Programming for AI

Abdul Haseeb

Web Scrapping

- Process to automatically extract the data from websites, web pages or online documents.
- Navigate websites, locate data and store it in a structured format

Web Scrapping Techniques

- **HTML Parsing:**
  - Analyze html structure of a web page to extract data
- **CSS Selectors:**
  - Use CSS selectors to target specific HTML elements
- **JavaScript Rendering:**
  - Executing java scripting code to access dynamic content
- **API Calls:**
  - Accessing website data through official APIs

Web Scrapping Tools

- Beautiful Soup:
  - Popular python library for HTML parsing
  - Used for small scale scrapping on static websites (don’t rely on js)
- Scrapy:
  - Full fledged web scrapping framework in python
  - Used for large scale scrapping
- Selenium:
  - Multi language
  - Automation tool for web browsers
  - Scrapping data from websites which rely heavily on dynamic content (js)

Import Beautiful Soup

- from bs4 import BeautifulSoup

Let’s Read HTML File

- with open(“website.html”) as file:

  contents=file.read()

Creating a BeautifulSoup Instance

- soup=BeautifulSoup(contents, “html.parser”)

Working with soup object

- print(soup.title)
- print(soup.title.string)
- print(soup)

Indented code

- print(soup.prettify())

Accessing the first tag of each type of element

- print(soup.a) #First anchor tag
- print(soup.li) #First li tag
- How to access all the p tags, all the a tags????

Use find_all()

- anchortags=soup.find_all(name=“a”)

- Returns a list of tags

getText()

- for tag in anchortags:

  print(tag.getText())

get

- for tag in anchortags:

  print(tag.get(“href”))

Finding a unique element with a particular id

- headings= soup.find(name=“h1”, class?=“name”)

- You can also search by class using class_ parameter of find.

Selecting using CSS Selectors

- Company_url=soup.select_one(selector=“p a”)

Selecting using CSS Selectors

- Selecting via id
- Company_url=soup.select_one(selector=“#name”)

Selecting using CSS Selectors

- Selecting via class
- Company_url=soup.select(“.heading”)
  - Will select all the tags with a class named heading and will return a list

Scrapping a Live Website

- Let’s head over to:
  - https://news.ycombinator.com/news

Try this code

```
response = requests.get("https://news.ycombinator.com/news")
yc_web_page = response.text

soup = BeautifulSoup(yc_web_page, "html.parser")
articles = soup.find_all(name="a", class_="storylink")
article_texts = []
article_links = []
for article_tag in articles:
    text = article_tag.getText()
    article_texts.append(text)
    link = article_tag.get("href")
    article_links.append(link)

article_upvotes = [score.getText() for score in soup.find_all(name="span", class_="score")]


print(article_texts)
print(article_links)
print(article_upvotes)
```

