FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

Department of Computer Engineering

Course, Subject & Experiment Details

Practical No:	2
Title:	Data Collection-Select the social media platforms of your choice (Twitter, Facebook, LinkedIn, YouTube, Web blogs etc), connect to and capture social media data for business (scraping, crawling, parsing).
Name of the Student:	Warren Fernandes
Roll No:	8940
Date of Performance:	31/01/2023
Date of Submission:	07/02/2023

Evaluation:

Sr. No.	Rubric	Grade
1	On time submission/completion (2)	
2	Preparedness (2)	
3	Skill (4)	
4	Output (2)	

Signature of the Teacher

Youtube Data Collection

The interesting story of our educational system | Adhitya Iyer | TEDxCRCE - https://www.youtube.com/watch?v=Fw1Fc_y_2Ek

Steps: **Get API Key** - Go to the *Google Cloud console • (h.ÜuZL.nn.åQle.clQu1gQQgleumL) and sign in with your Google Account

Click the project drop-down menu in the top bar and select or create the project you want to use.

Click the hamburger menu in the top left and select APIs & Services >Select Youtube Data V3 API> Credentials.

Click the Create credentials button and select API key.

The API key will be displayed in a pop-up window. You can click the RESTRICT KEY button to restrict the API keys usage, such as by IP address or referrer.

Click the COPY button to copy the API key to your clipboard.

