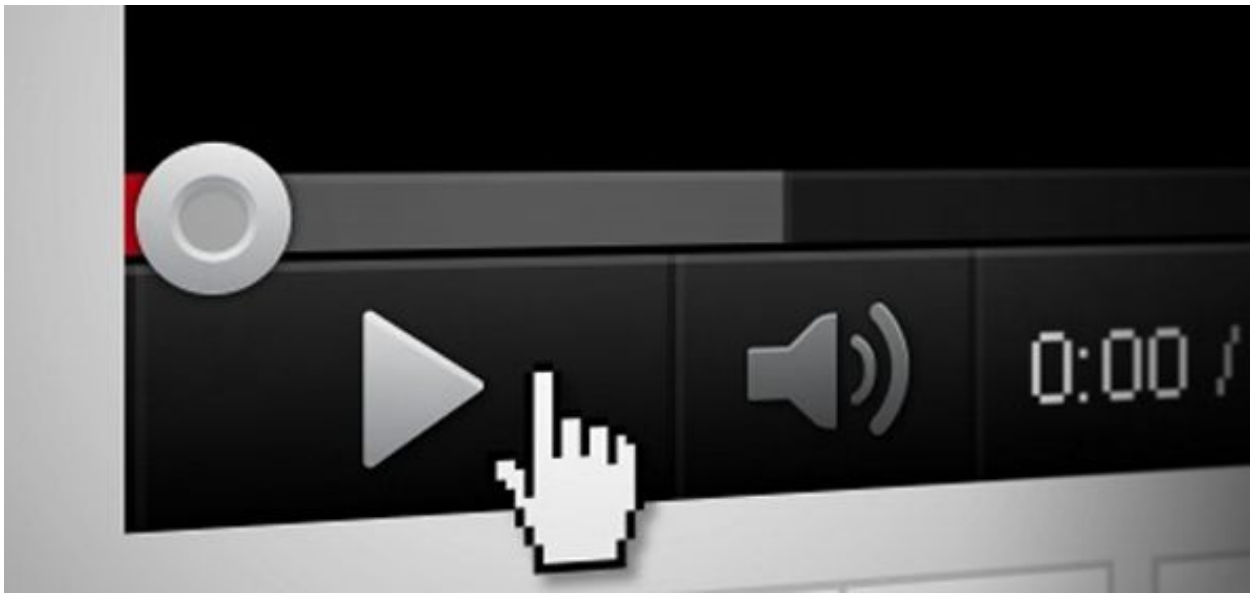
Information Retrieval and Extraction (IRE)

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

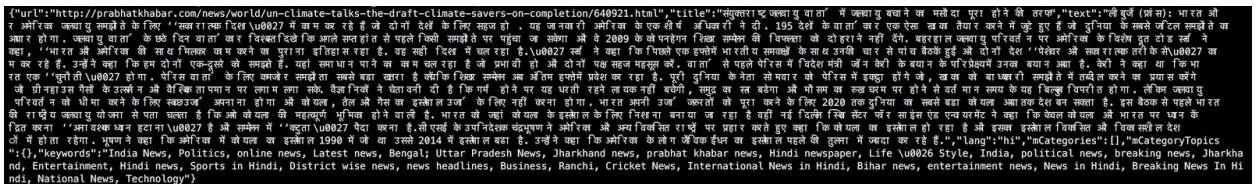
1. How to get videos around the article published time and rank them in terms of relevancy.
2. How to fit it in Indian Language case, the common case is that the article will be in Indian language while the video might be in English.

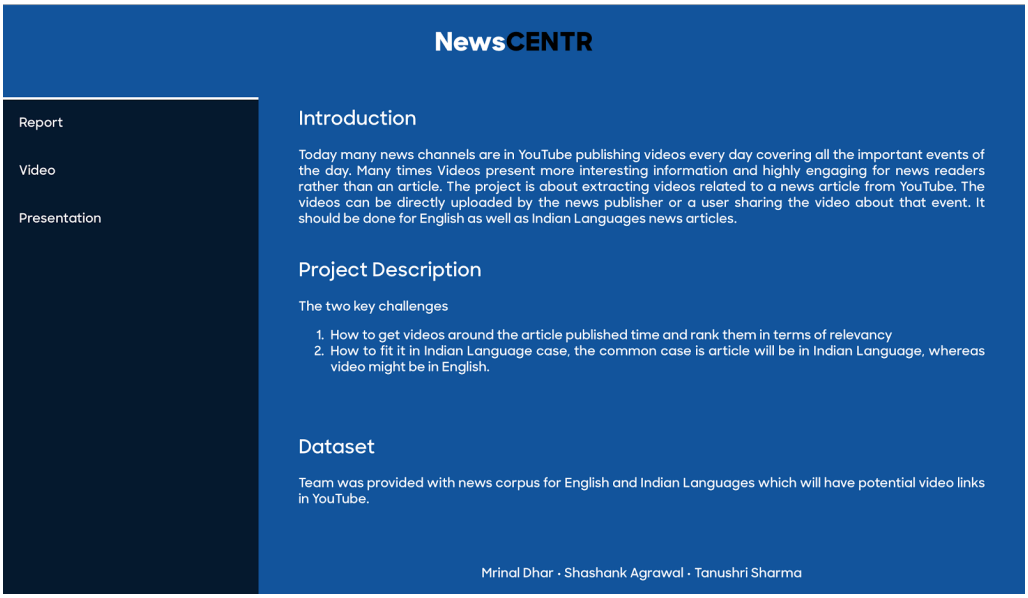
Approach to a solution

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.


