Amrita School of Engineering, Bengaluru
Dept of Computer Science & Engineering
B. Tech (CSE)-VII Sem (2017-2021 Batch)
Project Phase – I | Mid Term Report

YouTube Video Content Analysis and Summarization Automation

Team members:

Register No.	**Name**	**Section**
BL.EN.U4CSE17523	T V Premchand	C
BL.EN.U4CSE17524	Y V Sravan Kumar Reddy	C
BL.EN.U4CSE17551	T Yeswanth Sai	C

Project Advisor	Ms. Sasikala T.
Panel Members (No: 5)	Ms. Sangita Khare Ms. Thangam S.

INDEX

1. INTRODUCTION

1.1 Abstract

A software program will take the YouTube video link of the user's choice as input. And there are two types of content we are going to extract as; one is the summarized text content from the video using Natural Language Processing, and the second form of content is details of the YouTube video like; the name of the video, views count, the category under which the video had shared, etc. using web scraping process. This allows the user to save his time and improve efficiency in information extraction from an unstructured data source.

2. PROBLEM DEFINITION

To develop a software program that will allow the user to save his/her time and improve the efficiency in information extraction and analyzing the unstructured data sources.

Unstructured Data Source: YouTube Videos.

3. MOTIVATION

YouTube, one of the most significant online video sharing and streaming applications. We spend some noticeable amount of our time watching YouTube videos every single day, be it for education, sports, entertainment, or exploring our interests. Moreover, we know that we are watching it for information.

3.1 Aim

We want to develop a software program that automates the process of information extraction from a video and of the video.

3.2 Objective

The main objectives of the project are as follows:

a) To extract most from an unstructured date source: YouTube Videos
b) Prepare a well-structured data for further analysis
c) Extract the text from the audio connected video
d) Provide a detailed abstract/summary for the content extracted from the video source

4. PROJECT SPECIFICATIONS

The project will run in the software environment:

i. Windows (64 bit) 8.1 or higher
ii. Linux (latest version)

Programming Language: Python 3.6 and later versions

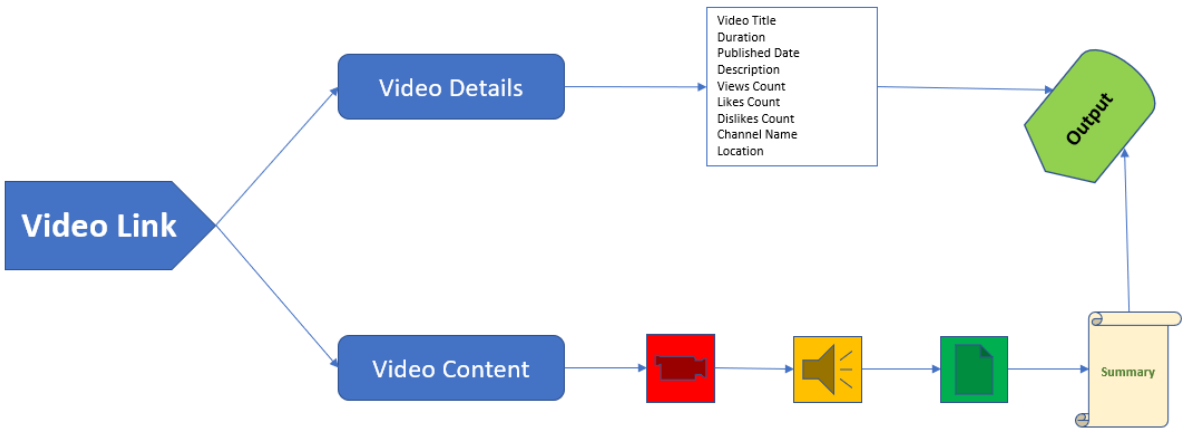
Chrome web driver: Chrome Version 87 or higher

Python Libraries:

- BeautifulSoup
- Contextlib
- OS
- Time
- Speech Recognition
- Wave
- YouTube_dl
- Selenium (for automation)

5. METHODOLOGY & IMPLEMENTATION

5.1 Frame Work Architecture



5.2 Implementation

The implementation includes the following steps:

1. The system takes YouTube video link as input from the user.
2. From the input taken we have two levels of details to be extracted:
    a. Video details
    b. Video content
3. Video details includes the details of YouTube video like video title, description, etc.
4. Video content includes the content present in the YouTube video.
5. After extracting the video content as text, we will summarize the text and add it to the output already contained i.e. video details.

