Hindi Inshorts Generation

A. Simriti Koul, *17BCE2211*, B. Arastoo Joshi, *17BCE2111*

Abstract— **Automatic summarization plays an important role in document processing system and information retrieval system. Generation of summary of a text document is a very important part of NLP. There are a number of scenarios where automatic construction of such summaries is useful. Text summarization is that process which converts a larger text into its shorter form maintaining its information. Summary of a longer text saves the reading time as it contains lesser number of lines but all important information of the original text document.**

In this paper we present a novel approach for text summarization of Hindi text document based on some linguistic rules. Dead wood words and phrases are also removed from the original document to generate the lesser number of words from the original text. Proposed system is tested on various Hindi inputs and accuracy of the system in form of number of lines extracted from original text containing important information of the original text document.

Introduction

A summary of a document is (much) shorter text that conveys the most important information from the source document. There are a number of outlines where automatic construction of such summaries is useful. E.g.: System of an information retrieval could present an automatically built summary in its list of retrieval outcome for this user to firstly decide which sentences are interesting and worth to be in the summary. We have considered Hindi as a language of study. It is written in the Devanagari script which has largest alphabet set. Hindi is an official language of India. It the native language of most people living in Delhi, Chhattisgarh, Himachal Pradesh, Chandigarh, Bihar, Jharkhand, Madhya Pradesh, Haryana, and Rajasthan. So for people who do not know English but want to read articles on the Internet, automatic summarization would play lion's role in it. While performing related search, it is observed that a lot of work has been done on English language as ample amount of resources are readily available for the same. Relatively very few have shown interest in the case of Hindi language. It motivated us for considering Hindi as a study language.

A. Abbreviations and Acronyms

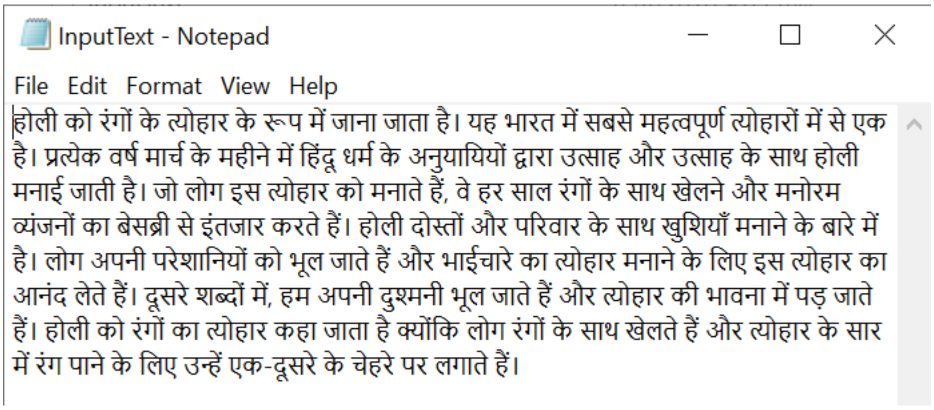
- NLP - Natural Language Processing.

B. Algorithm

```
import math
import networkx
import numpy
from nltk.tokenize.punkt import PunktSentenceTokenizer
from sklearn.feature_extraction.text import TfidfTransformer, CountVectorizer
def getrank(document):
        sentences = PunktSentenceTokenizer().tokenize(document)
        bow_matrix = CountVectorizer().fit_transform(sentences)
        normalized = TfidfTransformer().fit_transform(bow_matrix)
        similarity_graph = normalized * normalized.T
        nx_graph = networkx.from_scipy_sparse_matrix(similarity_graph)
        values = networkx.pagerank(nx_graph)
        sentence_array = sorted(((values[i], s) for i, s in enumerate(sentences)),
                                reverse=True)
```

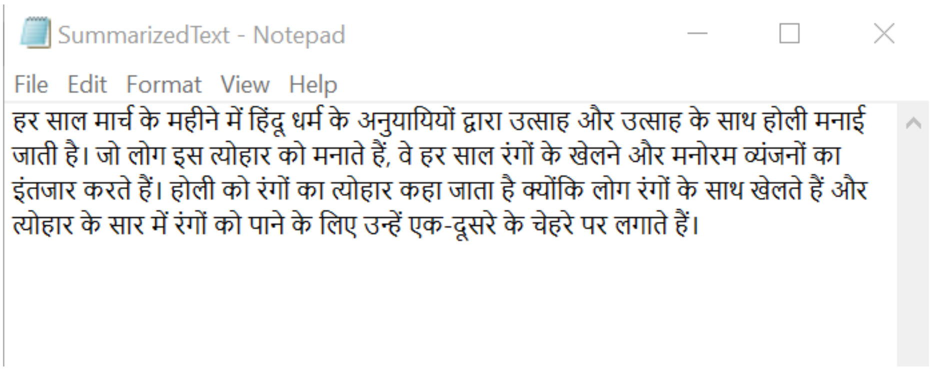
```
sentence_array = numpy.asarray(sentence_array)
freq_max = float(sentence_array[0][0])
freq_min = float(sentence_array[len(sentence_array) - 1][0])
temp_array = []
for i in range(0, len(sentence_array)):
if freq_max - freq_min == 0:
        temp_array.append(0)
else:
        temp_array.append((float(sentence_array[i][0]) - freq_min) / (freq_max -
                freq_min))
threshold = (sum(temp_array) / len(temp_array)) + 0.25
sentence_list = []
for i in range(0, len(temp_array)):
if temp_array[i] > threshold:
sentence_list.append(sentence_array[i][1])
seq_list = []
for sentence in sentences:
        if sentence in sentence_list:
                seq_list.append(sentence)
return seq_list
```

