YouTube Video Extraction

Problem Statement

The project is about extracting 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. It is to be done for English as well as Indian Language news articles.

Applications

Today many news channels are on YouTube publishing videos every day covering all the important events of the day. Many times Videos present more interesting information and highly engaging for news readers rather than an article. So we create a platform where a user can upload a news article (JSON format), and will receive links and thumbnails of relevant YouTube videos in a properly ranked order. The news article can be in an Indian language too. We chose Hindi for this project as it is the most widely used Indian language.

Challenges

1. **Semantic Extraction from the news article**

   Different words may mean the same and same words may mean different based on the context. Understanding the context and applying proper semantic extraction is a big challenge.

2. **Phrase Identification**

   Phrase Identification is important to make general words more specific. For example blood -> blood hound, blood test, blood brother, etc. Also identification of famous quotes as a single unit is necessary.

3. **Named and Place Entity Recognition**

   Name and place entity recognition is very important for news articles. Most of the news that we read is always about a person, group of people or organization or about an event at a particular place. Thus name and place recognition is important.

4. **Anaphora and co-references**

   It often happens in news articles that a particular person being called by pronoun eg “Pranab Mukherjee” vs “Honorable President of India”. Dealing with such things is in itself a complex and challenging problem.

5. **NLP on rich Hindi Language**

   Hindi is a very rich language with a very big set of grammar rules. Applying NLP and extracting data from Hindi language is a challenge.

6. **Ranking the relevant links**

   Once we have the relevant links with us, ranking them in proper order is a challenging task as there are a lot of parameters to consider. Some of them may include title, description, tags, static score based on trustworthiness of particular channels, timestamp, video quality, etc. Giving proper weightage to these parameters is a challenge.

Second Deliverable Details

In the second deliverable we had planned to complete the extraction from news article (basic NLP) for both English and Hindi and yielding the relevant links in properly ranked order.

Third Deliverable Details

We would apply advanced NLP techniques over the news article (some of them mentioned above in challenges) and improve our ranking function by considering more parameters. We would also try to incorporate the whole project in a proper end to end user web app.

Tools to be used

1. Django (Python)
2. NLTK
3. Indic NLP Library
4. More tools will be added as the project progresses

References

1. http://nlp.stanford.edu/software/CRF-NER.shtml
2. http://nlp.stanford.edu/software/dcoref.shtml
3. http://www.nltk.org/
4. http://backlinko.com/how-to-rank-youtube-videos
5. http://newspaper.readthedocs.org/en/latest/
6. http://knackforge.com/blog/selvam/auto-summarizing-news-articles-using-natural-language-processing-nlp

Team 19

Saurabh Jain (201301128)
Manohar Hari (201505551)
Sree Kavitha Parupalli (201356194)