{"url":"http://prabhatkhabar.com/news/world/un-climate-talks-the-draft-climate-savers-on-completion/640921.html","title":"संयुक्तराष्ट्र जलवायु वार्ता में जलवायु बचाने का मसौदा पूरा होने की तरफ","text":"ली बुर्ज (फ्रांस): भारत और अमेरिका जलवायु समझौते के लिए ''स्वरात्मक दिशा\u0027 में काम कर रहे हैं जो दोनों देशों के लिए सहज हो . यह जानकारी अमेरिका के एक शीर्ष अधिकारी ने दी. 195 देशों के वार्ता कर एक ऐसा खाका तैयार करने में जुटे हुए हैं जो दुनिया के सबसे जटिल समझौते का आधार होगा. जलवायु वार्ता के छठे दिन वार्ता कर विश्वास दिखे कि आगे सप्ताह से पहले किसी समझौते पर पहुंचा जा सकेगा और वे 2009 के कोपनहेगन शिखर सम्मेलन की विफलता को दोहराने नहीं देंगे. बहरहाल जलवायु परिवर्तन पर अमेरिका के विशेष दूत टोड स्टर्न ने कहा, ''भारत और अमेरिका की साथ मिलकर काम करने का पुराना इतिहास रहा है. वह सही दिशा में चल रहा है.\u0027\u0027 स्टर्न ने कहा कि पिछले एक हफ्तेमें भारतीय समझौते के साथ उनकी चार से पांच बैठकें हुई और दोनों देश ''पेरिस और सकारात्मक तरीके से\u0027 काम कर रहे हैं. उन्होंने कहा कि हम दोनों एक-दूसरे को समझते हैं. यहां लाखान पाने का काम चल रहा है जो प्रभावी हो और दोनों पक्ष सहज महसूस करें. वार्ता से पहले पेरिस में विदेश मंत्री जॉन केरी के बयान के परिप्रेक्ष्यमें उनका बयान आया है. केरी ने कहा था कि भारत एक ''घूरोती\u0027 होगा. पेरिस वार्ता के लिए वाकई समझौता उससे बड़ा खतरा है क्योंकि शिखर सम्मेलन अब अंतिम इसलेमें प्रवेश कर रहा है. पूरी दुनिया के नेता सोमवार को पेरिस में इकट्ठा होंगे जो, खास को बाध्यकारी समझौते में तब्दील करने का प्रयास करेंगे जो ग्रीनहाउस गैसों के उत्सर्जन अंतरराष्ट्रीय तापमान पर लगाम लगा सके. वैज्ञानिकों ने चेतावनी दी है कि गर्म होने पर यह धरती रहने लायक नहीं बचेगी. समुद्र का स्तर बढ़ेगा और मौसम का रुख चरम पर होने से वर्तमान समय के यह बिल्कुल विपरीत होगा. लेकिन जलवायु परिवर्तन को धीमा करने के लिए स्वछउर्ज अपनाना होगा और कोयला, तेल और गैस का इस्तेमाल उर्ज के लिए नहीं करना होगा. भारत अपनी उर्ज जरूरतों को पूरा करने के लिए 2020 तक दुनिया का सबसे बड़ा कोयला आधारित देश बन सकता है. इस बैठक से पहले भारत की राष्ट्रीय जलवायु योजना से पता चलता है कि ओ कोयला की महत्वपूर्ण भूमिका होने वाली है. भारत कोयला के इस्तेमाल के लिए निश्चित बनाया जा रहा है वहीं नई दिल्ली स्थित सेंटर फॉर साइंस एंड एन्वायरमेंट ने कहा कि कोयला और भारत पर ध्यान केंद्रित करना ''आंशिक ध्यान हटाना\u0027 है और सम्मेलन में ''बदतर\u0027 परिणाम पैदा करना है.सीएसई के उपनिदेशक चंद्रभूषण ने अमेरिका और अन्य विकसित राष्ट्रों पर प्रहार करते हुए कहा कि कोयला का इस्तेमाल हो रहा है और इसका इस्तेमाल विकसित और विकासशील देशों में होता रहेगा. भूषण ने कहा कि अमेरिका में कोयला का इस्तेमाल 1990 में जो था उससे 2014 में इस्तेमाल बढ़ा है. उन्होंने कहा कि अमेरिका के लोग जैविक ईंधन का इस्तेमाल पहले की तुलना में ज्यादा कर रहे हैं.","lang":"hi","mCategories":[],"mCategoryTopics":{},"keywords":"India News, Politics, online news, Latest news, Bengal; Uttar Pradesh News, Jharkhand news, prabhat khabar news, Hindi newspaper, Life \u0026 Style, India, political news, breaking news, Jharkhand, Entertainment, Hindi news, Sports in Hindi, District wise news, news headlines, Business, Ranchi, Cricket News, International News in Hindi, Bihar news, entertainment news, News in Hindi, Breaking News In Hindi, National News, Technology"}



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.

Stop words have been removed using data from a standard Hindi stopwords list.

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

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
बैंकॉक। भारतीय निशानेबाज जीतू राय ने पूर्व ओलंपिक चैंपियन पांग येइ 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 गृहमंत्री 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

Please select an article to read.

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
बैंकॉक। भारतीय निशानेबाज जीतू राय ने पूर्व ओलंपिक चैंपियन पांग येइ 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 गृहमंत्री 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

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 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

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.

There were some news articles, though, which we were unable to retrieve many relevant videos. The reason for this is the presence of rhetorics or style in the title of the article, for example: *"टी-20 विश्व कप : गेंद अब पीसीबी के पाले में"* is the title of one such article. Note that the title does not mention what the news is about in a direct manner, but uses a metaphor to describe it. Since different sources might explain the news in different manner, it is not possible to match this metaphor with another indirectly explained source.

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