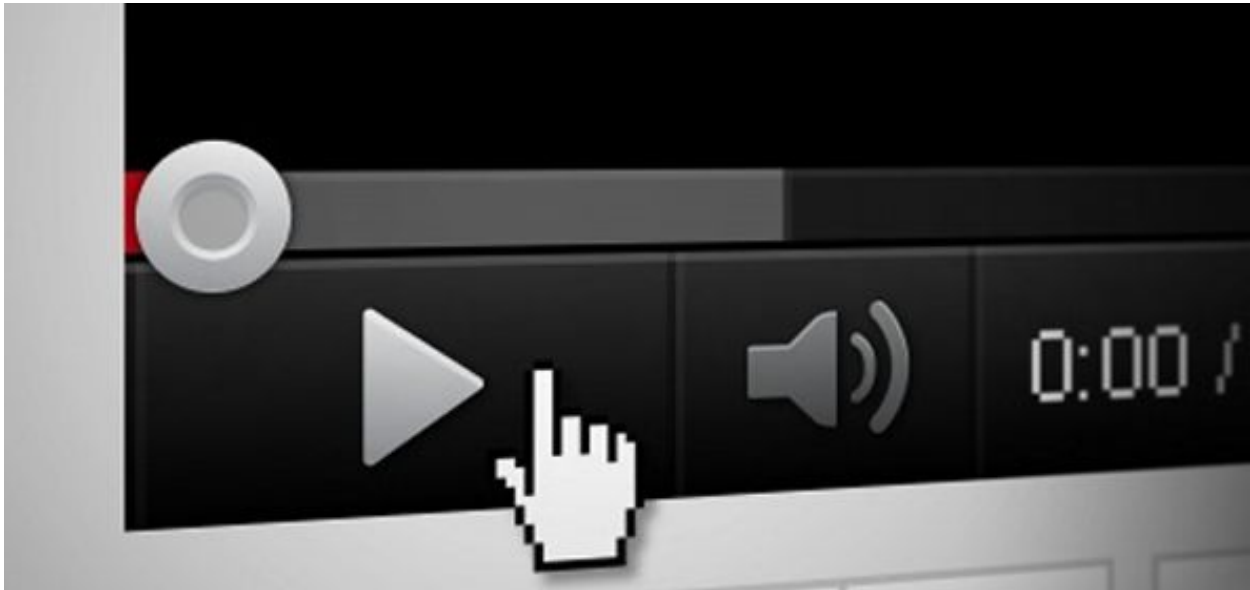


Youtube Video Extraction - Team 8

Mrinal Dhar (201325118) - Shashank Agrawal (201301056) - Tanushri Sharma (201505518)



Introduction

Many news channels today publish videos everyday covering the important events of the day. Videos are an interesting mode of presenting information that might otherwise seem dull in a text article and make it highly engaging for news readers and viewers.

The aim of this project is to create a system that extracts videos related to a news article from YouTube. This project also integrates a web server and online interface for the backend system, and renders the top 10 ranked videos for an article in the browser itself.

The videos might be uploaded by a news publisher or a user just sharing a video about the event. This is a bilingual system, supporting both English and Hindi.

Key challenges

1. How to get videos around the article published time and rank them in terms of relevancy.
2. How to fit it in Indian Language case, the common case is that the article will be in Indian language while the video might be in English.

