

# You**Tube** Video Extraction

---

Saurabh Jain (201301128)

Manohar Hari (201505551)

Sree Kavitha Parupalli (201356194)

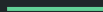
# Problem Statement

The aim of the project is to extract videos related to a news article from YouTube. The videos can be directly uploaded by the news publisher or a user sharing the video about that event. The news article will be given in JSON format. It is done for both Hindi and English news articles.

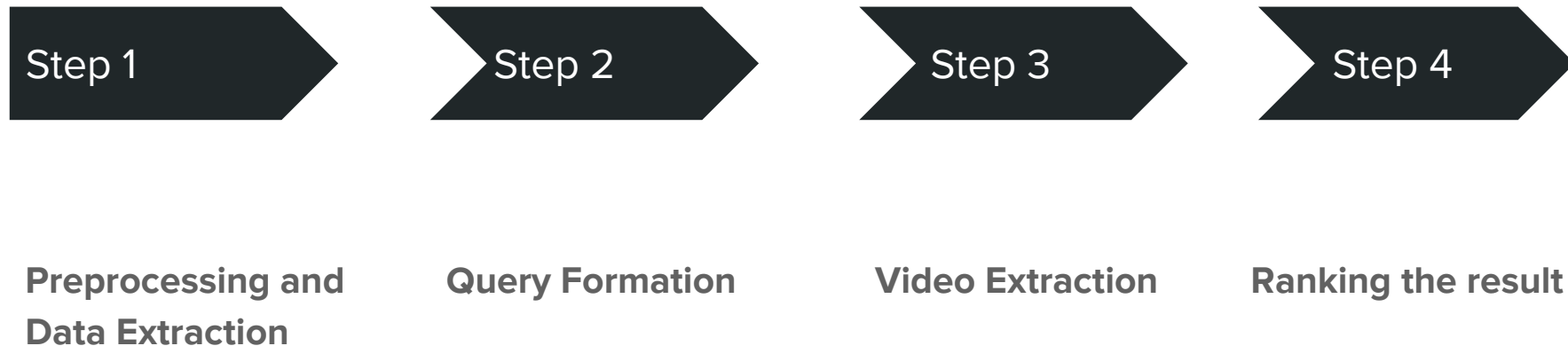


# The Solution

1. Data Extraction from News Article
2. Queries formation
3. YouTube Video Extraction based upon generated queries
4. Ranking the extracted videos



# How It Works



# Data Extraction

Textblob api is used to extract the following data from the text of the news article

1. Noun Phrases
2. Sentiment Analysis
3. Synsets

# Query Formation

Steps taken for query formation :

1. Getting the title as the base query.
2. Creating inverted index of the text data.
3. Retaining the terms in the title with higher idf value in the text and removing the ones with lower value. A phrase being considered as a single term.
4. Creating multiple queries using n-grams.
5. Fusing the formed queries with synsets if relevant result not found.

# Video Extraction

All the queries formed in the previous step are used to search the YouTube and the urls of all the resulting videos are stored.

Titles are extracted from the video urls using BeautifulSoup library.

Thumbnails of the videos are extracted from their respective video id.

