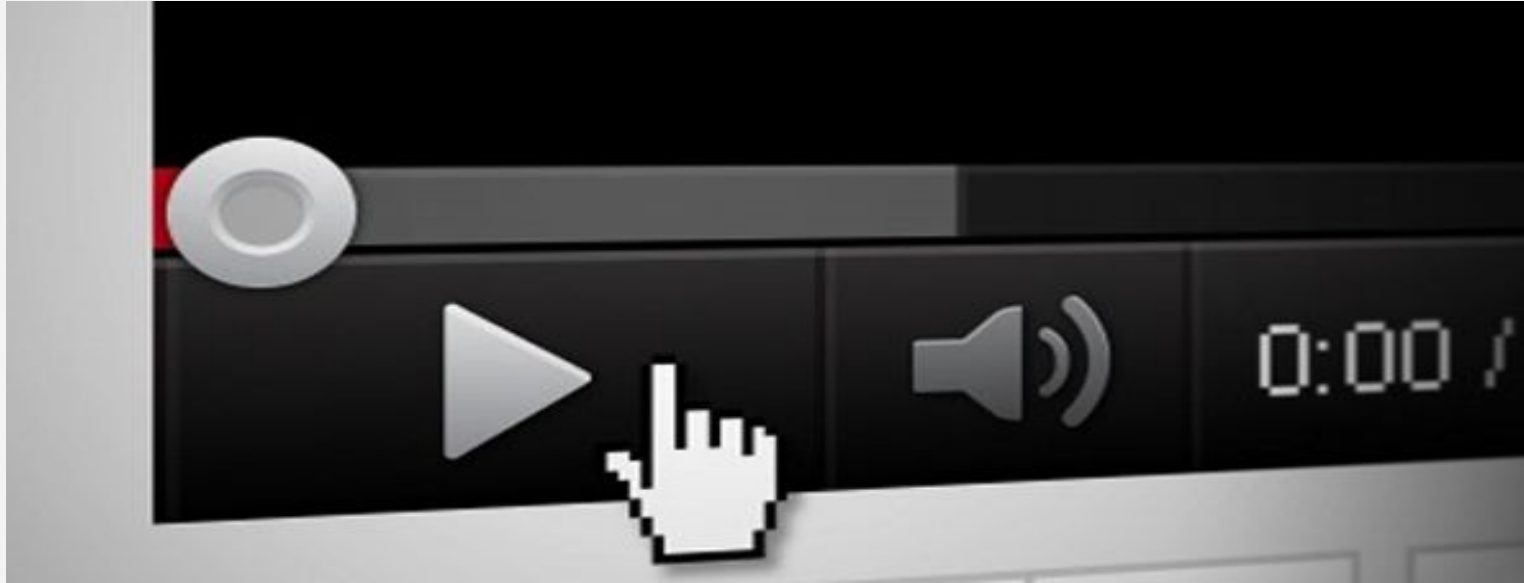


# YouTube Video Extraction



**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.

# Experimental Setup

- We were provided dataset containing 30,000 news articles in the given dataset, all of them in Hindi, provided in JSON format for easy and efficient parsing.
- Initially, we used the keywords that were already provided in JSON format in the dataset. A query was formulated using them for YouTube's search interface.
- We also used the keywords from the url of the article given in the dataset itself.

# Experimental Setup

- 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.
- For ranking the retrieved videos, we extracted keywords from the article text, performing stopwords removal on them, and compared them with the tags and description of the videos.

# Key Challenges

- How to get videos around the article published time and rank them in terms of relevancy.
- 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 used to overcome the challenges

- How to get videos around the article published time and rank them in terms of relevancy?

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.