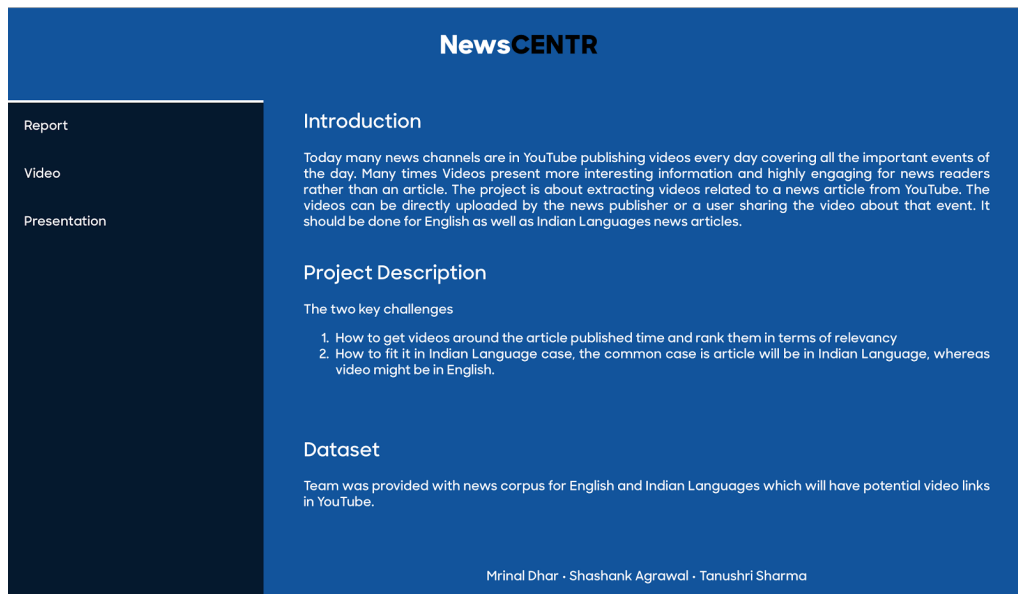
Approach to a solution

1. Since the articles from the dataset did not include the date or time of publishing, there was no way of retrieving only those videos which were published around the same time. So, our algorithm extracts the most recent videos relevant to the content of the article.
2. In order to extract videos for articles whose text might be in a different language, like Hindi, we used transliteration to generate the English counterpart of the search query, and extracted videos based on those terms.

Dataset

We were provided with news corpus for English and Indian Languages which will have potential video links in YouTube. There were about 30,000 news articles in the given dataset, all of them in Hindi, provided in JSON format for easy and efficient parsing.

[illegible]



Modules

The code has been divided into several modules to allow for efficient debugging and testing.

The first is the parser, which reads the JSON file with the articles and extracts all the required information into memory. This module also assigns an ID to every article and stores them in our data structure.

Then, the web server module gets into play. The web server is based on Python's BaseHTTPServer, and we build on that to serve our custom GET requests. These requests are served by a thread that shares the memory space with our article data structure.

An NER module has been used to categorize the words in the title of an article as named entities by calling the Hindi NER Module hosted on the LTRC web server.

A tokenizer module has been used to identify tokens from a sentence, both from the title and body text of an article.

Google's Translate API has been used, along with transliteration, to convert Hindi titles to English.

Google's YouTube API has been used to query YouTube and retrieve list of videos.

NewsCENTR

Please select an article to read.

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा :

ट्रंप

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह ...

जीतू को आईएसएसएफ विश्व कप में स्वर्ण

बैकाल। भारतीय निशानेबाज जीतू राय ने पूर्व ओलंपिक चैंपियन पांग वेइ को पछाड़ते हुए यहां आईएसएसएफ विश्व कप की 50 मीटर पिस्टल स्पर्धा में शुक्रवार को स्व ...

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा :

ट्रंप

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह ...

WT-20: भारत-पाक मैच के लिए अर्धसैनिक बल मुहैया करायेगी केंद्र सरकार

नई दिल्ली: भारत और पाकिस्तान के बीच धर्मशाला में होने वाले टी20 मैच को लेकर बनी अनिश्चितता के बीच मृगमयी राजनाथ सिंह ने आज कहा कि 19 मार्च को होने ...

टी-20 विश्व कप : गेंद अब पीसीबी के पाले में

नई दिल्ली: बीसीसीआई ने टी-20 विश्व कप के दौरान पाकिस्तानी टीम को

Mrinal Dhar · Shashank Agrawal · Tanushri Sharma

NewsCENTR

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा : ट्रंप

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा :

ट्रंप

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह ...

जीतू को आईएसएसएफ विश्व कप में स्वर्ण

बैकाल। भारतीय निशानेबाज जीतू राय ने पूर्व ओलंपिक चैंपियन पांग वेइ को पछाड़ते हुए यहां आईएसएसएफ विश्व कप की 50 मीटर पिस्टल स्पर्धा में शुक्रवार को स्व ...

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा :

ट्रंप

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह ...

WT-20: भारत-पाक मैच के लिए अर्धसैनिक बल मुहैया करायेगी केंद्र सरकार

नई दिल्ली: भारत और पाकिस्तान के बीच धर्मशाला में होने वाले टी20 मैच को लेकर बनी अनिश्चितता के बीच मृगमयी राजनाथ सिंह ने आज कहा कि 19 मार्च को होने ...

