YouTube Video Extraction

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

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 stopword 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.

Thank You