You can use the API key in your application to access the Google Cloud APIs. Be sure to keep the API key confidential, as it can be used to access your Google Cloud resources.

```
import requests

video_id = "Fw1Fc_y_2Ek"

api_key = "AIzaSyDgqBDVRZ_gl1Z46qWgEJembokeMTPjrrY"

# Retrieve video information
video_info_url = f"https://www.googleapis.com/youtube/v3/videos?part=snippet&id={video_id}&key={api_key}"
video_info_response = requests.get(video_info_url)
video_info_data = video_info_response.json()

video_info_data
```


```
{'kind': 'youtube#videoListResponse',
 'etag': 'k-DLvT03J8lym6u5lYyfVLsKvFU',
 'items': [{'kind': 'youtube#video',
   'etag': '33VDJhtOgcwlAEVBnNTD49e3H2Y',
   'id': 'Fw1Fc_y_2Ek',
   'snippet': {'publishedAt': '2016-08-17T18:38:44Z',
    'channelId': 'UCsT0YIqwnpJCM-mx7-gSA4Q',
    'title': 'The interesting story of our educational system | Adhitya Iyer | TEDxCRCE',
    'description': "A story surrounding more than a million students in India built around a booming IT industry, a machine like education system, a race to a seat amongst the top engineering colleges and student suicides in our system. The Indian education system in all its complexity has turned out to be one of the most fascinating educational stories in the world. Watch Adhitya Iyer in his talk share his travel around the country discovering the truth behind the common engineer in India.\n\nAdhitya Iyer completed his engineering degree from Mumbai University, where he studied in an institute interestingly named after Sardar Patel and was inaugurated by Nehru. He was listed as one of India's Top 30 student entrepreneurs by the NEN. He later spent 2 years in Bangalore selling chai, before he set out to write a non-fiction on the complex life of India's engineers is Kickstarter funded and is India's highest crowdfunded book. He thinks the society needs to be re-imagined and that humans can't be spending more than half their life doing something they seldom enjoy.\n\nThis talk was given at a TEDx event using the TED conference format but independently organized by a local community. Learn more at http://ted.com/tedx",
    'thumbnails': {'default': {'url': 'https://i.ytimg.com/vi/Fw1Fc_y_2Ek/default.jpg',
      'width': 120,
      'height': 90},
     'medium': {'url': 'https://i.ytimg.com/vi/Fw1Fc_y_2Ek/mqdefault.jpg',
      'width': 320,
      'height': 180},
     'high': {'url': 'https://i.ytimg.com/vi/Fw1Fc_y_2Ek/hqdefault.jpg',
      'width': 480,
      'height': 360},
     'standard': {'url': 'https://i.ytimg.com/vi/Fw1Fc_y_2Ek/sddefault.jpg',
      'width': 640,
      'height': 480},
     'maxres': {'url': 'https://i.ytimg.com/vi/Fw1Fc_y_2Ek/maxresdefault.jpg',
      'width': 1280,
      'height': 720}},
    'channelTitle': 'TEDx Talks',
    'tags': ['TEDxTalks',
     'English',
     'India',
     'Education',
     'Change',
     'Education reform',
     'Students'],
    'categoryId': '29',
    'liveBroadcastContent': 'none',
    'localized': {'title': 'The interesting story of our educational system | Adhitya Iyer | TEDxCRCE',
     'description': "A story surrounding more than a million students in India built around a booming IT industry, a machine like education system, a race to a seat amongst the top engineering colleges and student suicides in our system. The Indian education system in all its complexity has turned out to be one of the most fascinating educational stories in the world. Watch Adhitya Iyer
```

```
in his talk share his travel around the country discovering the truth behind the common engineer in India.\n\nAdhitya Iyer
completed his engineering degree from Mumbai University, where he studied in an institute interestingly named after Sardar Patel
and was inaugurated by Nehru. He was listed as one of India's Top 30 student entrepreneurs by the NEN. He later spent 2 years in
Bangalore selling chai, before he set out to write a non-fiction on the complex life of India's engineers is Kickstarter funded
and is India's highest crowdfunded book. He thinks the society needs to be re-imagined and that humans can't be spending more than
half their life doing something they seldom enjoy.\n\nThis talk was given at a TEDx event using the TED conference format but
independently organized by a local community. Learn more at http://ted.com/tedx"},
    'defaultAudioLanguage': 'en'}}],
 'pageInfo': {'totalResults': 1, 'resultsPerPage': 1}}
```

```
# Retrieve video compnts
comments_url = f"https://www.googleapis.com/youtube/v3/commentThreads?part=snippet&videoId={video_id}&key={api_key}"
comments_response= requests.get(comments_url)
comments_data = comments_response.json()

comments_data
```

```
          'authorChannelUrl': 'http://www.youtube.com/channel/UCoRKNk0I51fs-6bnK_CHA6g',
          'authorChannelId': {'value': 'UCoRKNk0I51fs-6bnK_CHA6g'},
          'canRate': True,
          'viewerRating': 'none',
          'likeCount': 0,
          'publishedAt': '2022-10-12T14:23:15Z',
          'updatedAt': '2022-10-12T14:23:15Z'}},
       'canReply': True,
       'totalReplyCount': 0,
       'isPublic': True}},
     {'kind': 'youtube#commentThread',
      'etag': '5z9a-bDrgfijLlqXmPCAnhPJ5WM',
      'id': 'UgwQ_z8jbqQstI4J7ZV4AaABAg',
      'snippet': {'videoId': 'Fw1Fc_y_2Ek',
       'topLevelComment': {'kind': 'youtube#comment',
        'etag': 'L62dohS9RCNmlNzmPSF0VH4W6jg',