टी-20 विश्व कप : गेंद अब पीसीबी के पाले में

नई दिल्ली: बीसीसीआई ने टी-20 विश्व कप के दौरान पाकिस्तानी टीम को



अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह राष्ट्रपति बन गए तो वह एच1 बी वीजा कार्यक्रम को खत्म कर देंगे।' हालांकि, अपने इस बयान से एक घंटा पहले उन्होंने इसके उलट कहा था कि अमेरिका को

Mrinal Dhar · Shashank Agrawal · Tanushri Sharma

NewsCENTR

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा :

ट्रंप

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह ...

जीतू को आईएसएसएफ विश्व कप में स्वर्ण

बैकग्राउंड। भारतीय निशानेबाज जीतू राव ने पूरे ओलंपिक सेमिफाइनल पांग वेड को पछाड़ते हुए यहां आईएसएसएफ विश्व कप की 50 मीटर स्मिथल स्पर्धा में शुक्रवार को स्व ...

राष्ट्रपति बना तो एच1 बी वीजा खत्म कर दूंगा :

ट्रंप

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रबल दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह ...

WT-20: भारत-पाक मैच के लिए अर्धसैनिक बल मुहैया करायेगी केंद्र सरकार

नई दिल्ली। भारत और पाकिस्तान के बीच धर्मशाला में होने वाले टी20 मैच को लेकर बनी अनिश्चितता के बीच नूतनमयी राजनाथ सिंह ने आज कहा कि 19 मार्च को होने ...

टी-20 विश्व कप : गेंद अब पीसीबी के पाले में

नई दिल्ली। बीसीसीआई ने टी-20 विश्व कप के दौरान पाकिस्तानी टीम को

इसलिए जरूरी है, क्योंकि पाकिस्तान के पास परमाणु हथियार हैं और उनकी सुरक्षा जरूरी है। उन्होंने कहा, मुझे लगता है कि हमारी सेना को कुछ समय के लिए अफगानिस्तान में रहना होगा। क्योंकि पाकिस्तान के परमाणु हथियारों ने पूरा खेल बदल दिया है। पिछले साल ट्रंप ने पाकिस्तान को दुनिया का सबसे खतरनाक देश बताया था। एक साक्षात्कार में उन्होंने संकेत दिया था कि पाकिस्तान के परमाणु निशस्त्रीकरण की जरूरत है। उधर, ट्रंप ने एक शीर्ष अमेरिकी सीनेटर को अपने राष्ट्रीय सुरक्षा दल का प्रमुख नियुक्त करने की घोषणा की है। एक बयान में ट्रंप ने कहा कि व्यापार और आमजन जैसे मुद्दों पर उन्हें सलाह दे चुके जेफ सेशन्स उनकी राष्ट्रीय सुरक्षा सलाहकार समिति के प्रमुख होंगे।



Donald Trump Changed His ...



Dr. Michio Kaku America Ha...



Trump's Softened Stance on ...



PM Modi instrumental in op...

Mrinal Dhar · Shashank Agrawal · Tanushri Sharma

Experiments

Our initial approach was to use the keywords that were already provided in JSON format in the dataset. Some of the articles had keywords, both in English and Hindi, and they were used to build the query string for YouTube's search interface.

The URLs for these articles were also given in the dataset. Since many Content Management Systems (CMS), which the news portals are based on, give the title of the article in the URL itself, in English, we tokenized the URLs to extract terms to be used in place of the keywords for the articles where there were no given keywords.

Next, we converted the title of the article from Hindi to English using transliteration. Using this as query text, we were able to extract more videos.

Another approach that we tried was extracting keywords from the article text, performing stopword removal on them, and using that to rank the videos obtained by comparing the tags and description of the videos.

Finally, we used a Named Entity Recognition module to extract the proper nouns in the article title. This allowed us to improve the quality of our search query, and thereby the overall list of videos retrieved from YouTube.

Results & Analysis

We were able to obtain better results using our modified approach in Phase 3. By using the title of the article, along with keywords from the text or provided otherwise, we were able to achieve a better ranking algorithm for the videos along with a larger relevant set of YouTube videos.

Reference links

Link to GitHub **repository**: <https://github.com/mrinaldhar/newscentr>

Link to GitHub **webpage**: <http://mrinaldhar.github.io/newscentr/>

Link to YouTube **video**: <https://youtu.be/qAzP9lYWc5k>

Link to SlideShare **presentation**: <https://www.slideshare.net/secret/jcc7dPCP4rtdZv>

Link to DropBox **folder**: <https://goo.gl/k9lstN>