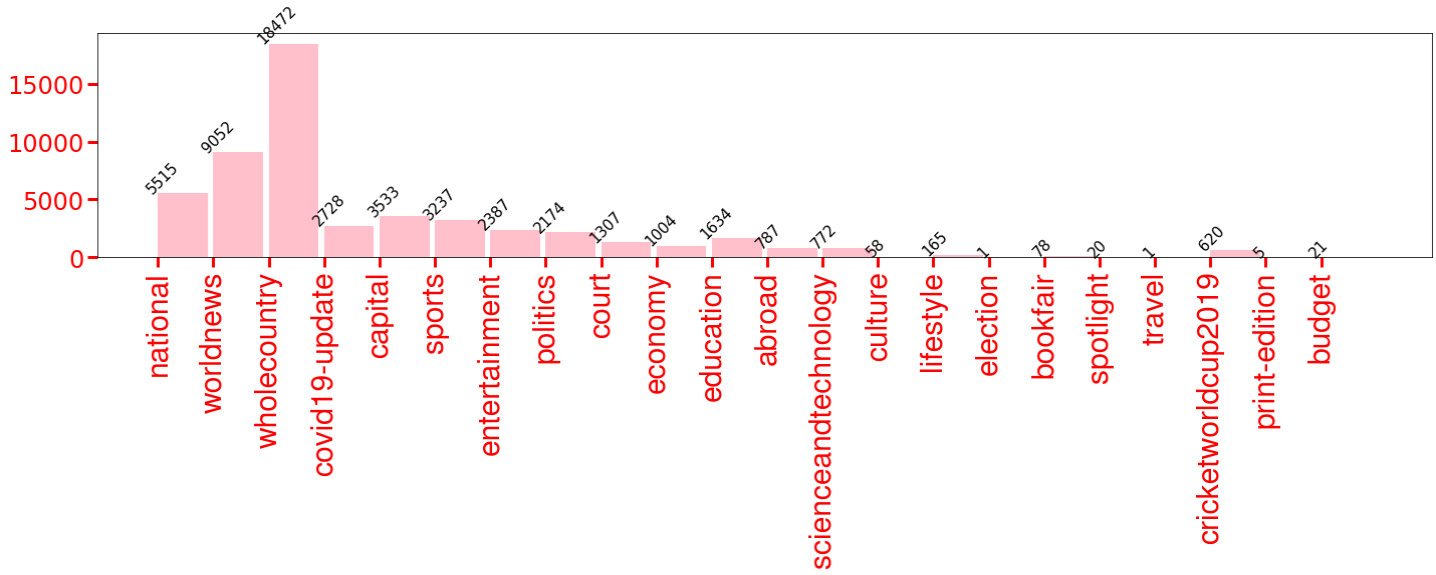
#unique trigram ==========> 8218523

#Total Trigram ==========> 24030358

Category Summary: **Total 22 Categories**

Type	News Count	Type (Bengali Version)

'national':	5515,	জাতীয়
'worldnews':	9052,	আন্তর্জাতিক
'wholecountry':	18472,	সারাদেশ
'covid19-update':	2728,	কোভিড-১৯
'capital':	3533,	রাজধানী
'sports':	3237,	খেলা
'entertainment':	2387,	বিনোদন
'politics':	2174,	রাজনীতি
'court':	1307,	কোর্ট
'economy':	1004,	অর্থনীতি
'education':	1634,	শিক্ষা
'abroad':	787,	বিদেশ
'Scienceandtechnology':	772,	বিজ্ঞান ও টেক
'culture':	58,	সংস্কৃতি
'lifestyle':	165,	জীবনযাপন
'bookfair':	78,	বইমেলা
'cricketworldcup2019':	620	ক্রিকেট বিশ্বকাপ -১৯



So, the methodology is consisted of following procedures:

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.

f.Small documents removal

Article of smaller lengths are removed in this case.