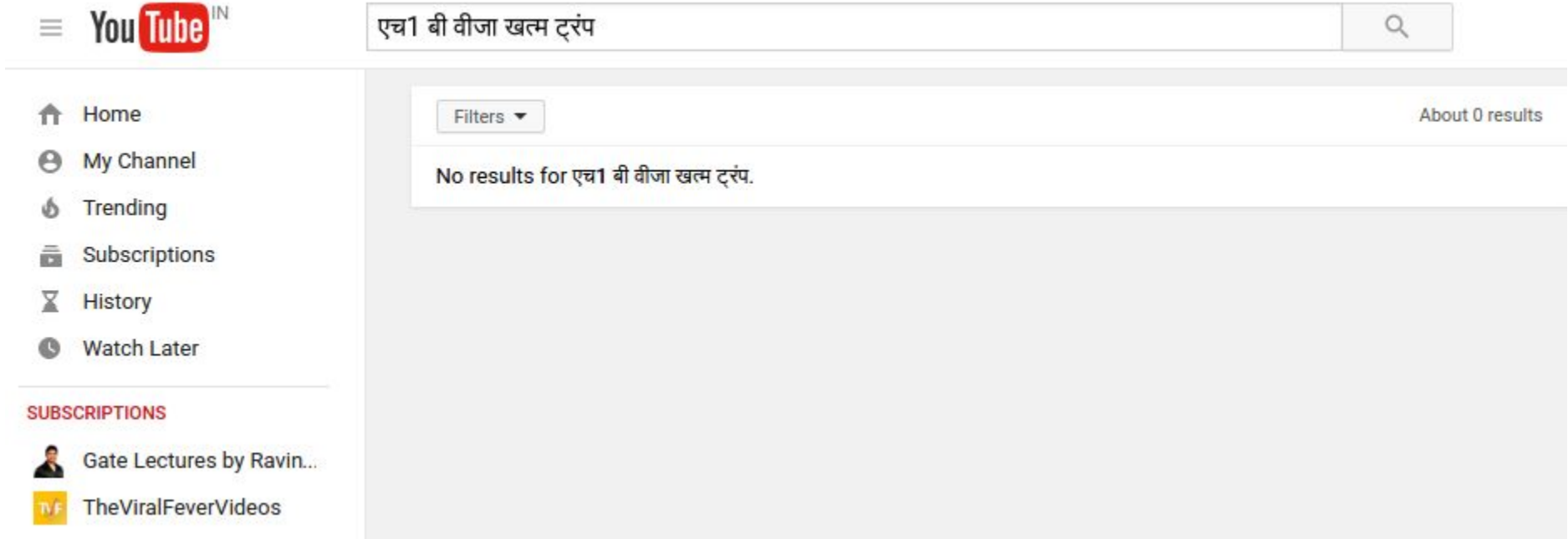
# Ranking the videos

The extracted videos were ranked using the following algorithm :


1. A rank value of the video is computed by finding the similarity between the query term and the title of the video.
2. Noun terms in title are given more weightage.
3. Titles of videos in Hindi are given more priority than English ones if the news article is in Hindi.



# Example





# Example



- Home
- My Channel
- Trending
- Subscriptions
- History
- Watch Later

**SUBSCRIPTIONS**


-  Gate Lectures by Ravin...
-  TheViralFeverVideos

- Browse channels
- Manage subscriptions


एच1 बी वीजा ट्रंप

Filters


About 1 results




अमेरिका में फर्जी वीजा मामले में **10** भारतीय गिरफ्तार  
NyusuDigital Media Pvt Ltd  
6 days ago • 2 views  
अमेरिकी सरकार ने स्टिंग ऑपरेशन की मदद से एक साल से यूएस में चल रहे एक वीजा...  
**NEW**



डोनाल्ड ट्रम्प के एच 1 बी वीजा नीति पर न्यूट **Gingrich**  
Fox News ✓  
1 month ago • 6,255 views  
राष्ट्रपति पद के उम्मीदवार की स्थिति पर एक करीब देखो, पूर्व हाउस स्पीकर 'केली फ्राइल' पर प्रतिक्रिया करते हैं