In this whole implementation process, we will use some of the major techniques and algorithms given below.

- Web Scraping
- Word Sense Disambiguation
- Sentence Ranking Algorithm (Reference from Google's Page Rank Algorithm)
- Speech Recognition
- Speech to text conversion

6. STUDY ON LITERATURE SURVEY

6.1 Research Paper - 1

Title: Comments Scraping Application for Review YouTube Content

Authors: Viny Christinti M., Walda, Tri Sutrisno

Description: In this research they have built a scraping application to obtain comment data on YouTube. YouTube also consists of existing structures in HTML, we can see comments that are visible on the web. But when the source is seen in the form of HTML structure so they have created that application that only takes comments without taking other data that is not needed. They did validation for this comment collected by them for data mining and further analysis. To get most out of the YouTube Videos (Unstructured Data Source), we started collecting more content in detail about the YouTube Video.

6.2 Research Paper – 2

Title: An Overview on Web Scraping Techniques Ans Tools

Authors: Anand V. Saurkar, Kedar G. Pathare, Shweta A. Gode.

Description: In this research they've used several Web Scraping software tools like Mozenda, Visual web ripper, web content extractor, imort.io, scrapy, etc. which were very simple but with limited data mining extension for facilizing online extraction of data for researcher in the format of spreadsheets. BeautifulSoup is one of the Python libraries where we can use HTML Parsing technique and use the markup data as user needs. Scrapy gives a result in the form of spreadsheet for a website which we select for scraping.

6.3 Research Paper – 3

Title: Web Information Retrieval Using Python and BeautifulSoup

Authors: Pratiksha Ashiwal, S. R. Tandan, Priyanka Tripathi, Rohit Miri

Description: In this research paper they used Python programming language and BeautifulSoup for extracting the web information. They have directly parsed the HTML page using the scrapers built based on BeautifulSoup and Python which results in missing some dynamic data for the target information from the HTML data in the domain. To make the scraper work more efficiently, downloading the scraped markup language text in to local file and then parsing it for data pre-processing is the better idea.

6.4 Research Paper – 4

Title: Automatic Text Summarization Using Natural Language Processing

Authors: Pratibha Devihosur, Naseer R

Description: In this research paper they've built an unsupervised learning system. The process for summarizing the text includes four main steps as data input, data pre-processing, using Lesk's algorithm, and generating summary. Along with this they've used the wordnet as an online semantic lexicon. The disadvantages of using this method is disambiguating a sentence which have multiple ambiguous words in the same sentence and for disambiguating the sense of the word they are not disambiguating the other context words.

6.5 Research Paper – 5

Title: Text Summarization using Sentence Scoring Method

Authors: T. Sri Rama Raju, Bhargav Allarpu

Description: In this research they've used sentenced scoring method which involves pre-processing, word frequencies and sentence ranking. The user can specifically choose how many summary points can be taken from the whole context. The process involves tokenization, stemming, frequencies of words, etc. To improve efficiency in pre-processing we additionally used lemmatization which will reduce the task of finding frequencies of every word in the text collection.

7. CONCLUSION

After performing the web scraping, text extraction and we will generate the multi-point summary based on the text extracted from the YouTube video. By using python libraries, we will take the structured output in to document file.

8. REFERENCES

[1] IRJET Volume: 04 Issue: 08 | Aug -2017, **"Automatic Text Summarization Using Natural Language Processing",** Pratibha Devihosur1, Naseer

[2] IRJET Volume: 04 Issue: 04 | Aug -2017, **"Text Summarization using Sentence Scoring Method",** T. Sri Rama Raju, Bhargav Allarpu

[3] International Journal on Future Revolution in Computer Science & Communication Engineering, Volume: 4, Issue: 4, **"An Overview on Web Scraping Techniques Ans Tools"**, Anand V. Saurkar, Kedar G. Pathare, Shweta A. Gode

[4] iJRASET Volume: 4, Issue: VI, June 2016, **"Web Information Retrieval Using Python and BeautifulSoup"**, Pratiksha Ashiwal, S.R.Tandan , Priyanka Tripathi , Rohit Miri

[5] IOP Conf. Series: Materials Science and Engineering, **"Comments Scraping Application for Review YouTube Content"**, Viny Christinti M., Walda, Tri Sutrisno