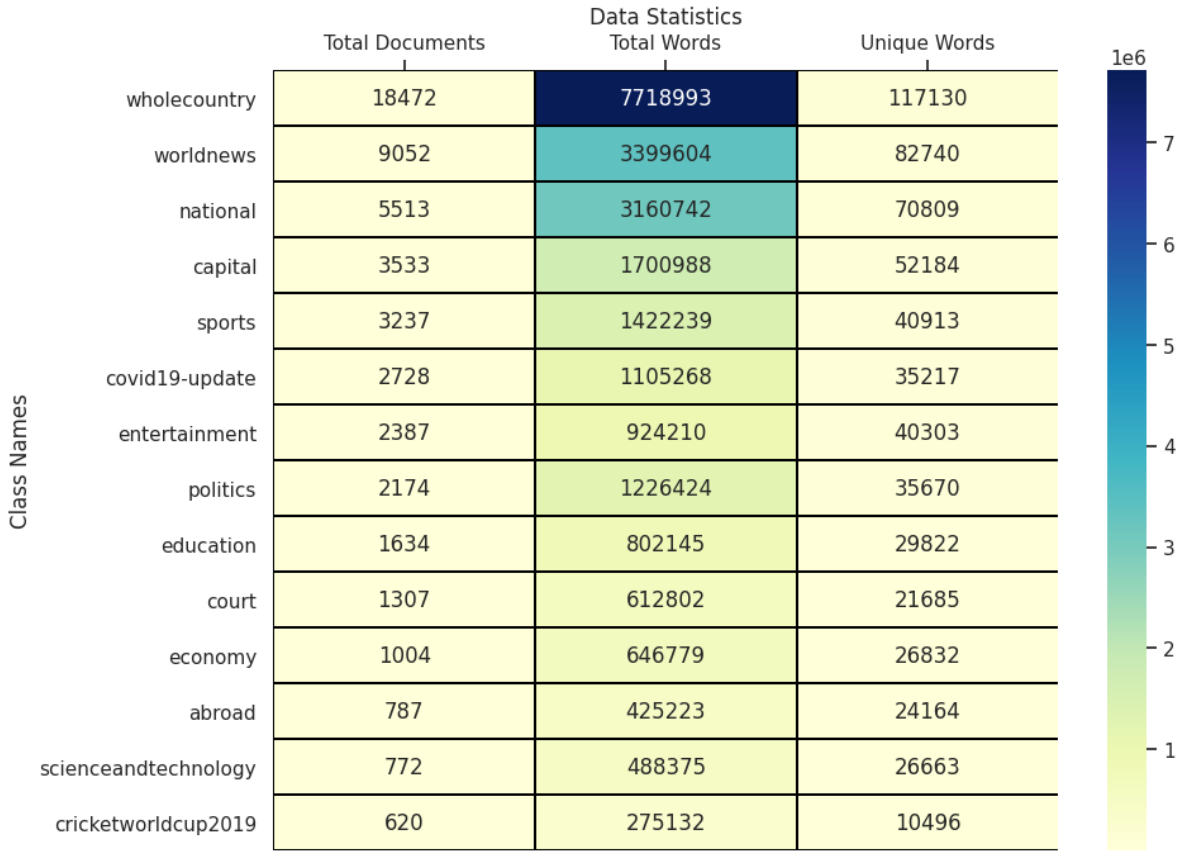


Fig:After pre-processing dataset

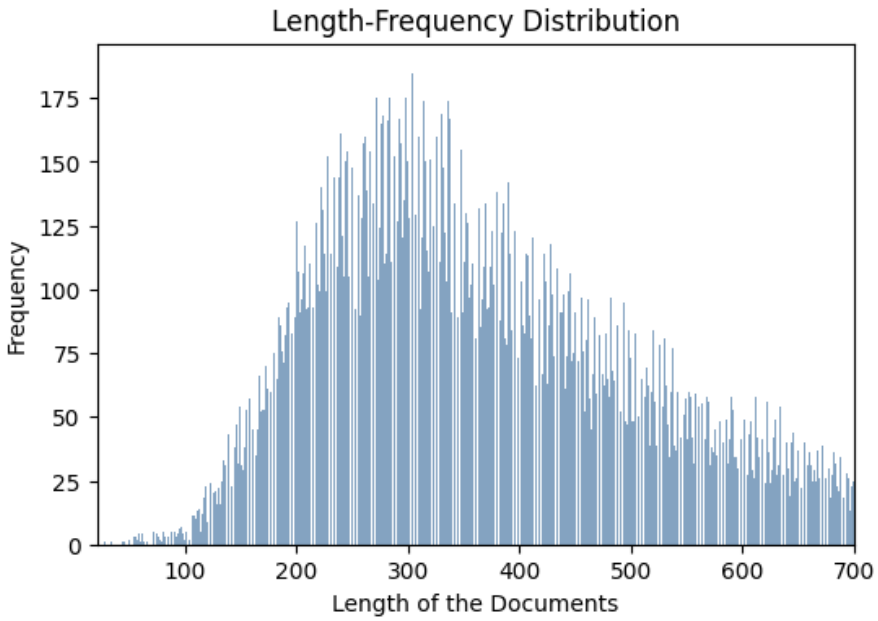


Fig:Lengths of documents

B. Bangla Article Summary using sentence ranking

In this method we have used sentence ranking.We have counted which words are important.Then for each sentences we have gone with the same approach to find out which word is most important. Next the sentences are ranked based on the important words.