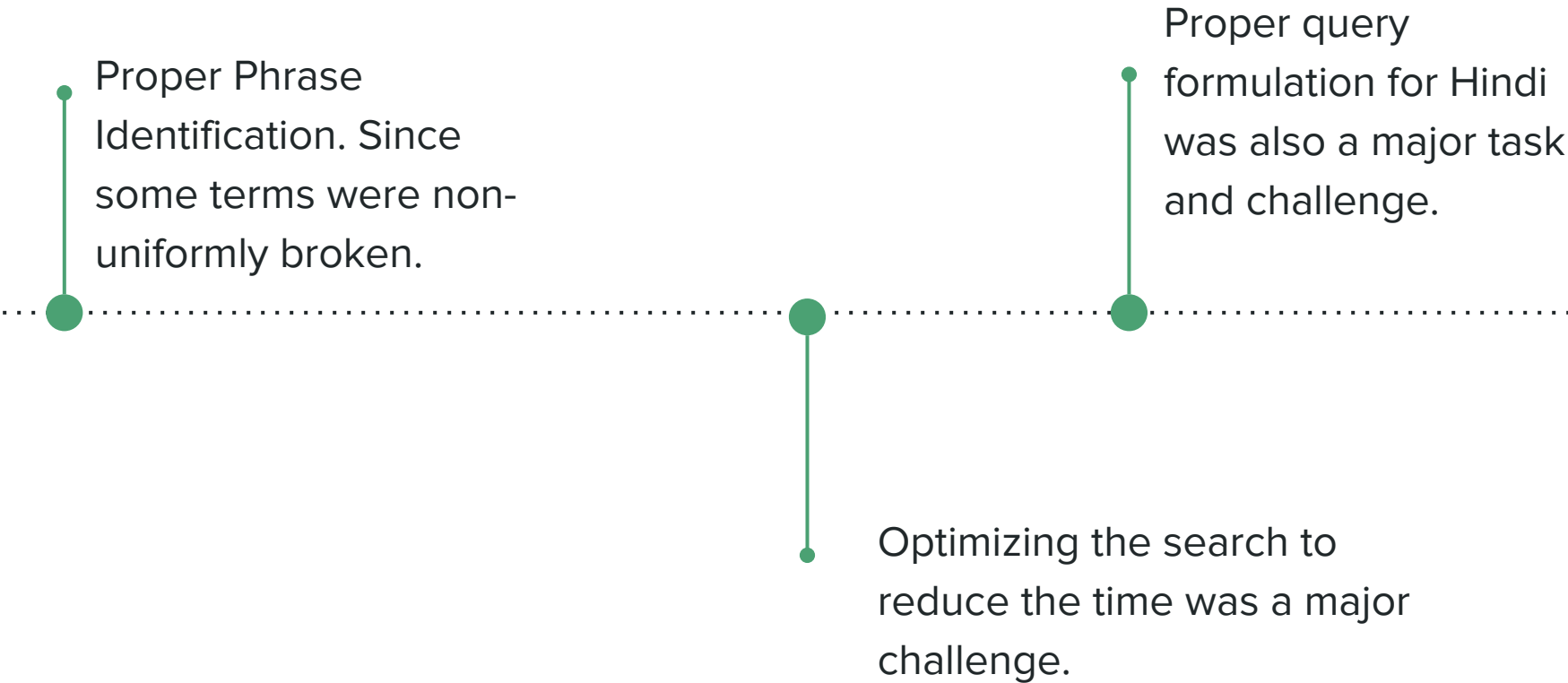


एच 1 बी वीजा खतम डोनाल्ड ट्रम्प के फिलप फ्लॉप उसे विवाद में भूमि  
The Economic Times ✓  
1 month ago • 3,980 views  
इकोनॉमिक टाइम्स एक अंग्रेजी भाषा के भारतीय दैनिक समाचार पत्र वेबसाइट, कोलमैन एंड कंपनी लिमिटेड द्वारा प्रकाशित, अधिक है ...



ट्रम्प: 'मैं बदल रहा हूँ वीजा पर अत्यधिक कुशल श्रमिकों के लिए  
Tea Party 2  
1 month ago • 970 views  
रिपब्लिकन डोनाल्ड ट्रम्प **now** कहते हैं, "हम इस देश में अत्यधिक कुशल लोगों की जरूरत है," तो उन्होंने कहा, "बदल" है और "नरम" उसकी ...

# Challenges Faced



Proper Phrase Identification. Since some terms were non-uniformly broken.

Proper query formulation for Hindi was also a major task and challenge.

Optimizing the search to reduce the time was a major challenge.

## Drawbacks of the System

Given a title and text, if the text in hindi is not consistent while writing, it will not fetch a proper result.

Example: if Title is : एच1 बी वीजा

And the text has the word: एच1बी

These words become different and it would not be fetched as important word, or a noun.

# Future Scopes

1. Can be more optimized.
2. Can be extended to other languages as well.
3. Semantic meaning can be extracted which can be applied on scholarly articles.
4. Videos from other websites can also be included.

A close-up photograph of a person's hand, wearing a dark suit sleeve, adjusting a white slider on a professional audio mixing console. The hand is positioned in the lower-left quadrant, with the index finger and thumb touching the slider. The mixing console has various knobs and sliders, with some red-tipped knobs visible on the right. The background is heavily blurred, showing out-of-focus lights in shades of green, blue, and red, suggesting a stage or concert environment. The text "Thank You" is centered in the middle of the image in a white, sans-serif font.

Thank You