Excercise

- Print the title and link for the story

Ethics in Web Scrapping

Web scraping is now legal

Here's what that means for Data Scientists

Tom Waterman Follow

Jan 29 · 3 min read ★

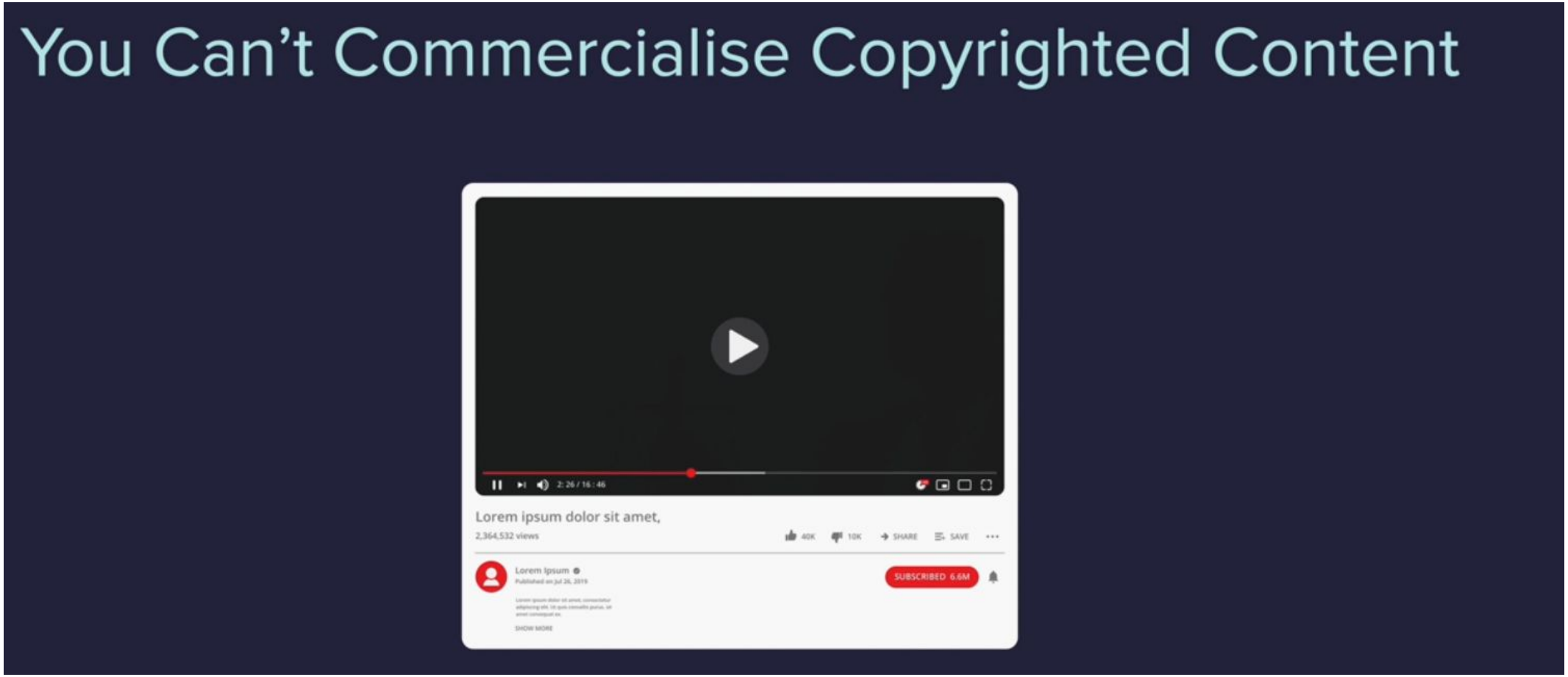
This story was sponsored by https://www.corbettanalytics.com.

In late 2019, the US Court of Appeals denied LinkedIn's request to prevent HiQ, an analytics company, from scraping its data.

The decision was a historic moment in the data privacy and data regulation era. It showed that any data that is *publicly available* and *not copyrighted* is fair game for web crawlers.