        'id': 'UgwQ_z8jbqQstI4J7ZV4AaABAg',
        'snippet': {'videoId': 'Fw1Fc_y_2Ek',
         'textDisplay': 'thanks',
         'textOriginal': 'thanks',
         'authorDisplayName': 'Anupam Bhise',
         'authorProfileImageUrl': 'https://yt3.ggpht.com/ytc/AL5GRJWRZEkUblUbPA7AlsjfWccS4wZK0UBNWTeOBlCRcA=s48-c-k-c0x00ffffff-no-
rj',
         'authorChannelUrl': 'http://www.youtube.com/channel/UCpGh6rJ4Pwl2jivGsGuillQ',
         'authorChannelId': {'value': 'UCpGh6rJ4Pwl2jivGsGuillQ'},
         'canRate': True,
         'viewerRating': 'none',
         'likeCount': 0,
         'publishedAt': '2022-09-30T05:30:44Z',
         'updatedAt': '2022-09-30T05:30:44Z'}},
       'canReply': True,
       'totalReplyCount': 0,
       'isPublic': True}},
     {'kind': 'youtube#commentThread',
      'etag': 'cnwvO3uPSu5C4goTA759khlZRBo',
      'id': 'UgwkwKuV1dOMSg_hFAZ4AaABAg',
      'snippet': {'videoId': 'Fw1Fc_y_2Ek',
       'topLevelComment': {'kind': 'youtube#comment',
        'etag': '6wgPgHpp6RYMEYO-pk9POswuIUM',
        'id': 'UgwkwKuV1dOMSg_hFAZ4AaABAg',
        'snippet': {'videoId': 'Fw1Fc_y_2Ek',
         'textDisplay': 'Engineer by only option not by choice !!!!!<br>That&#39;s what he meant by saying <br>His biggest mistake
was that he was good at studies',
         'textOriginal': "Engineer by only option not by choice !!!!!\nThat's what he meant by saying \nHis biggest mistake was
that he was good at studies",
         'authorDisplayName': 'CLARITY-SPREADER',
         'authorProfileImageUrl': 'https://yt3.ggpht.com/sD3m74aYzdAwJmcGqBR9tyilmwADVosm_2evziBHJfAu37oBM_PQlxa2-
PUyzL0a8RCKTzCRgQ=s48-c-k-c0x00ffffff-no-rj',
         'authorChannelUrl': 'http://www.youtube.com/channel/UCjCqTrBzSpuKwZpxyEmNgEA',
         'authorChannelId': {'value': 'UCjCqTrBzSpuKwZpxyEmNgEA'},
         'canRate': True,
         'viewerRating': 'none',
         'likeCount': 0,
         'publishedAt': '2022-09-27T09:01:23Z',
         'updatedAt': '2022-09-27T09:02:16Z'}},
       'canReply': True,
       'totalReplyCount': 0,
       'isPublic': True}}]}
```

```
# Extract the carnents
comments = [item["snippet"]["topLevelComment"]["snippet"]["textOriginal"] for item in comments_data["items"]]

comments
```

```
['Engineering give attendance 🤣',
 '*be an engineer first & then decide what to do with life* 💯',
 'Best talk ever Now.',
 'India is affected by British education system',
 'He is real hero I watched first video which is useful after independence',
 'currently i am a jee aspirat',
 'Bhai m kyu padh raha hu mujhe to English aati h  hehe 😁',
```

```
 "His jokes are actually good boomers don't get 😂",
 'I am 17 now I am in 12,Iam a biomaths student,and The only advice I get from elders( uncle, aunty, relatives,..etc) prepare for
NEET or JEE(For to become Doctor or engineer,)Yes I know one thing that Indians mindset is the person who success in their life is
only when they become doctor or engineer.\nNB: Science is my favourite subject,but now Iam not interest in it,so ofcourse I drop
it after I complete my 12 th and change my stream',
 'I am really lucky to have parents who encourage me go to Shantiniketan Vishwa Vidyalaya to study art because I am good at art.',
 "And the reason why I'm not on that stage giving a TED talk like Adhitya Iyyer, is because unlike him I followed my heart right
from school when he was still searching for his. So now, he's an engineer giving a TED talk, and I'm living my life the way I
wanted. 😁 I mean that takes courage... Ask the kids who are in this list not wanting to do the usual but don't know how to do the
otherwise. You've got to be a little rebel, a little more courageous, and little or actually, a lot of faith in yourself. \n\nFor
kids who don't want to follow the norm, please don't. 10 years down the line it won't matter, except that you're happy doing what
you're doing. Honestly, noone would really bother if you're happy, which is exactly why you need to be happy. Cheerio! ❤️",
 "I also felt the same. Being 'good at studies' actually opens up lot of options. When you get lot of options there's only lesser
chance you choose the right option. I told my mom that same dialogue he said 2 years before when I realized this. \nI only wanted
just 2 option or just one way then I would have not get confused. I would have chosen a direct path to get a professional job
(atleast happy job)",
 'I am not good at studies and not in math either but I am an engineer and now I am broke and poor',
 'Nice story.',
 'Bonkers!',
 "You forgot about India's obsession with hard work. We're being taught that hard work makes us rich and successful. And most of
the Indian students' ambition is to become a doctor, engineer or civil service. These subjects require lots of hard work that will
eventually lead you to six digit salary and social Upliftment. So we're designed to dream in a specific way by ignoring our
interests and passion.\nAnd it is for this reason we're giving more importance to maths and science as it requires hard work.",
 'Thumbnail may Arjun kapoor ki pic h',
 'In Israel, it is a must to work in army at a point of time. In india it is a must to work as an Engineer.',
 'thanks',
 "Engineer by only option not by choice !!!!!\nThat's what he meant by saying \nHis biggest mistake was that he was good at
studies"]
```

