

# Beautiful Soup



# AGENDA

## 40- 50 min

- I. What is web scraping?
- II. Tools to webscrape
- III. Beginner Exercise with BeautifulSoup4
- IV. Demonstration of Selenium + BeautifulSoup4
- V. Q&A

---

# Web Scrapping

Extracting information from a website (HTML) and parsing it in a readable format (this is called getting soup).

We will be using a package in the Python Library called BeautifulSoup which parses HTML and XML.

Some Best Practices:

- Check the Robots.txt and make sure User Agent: \* is not under disallow
- Scrapping too many pages at once quickly can get your IP blocked, use `time.sleep(x)` timers
- Determine the best tool

# Tools

- I. BeautifulSoup4 (web scrape)\*
- II. Requests\*
- III. Selenium (test websites)\*
- IV. WebDriver for Selenium
- V. Scrapy (advanced, Selenium + BeautifulSoup4)\*
- VI. Browser Developer Tool

\*use python package manager  
pip3, pip, conda to install these

---

Python 3: `pip3 install beautifulsoup4`

Python 2.7: `pip install beautifulsoup4`

`conda install -c anaconda beautifulsoup4`

# IMDB

## Monty Python & The Holy Grail

# Our Beginner Web Scrapping Exercise

1. Check the Robots.txt of IMDB
2. Parse the HTML 1 url of IMDB: Monty Python Holy Grail's IMDB if no issues seen in Robots.txt of IMDB
3. Let's web scrape the Title, Content Rating, Actors, Description of this movie
4. Let's create a function to add this information to the key of the ID of this IMDB Page eg: if we put tt0071853 into the function it will return the information in 4 as a dictionary.