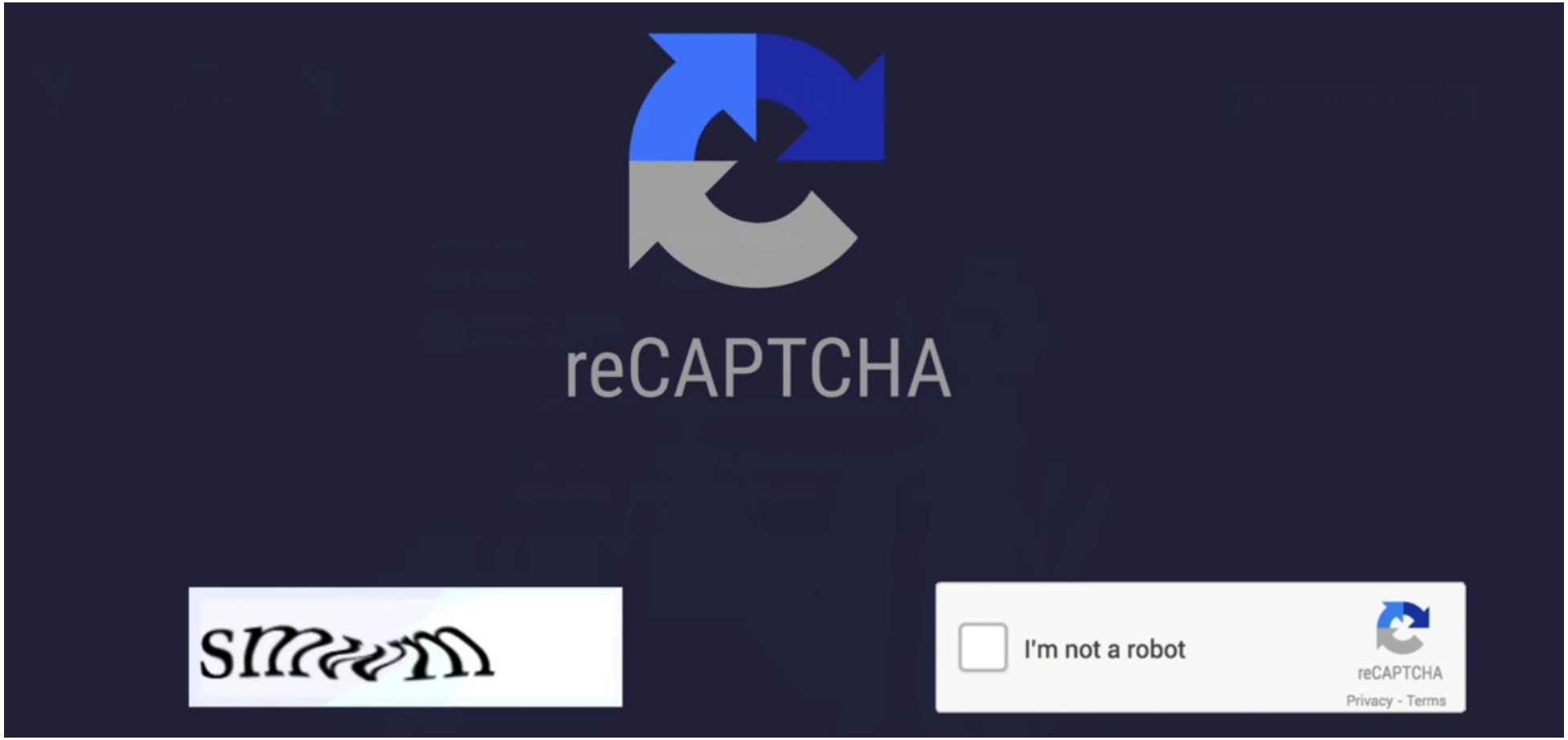
Top highlight

You can’t use copyrighted video as your own



After Login you can’t scrape

Prevent code/bot from scrapping



If there is an API, Go for it

In a minute scrape only for once

At the end place robots.txt, to see what they allow and what they don’t

- Linkedin.com/robots.txt

100 best movies of all the time

empireonline.com/movies/features/best-movies-2/

EMPIRE Movies Gaming PILOT Podcasts Awards Shopping Subscribe Win

1 of 100

100) Stand By Me

1986

Rob Reiner's adaptation of Stephen King's novella The Body is a stirring, touching adventure film which knows the real world is exciting and scary enough just as it is. It's also a coming-of-age movie which celebrates the intensity of childhood friendship, while gently mourning the transience of such bonds. Which is why, unlike its central character, it'll never get old.

Read Empire's review of Stand By Me
Buy the film now

Web Scrapping using Beautiful Soup, Final Task

- Scrape the Title and Year of each movie and store it in a file called movies.txt