# Approach used to overcome the challenges

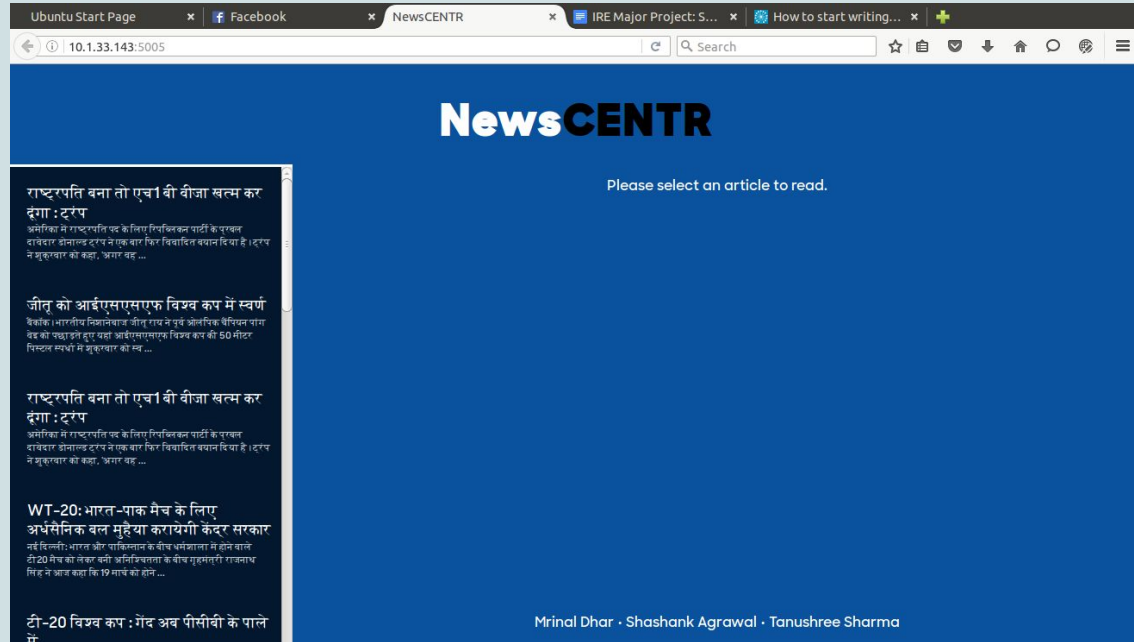
- How to fit it in Indian Language case, the common case is that the article will be in Indian language while the video might be in English?

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. 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.

# Demonstration of Results

We hosted a server in which a user can select an article of his/her choice. The application will then display top 10 relevant Youtube videos.

User is prompted to select an article to read =>





# News article corresponding to the article selected by user

Ubuntu Start Page

Facebook

NewsCENTR

IRE Major Project: S...

How to start writing...

+

10.1.33.143:5005

Search

NewsCENTR

**राष्ट्रपति बना तो एच1वी वीजा खत्म कर दूंगा : ट्रंप**  
अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रमुख 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**  
बेकोंक। भारतीय निशानेबाज जीतू राय ने पूर्व ओलंपिक चैंपियन वांग वेइ को पछाड़ते हुए यहां आईएसएसएफ विश्व कप की 50 मीटर पिस्टल स्पर्धा में शुक्रवार को स्व ...

**राष्ट्रपति बना तो एच1वी वीजा खत्म कर दूंगा : ट्रंप**  
अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रमुख 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 गृहसूत्री राजनाथ सिंह ने आज कहा कि 19 मार्च को होने ...