C. Application on a hindi paagraph and its result



होली को रंगों के त्योहार के रूप में जाना जाता है। यह भारत में सबसे महत्वपूर्ण त्योहारों में से एक है। प्रत्येक वर्ष मार्च के महीने में हिंदू धर्म के अनुयायियों द्वारा उत्साह और उत्साह के साथ होली मनाई जाती है। जो लोग इस त्योहार को मनाते हैं, वे हर साल रंगों के साथ खेलने और मनोरम व्यंजनों का बेसब्री से इंतजार करते हैं। होली दोस्तों और परिवार के साथ खुशियाँ मनाने के बारे में है। लोग अपनी परेशानियों को भूल जाते हैं और भाईचारे का त्योहार मनाने के लिए इस त्योहार का आनंद लेते हैं। दूसरे शब्दों में, हम अपनी दुश्मनी भूल जाते हैं और त्योहार की भावना में पड़ जाते हैं। होली को रंगों का त्योहार कहा जाता है क्योंकि लोग रंगों के साथ खेलते हैं और त्योहार के सार में रंग पाने के लिए उन्हें एक-दूसरे के चेहरे पर लगाते हैं।

Fig. Input Text



हर साल मार्च के महीने में हिंदू धर्म के अनुयायियों द्वारा उत्साह और उत्साह के साथ होली मनाई जाती है। जो लोग इस त्योहार को मनाते हैं, वे हर साल रंगों के खेलने और मनोरम व्यंजनों का इंतजार करते हैं। होली को रंगों का त्योहार कहा जाता है क्योंकि लोग रंगों के साथ खेलते हैं और त्योहार के सार में रंगों को पाने के लिए उन्हें एक-दूसरे के चेहरे पर लगाते हैं।

Fig. Summarized Text

CONCLUSION

We have considered text summarization which is based on sentence extraction method. Using statistics and normalization, we have tried to achieve context conservation. In the flow of proposed approach first feature extraction comes, and then sentence scoring and lastly selection of higher ranked sentences as a summary. Six statistical and two linguistic features are used for this single document summarization. We have considered the Hindi, an official language of India, as a language of study.

REFERENCES

Vipul Dalal, Latesh Malik, *Data Clustering Approach for Automatic Text Summarization of Hindi Documents using Particle Swarm Optimization and Semantic Graph*, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-7 Issue-3, July, 2017. Accessed on: December, 21, 2019.

- Deepali P Kadam, Mrs. Nita Patil, Mrs. Archana Gulathi, A Comparative Study of Hindi Text Summarization Techniques: Genetic Algorithm and Neural Network, International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4, Special Issue March 2015.
- Jathavedan M, Dhanya P.M, Comparative Study of Text Summarization in Indian Languages, International Journal of Computer Applications (0975 – 8887), Volume 75– No.6, August 2013.
- Krish Perumal, Bidyut Baran Chaudhuri, Language Independent Sentence Extraction Based Text Summarization, Proceedings of ICON-2011: 9th International Conference on Natural Language Processing.
- Sheetal Shimpikar, Sharvari Govilkar ,A Survey of Text Summarization Techniques for Indian Regional Languages , International Journal of Computer Applications, Foundation of Computer Science (FCS), NY, USA, Volume 165 - Number 11, 2017.
- Barzilay and Mckeown, “Sentence fusion for multi document news summarization”, Computer Linguistics, vol-31, pp. 297-338, 2005.
- Genest and Lapalme, “Framework for abstractive summarization using text to text generation”, in Proceeding of workshops on Monolingual Text to Text generation., 2011, pp.64-73, information item.
- Pierre-Etienne Genest, Guy Lapalme, “Framework for Abstractive Summarization using Text-to-Text Generation”, Workshop on Monolingual TextTo-Text Generation, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 64–73,Portland, Oregon, 24 June 2011. c 2011 Association for Computational Linguistics.
- M. S. Binwahlan, Salim, N., & Suanmali, L.: “Swarm Based Text Summarization”, Computer Science and Information Technology – Spring Conference, 2009. IACSITSC '09. International Association of, 2009, pp. 145150.
- Upendra Mishra, Chandra Prakash.: MAULIK: “An Effective Stemmer for Hindi Language”, International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397, Vol. 4 No. 05 May 2012.