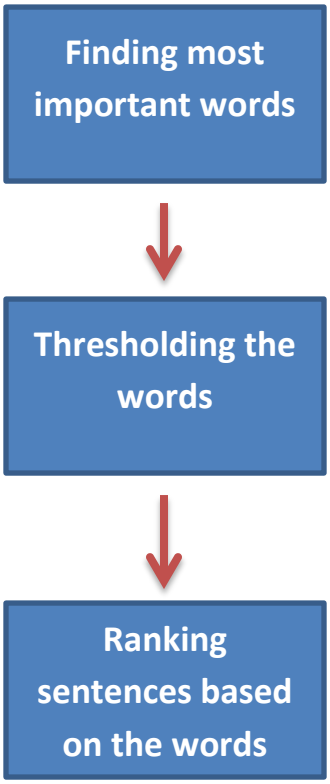
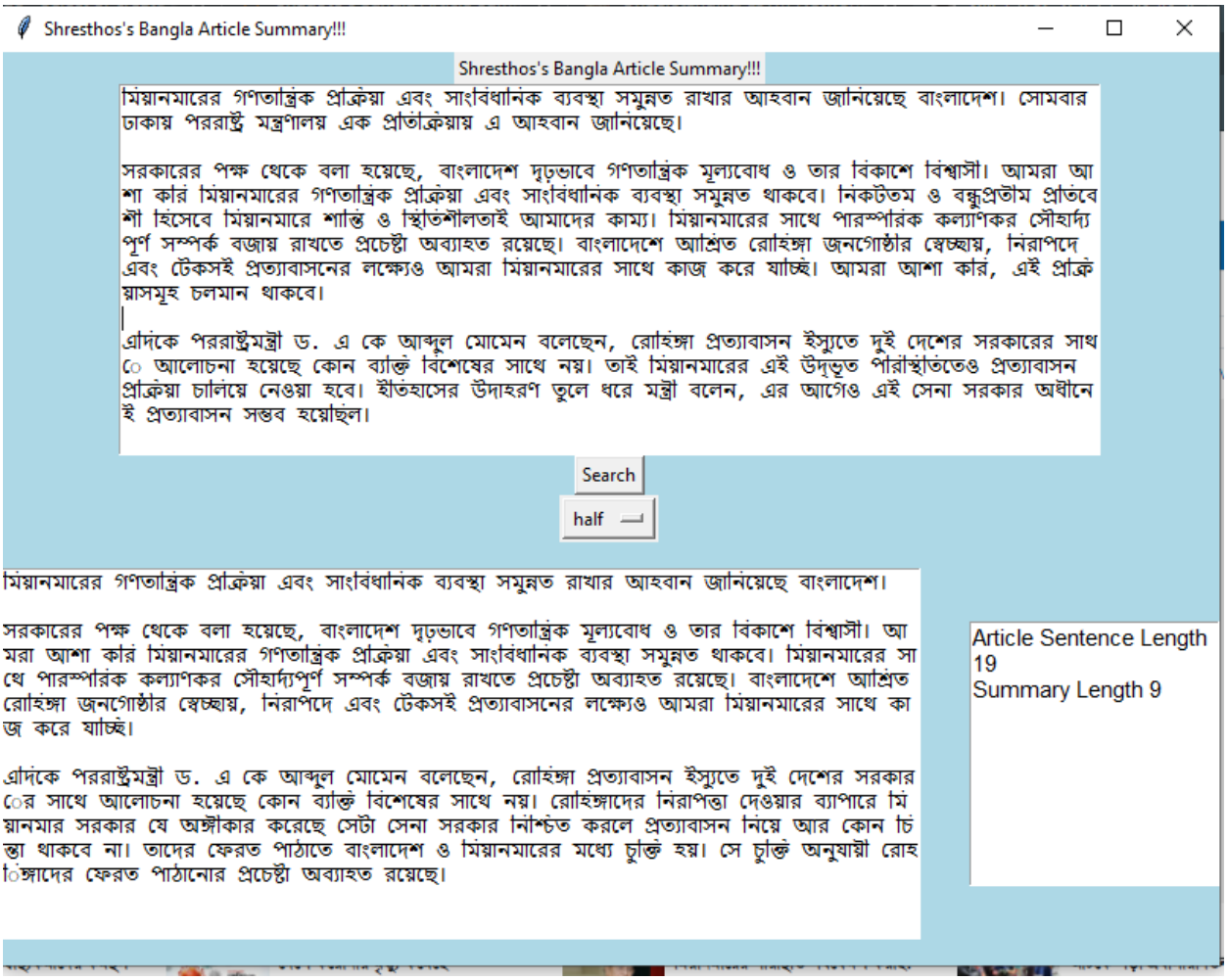
