2. Web scraping with crawl

(1) First, we start from the IMDB page of Harry Potter and the Sorcerer’s Stone.

(2) Secondly, we use the function parse () to read the data from the starting IMDB page and then navigate to the full Cast and Crew page. The function has two parts, and therefore, two command lines:

a) navigate to the next link, i.e., the full Cast and Crew page.

b) use “request” to go to the next link and we use the next function to retrieve more specific data about the actors.

(3) Thirdly, we write a function called parse\_full\_credits () to scrape all the actor names and go to their IMDB pages.

a) first the function yields a scrapy. Request for the page of each actor listed on the page.

b) then once we get into the actors’ individual pages, we use the next function to scrape what we need.

(4) Finally, we write a function parse\_actor\_page (), which scrapes the actor/actress’s Filmography section and log all the movies or TV shows that he/she has worked on.

a) first, we extract the actors’ names.

b) then, we extract the movies the actors have worked on before from their pages.

c) Lastly, we yield a dictionary to for each of all the movies or TV shows that he/she has worked on.