Sentiment Analysis of the comments

Using TextBlob we will find the polarity of the comment and classify it into three classes ie. Positive, Negative and Neutral.

```
from textblob import TextBlob

def get_comment_sentiment(comment):
  analysis = TextBlob(comment)
  if analysis.sentiment.polarity > 0:
    return "Positive"
  elif analysis.sentiment.polarity == 0:
    return "neutral"
  else:
    return "negative"

comment_list = []
sentiment_list = []
for comment in comments:
  sentiment = get_comment_sentiment(comment)
  comment_list.append(comment)
  sentiment_list.append(sentiment)
  print(f"{comment} : {sentiment}")
```

```
Engineering give attendance 🤣 : neutral
*be an engineer first & then decide what to do with life* 💯 : Positive
Best talk ever Now. : Positive
India is affected by British education system : neutral
He is real hero I watched first video which is useful after independence : Positive
currently i am a jee aspirat : neutral
Bhai m kyu padh raha hu mujhe to English aati h  hehe 😁 : neutral
His jokes are actually good boomers don't get 😂 : Positive
I am 17 now I am in 12,Iam a biomaths student,and The only advice I get from elders( uncle, aunty, relatives,..etc) prepare for NEE
NB: Science is my favourite subject,but now Iam not interest in it,so ofcourse I drop it after I complete my 12 th and change my st
I am really lucky to have parents who encourage me go to Shantiniketan Vishwa Vidyalaya to study art because I am good at art. : Po
And the reason why I'm not on that stage giving a TED talk like Adhitya Iyyer, is because unlike him I followed my heart right from

For kids who don't want to follow the norm, please don't. 10 years down the line it won't matter, except that you're happy doing wh
I also felt the same. Being 'good at studies' actually opens up lot of options. When you get lot of options there's only lesser cha
I only wanted just 2 option or just one way then I would have not get confused. I would have chosen a direct path to get a professi
I am not good at studies and not in math either but I am an engineer and now I am broke and poor : negative
Nice story. : Positive
Bonkers! : neutral
You forgot about India's obsession with hard work. We're being taught that hard work makes us rich and successful. And most of the
And it is for this reason we're giving more importance to maths and science as it requires hard work. : Positive
Thumbnail may Arjun kapoor ki pic h : neutral
In Israel, it is a must to work in army at a point of time. In india it is a must to work as an Engineer. : neutral
thanks : Positive
Engineer by only option not by choice !!!!!
That's what he meant by saying
His biggest mistake was that he was good at studies : Positive
```

```
import pandas as pd

sentiment_df = pd.DataFrame({"Comments": comment_list,"Sentiment": sentiment_list})
```

```
sentiment_df.head()
```

	Comments	Sentiment
0	Engineering give attendance 🤣	neutral
1	*be an engineer first & then decide what to do...	Positive
2	Best talk ever Now.	Positive
3	India is affected by British education system	neutral
4	He is real hero I watched first video which is...	Positive

Web Scraping

Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform web scraping to obtain data from websites.