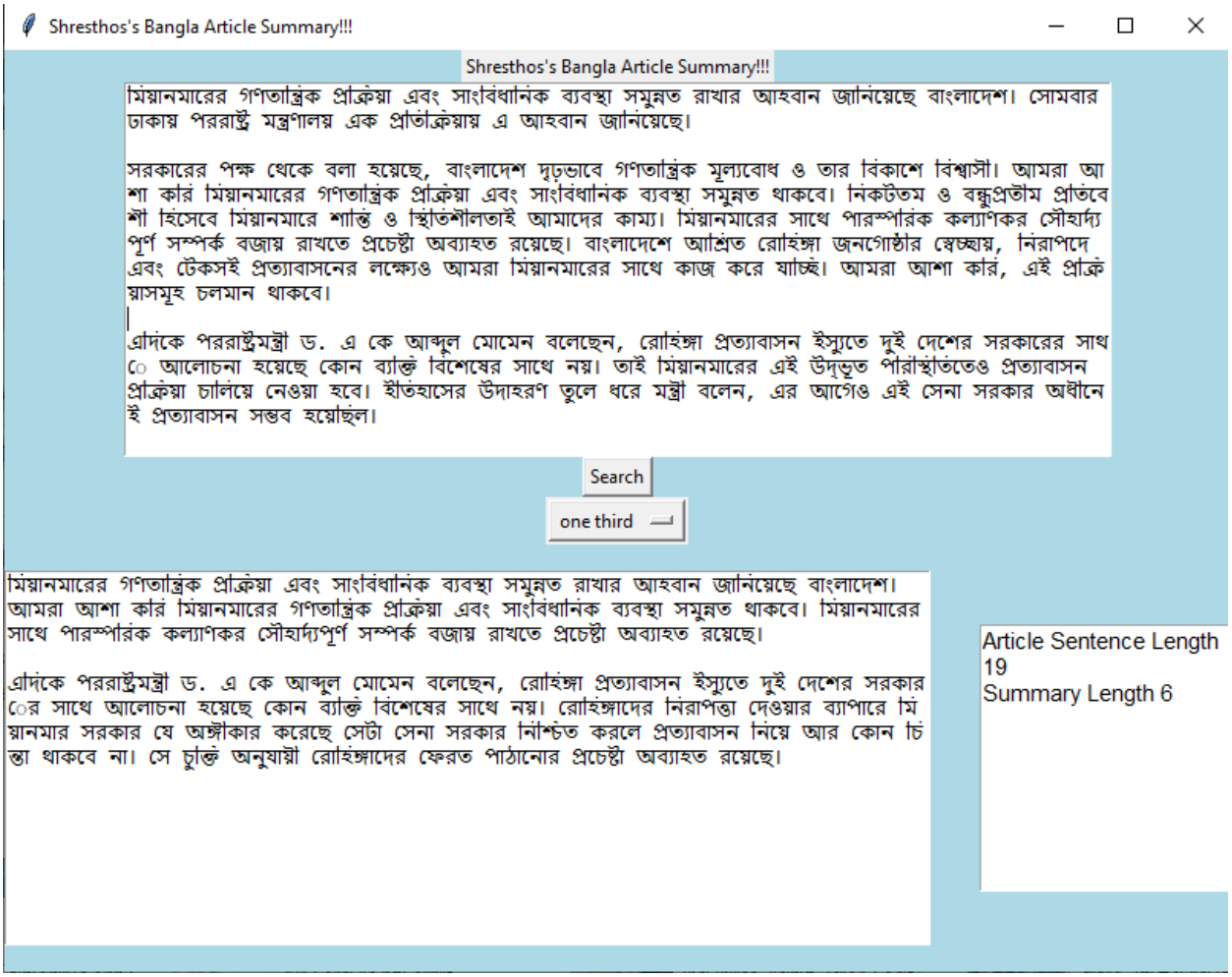