**टी-20 विश्व कप : गेंद अब पीसीवी के पाले में**

अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के प्रमुख दावेदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुक्रवार को कहा, 'अगर वह राष्ट्रपति बन गए तो वह एच1वी वीजा कार्यक्रम को खत्म कर देंगे।' हालांकि, अपने इस बयान से एक घंटा पहले उन्होंने इसके उलट कहा था कि अमेरिका को उच्च कौशल वाले विदेशीकर्मियों की जरूरत है। रिपब्लिकन पार्टी द्वारा आयोजित एक परिचर्चा में ट्रंप ने कहा, एच1वी वीजा न तो उच्च कौशल के लिए है और न ही आव्रजन के लिए। इससे अस्थायी विदेशी कर्मी आयात किए जाते हैं और उसका साफ-साफ मकसद अमेरिकियों की जगह सन्ते में काम करने वाले लोगों को लगाना। यही, नहीं उन्होंने बाहर से पढ़ने आने वाले छात्रों पर निशाना साधा। उन्होंने कहा, हमारी एक सबसे बड़ी समस्या है कि लोग हमारे सबसे अच्छे कॉलेजों में पढ़ने आ जाते हैं। वे हार्वर्ड जाएंगे, स्टैनफोर्ड जाएंगे, वार्टन जाएंगे। जैसे ही उनकी पढ़ाई पूरी हुई उन्हें बाहर कर दिया जाएगा। फ्लोरिडा से आने वाले सीनेट के सदस्य माका े रुबियो ने ट्रंप के इस बयान की आलोचना की है। उन्होंने ट्रंप के विदेश नीति पर सवाल उठाते हुए कहा कि रिपब्लिकन उम्मीदवार को इस मुद्दे सहित कई मुद्दों का ज्ञान ही नहीं है। एच1वी वीजा भारतीय सॉफ्टवेयर पेशेवरों में बड़ा लोकप्रिय है। अमेरिकी कंपनियां इसी अल्प कालिक वीजा पर भारतीयों को काम के लिए आमंत्रित करती हैं। राष्ट्रपति पद की डेमोक्रेटिक उम्मीदवारी की प्रमुख दावेदार हिलेरी क्लिंटन ने ट्रंप के नफरत से भरे नजरिये की आलोचना करते हुए उन्हें भयावह करार दिया हिलेरी ने कहा, जिस तरह से वह अपनी सहूलियत के हिसाब से महिलाओं, अश्वेतों और सभी देशों का अपमान करते हैं। मुझे उससे नफरत है। अमेरिकियों को प्रभावित करने वाले जटिल मुद्दों पर उनमें समझ की भारी कमी है। डोनाल्ड ट्रंप ने कहा है कि अमेरिका का अफगानिस्तान में रहना 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

Mrinal Dhar · Shashank Agrawal · Tanushree Sharma

# List of Youtube videos extracted

Ubuntu Start Page x Facebook x NewsCENTR x IRE Major Project: S... x How to start writing... x

10.1.33.143:5005 Search

## NewsCENTR

**राष्ट्रपति बना तो एच।बी.वी.जा खत्म कर दूंगा : ट्रंप**  
अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के पूर्वकान दाविदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुरुवार को कहा, 'अगर वह ...

**जीतू को आईएसएसएफ विश्व कप में स्वर्ण**  
बेकोक। भारतीय निशानबाज जीतू राय ने पूर्व ओलंपिक बेरियन पांग ब्रेड को पछाड़ते हुए यहां आइएसएसएफ विश्व कप की 50 मीटर निम्नतल स्पर्धा में शुरुवार को स्व ...

**राष्ट्रपति बना तो एच।बी.वी.जा खत्म कर दूंगा : ट्रंप**  
अमेरिका में राष्ट्रपति पद के लिए रिपब्लिकन पार्टी के पूर्वकान दाविदार डोनाल्ड ट्रंप ने एक बार फिर विवादित बयान दिया है। ट्रंप ने शुरुवार को कहा, 'अगर वह ...

**WT-20: भारत-पाक मैच के लिए अर्धसैनिक बल मुहैया करायेगी केंद्र सरकार**  
नई दिल्ली: भारत और पाकिस्तान के बीच भूमजाला में होने वाले टी-20 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**



**Trump's Softened Stance on Visas...**



**Obama's immigration plan doesn't ...**



**American Jobs Lost to H1B Visa - ...**



**Hillary Clinton Reaffirms Call for Sil...**

Mrinal Dhar · Shashank Agrawal · Tanushree Sharma

# 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 DropBox **shared folder**: <https://goo.gl/k9lstN>

**Thank You**