Using Boilerpipe

```
from boilerpipe.extract import Extractor
URL = "https://www.freepressjournal.in/education/pushing-your-limits-engineering-college-students-assemble-ted-talks-in-bandra"
extractor=Extractor(extractor="ArticleExtractor",url=URL)
print(extractor.getText())
```

```
Home Education Pushing your limits: Engineering college students assemble TED talks in Bandra
Pushing your limits: Engineering college students assemble TED talks in Bandra
TEDxCRCE, an independently organized TED event is bringing speakers from various walks of life and cultures together to connect and
Staff Reporter Updated: Tuesday, January 10, 2023, 05:37 PM IST
Follow us on
Mumbai: To ignite the spark in our minds and to further our motto “Ideas worth spreading," Fr. Conceicao Rodrigues College of Engin
TEDxCRCE, an independently organized TED event is bringing speakers from various walks of life and cultures together to connect and
Read Also
Discussions brew over struggles behind success at Mithibai College's 'Coffee with Colosseum
TEDxCRCE will be held on January 13, 4 pm, at the Samvaad Auditorium of the host college in Bandra, where speakers will try to shed
The theme for the conference is ‘Bending Rules’, defining a concept that encourages creativity. Today, almost every task has predef
Among the key speakers are, Tannaz Irani, Actress, and Nakash Aziz, a Music Composer. Tannaz is known for her roles in movies and i
(If you have a story in and around Mumbai, you have our ears, be a citizen journalist and send us your story  here . )
(To receive our E-paper on WhatsApp daily, please click here.    To receive it on Telegram, please click here . We permit sharing o
RECENT STORIES
```

Using FeedParser

```
import feedparser
FEED_URL="http://feeds.feedburner.com/oreilly/radar/atom"
fp=feedparser.parse(FEED_URL)
for e in fp.entries:
  print(e.title)
```

```
Automating the Automators: Shift Change in the Robot Factory
Digesting 2022
Radar Trends to Watch: January 2023
What Does Copyright Say about Generative Models?
Radar Trends to Watch: December 2022
AI’s ‘SolarWinds Moment’ Will Occur; It’s Just a Matter of When
Technical Health Isn’t Optional
Healthy Data
Formal Informal Languages
Radar Trends to Watch: November 2022
What We Learned Auditing Sophisticated AI for Bias
The Collaborative Metaverse
What Is Hyperautomation?
Radar Trends to Watch: October 2022
The Problem with Intelligence
```

Web Parsing

Data parsing is the process of transforming a sequence (unstructured data) into a tree or parse tree (structured data) that’s easier to read, understand and use.

```
import requests
import lxml
from bs4 import BeautifulSoup
```

```
url = "https://tedxcrce.com/speakers.html"
headers = {
  'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/63.0.3239.132 Safari/537.36 QIHU 360SE
}
f = requests.get(url, headers = headers)
movies_lst = []
soup = BeautifulSoup(f.content, 'lxml')
speaker_box = soup.find_all("div", class_="speaker-box")
speakers=[]
for speaker in speaker_box:
  speaker_name = speaker.find("span", class_="text-uppercase")
  speakers.append(speaker_name.text.strip())
for speaker in speakers:
  print(speaker)
```

```
Prateek Sethi
Dr. Ashok Johari
Tannaz Irani
Raj Aditya Kapoor
Nakash Aziz
Pooja Taparia
Ajay Prabhakar
Aruna Varanasy
Raveena Tandon
Mark Manuel
Gaurav Kotian
Kavita Seth
Dr. Prashant Warier
Siddharth Randeria
Vibhas Sen
Rithvik Dhanjani
Ankur and Akshay
Col Subin Balakrishnan
Subarna Ghosh
Ramesh Sippy
Adhitya Iyer
Prabir Jha
Sameer Ganpathy
Kalpana Morparia
Sarah Jane Dias
Triveni Acharya
```