Fig:Sentence Ranking approach

B. Application UI




Shresthos's Bangla Article Summary!!!
Shresthos's Bangla Article Summary!!!
মিয়ানমারের গণতান্ত্রিক প্রক্রিয়া এবং সাংবিধানিক ব্যবস্থা সমুন্নত রাখার আহবান জানিয়েছে বাংলাদেশ। সোমবার ঢাকায় পররাষ্ট্র মন্ত্রণালয় এক প্রতিক্রিয়ায় এ আহবান জানিয়েছে।
সরকারের পক্ষ থেকে বলা হয়েছে, বাংলাদেশ দৃঢ়ভাবে গণতান্ত্রিক মূল্যবোধ ও তার বিকাশে বিশ্বাসী। আমরা আশা করি মিয়ানমারের গণতান্ত্রিক প্রক্রিয়া এবং সাংবিধানিক ব্যবস্থা সমুন্নত থাকবে। নিকটতম ও বন্ধুপ্রতীম প্রতিবেশী হিসেবে মিয়ানমারে শান্তি ও স্থিতিশীলতাই আমাদের কাম্য। মিয়ানমারের সাথে পারস্পরিক কল্যাণকর সৌহার্দ্য পূর্ণ সম্পর্ক বজায় রাখতে প্রচেষ্টা অব্যাহত রয়েছে। বাংলাদেশে আশ্রিত রোহিঙ্গা জনগোষ্ঠীর স্বেচ্ছায়, নিরাপদে এবং টেকসই প্রত্যাবাসনের লক্ষ্যেও আমরা মিয়ানমারের সাথে কাজ করে যাচ্ছি। আমরা আশা করি, এই প্রক্রিয়াসমূহ চলমান থাকবে।
এদিকে পররাষ্ট্রমন্ত্রী ড. এ কে আব্দুল মোমেন বলেছেন, রোহিঙ্গা প্রত্যাবাসন ইস্যুতে দুই দেশের সরকারের সাথে আলোচনা হয়েছে কোন ব্যক্তি বিশেষের সাথে নয়। তাই মিয়ানমারের এই উদ্ভূত পরিস্থিতিতেও প্রত্যাবাসন প্রক্রিয়া চালিয়ে নেওয়া হবে। ইতিহাসের উদাহরণ তুলে ধরে মন্ত্রী বলেন, এর আগেও এই সেনা সরকার অধীনেই প্রত্যাবাসন সম্ভব হয়েছিল।
Search
one third
মিয়ানমারের গণতান্ত্রিক প্রক্রিয়া এবং সাংবিধানিক ব্যবস্থা সমুন্নত রাখার আহবান জানিয়েছে বাংলাদেশ। আমরা আশা করি মিয়ানমারের গণতান্ত্রিক প্রক্রিয়া এবং সাংবিধানিক ব্যবস্থা সমুন্নত থাকবে। মিয়ানমারের সাথে পারস্পরিক কল্যাণকর সৌহার্দ্যপূর্ণ সম্পর্ক বজায় রাখতে প্রচেষ্টা অব্যাহত রয়েছে।
এদিকে পররাষ্ট্রমন্ত্রী ড. এ কে আব্দুল মোমেন বলেছেন, রোহিঙ্গা প্রত্যাবাসন ইস্যুতে দুই দেশের সরকারের সাথে আলোচনা হয়েছে কোন ব্যক্তি বিশেষের সাথে নয়। রোহিঙ্গাদের নিরাপত্তা দেওয়ার ব্যাপারে মিয়ানমার সরকার যে অঙ্গীকার করেছে সেটা সেনা সরকার নিশ্চিত করলে প্রত্যাবাসন নিয়ে আর কোন চিন্তা থাকবে না। সে চুক্তি অনুযায়ী রোহিঙ্গাদের ফেরত পাঠানোর প্রচেষ্টা অব্যাহত রয়েছে।
Article Sentence Length
19
Summary Length 6