Web Crawling

Web crawling is a powerful technique to collect data from the web by finding all the URLs for one or multiple domains. A web crawler starts with a list of URLs to visit, called the seed. For each URL, the crawler finds links in the HTML, filters those links based on some criteria and adds the new links to a queue. All the HTML or some specific information is extracted to be processed by a different pipeline.

```
from bs4 import BeautifulSoup
import requests

pages_crawled = []


def crawler(url):
    print(url)
    page = requests.get(url)
    soup = BeautifulSoup(page.text, 'html.parser')
    links = soup.find_all('a')

    for link in links:
        print(link['href'])
        if 'href' in link.attrs:
            if link['href'].startswith('/wiki') and ':' not in link['href']:
                print(link['href'])
                if link['href'] not in pages_crawled:
                    new_link = f"https://en.wikipedia.org{link['href']}"
                    pages_crawled.append(link['href'])
                    # try:
                    #     with open('data.csv', 'a') as file:
                    #         file.write(f'{soup.title.text}; {soup.h1.text}; {link["href"]}\n')
                    #     crawler(new_link)
                    # except:
                    #     continue


crawler('https://www.freepressjournal.in/')
```

https://marathi.freepressjournal.in/nation/asaram-bapu-gets-life-term-imprisonment-in-rape-case
https://marathi.freepressjournal.in/nation/union-budget-2023-fm-nirmala-sitharaman-tables-economic-survey
https://marathi.freepressjournal.in/maharashtra/mpsc-student-protest-successful
https://marathi.freepressjournal.in/
https://www.freepressjournal.in/
https://www.facebook.com/FreePressJournal/
https://twitter.com/fpjindia?lang=en
https://www.linkedin.com/company/the-free-press-journal-newspaper
https://www.instagram.com/freepressjournal/?hl=en
https://www.youtube.com/channel/UCM1VqjE2rshkzKfp4zRVwKA
https://www.freepressjournal.in/analysis
https://www.freepressjournal.in/mumbai
https://www.freepressjournal.in/mumbai
https://www.freepressjournal.in/indore
https://www.freepressjournal.in/bhopal
https://www.freepressjournal.in/delhi
https://www.freepressjournal.in/education
https://www.freepressjournal.in/entertainment
https://www.freepressjournal.in/bollywood
https://www.freepressjournal.in/hollywood
https://www.freepressjournal.in/movie-review-entertainment
https://www.freepressjournal.in/entertainment
https://www.freepressjournal.in/regional-film-news-entertainment
https://www.freepressjournal.in/television-entertainment
https://www.freepressjournal.in/brandsutra
https://www.freepressjournal.in/corporate-corner
https://www.freepressjournal.in/fpj-initiatives
https://www.freepressjournal.in/horoscope
https://www.freepressjournal.in/legal
https://www.freepressjournal.in/science
https://www.freepressjournal.in/spirituality
https://www.freepressjournal.in/sports
https://www.freepressjournal.in/cricket
https://www.freepressjournal.in/sports
https://www.freepressjournal.in/lifestyle
https://www.freepressjournal.in/health
https://www.freepressjournal.in/travel
https://www.freepressjournal.in/food
https://www.freepressjournal.in/
https://www.freepressjournal.in/photos
https://www.freepressjournal.in/tech
https://www.freepressjournal.in/videos
https://www.freepressjournal.in/viral
https://www.freepressjournal.in/weekend
https://www.freepressjournal.in/about-us
https://www.freepressjournal.in/editorial_policy
https://www.freepressjournal.in/careers
https://www.freepressjournal.in/disclaimer
https://www.freepressjournal.in/privacy-policy
https://www.freepressjournal.in/contact-us

Conclusion

Web Scraping, Parsing and Crawling executed successfully.