Scraping Job Posts

Domingos Lopes

Bootcamp #8
New York City Data Science Academy

Introduction

- Task:

  Scrape job posts from multiple sources, and use the data to build a job recommendation system.

First Approach

- Get a list of the most important companies.
- Scrape each one individually.

Main advantage:

- Cleaner data

Main disadvantages:

- A lot of spiders to create
- Potentially requires high maintenance

Second Approach

- Scrape job posts aggregators

  Main advantages:

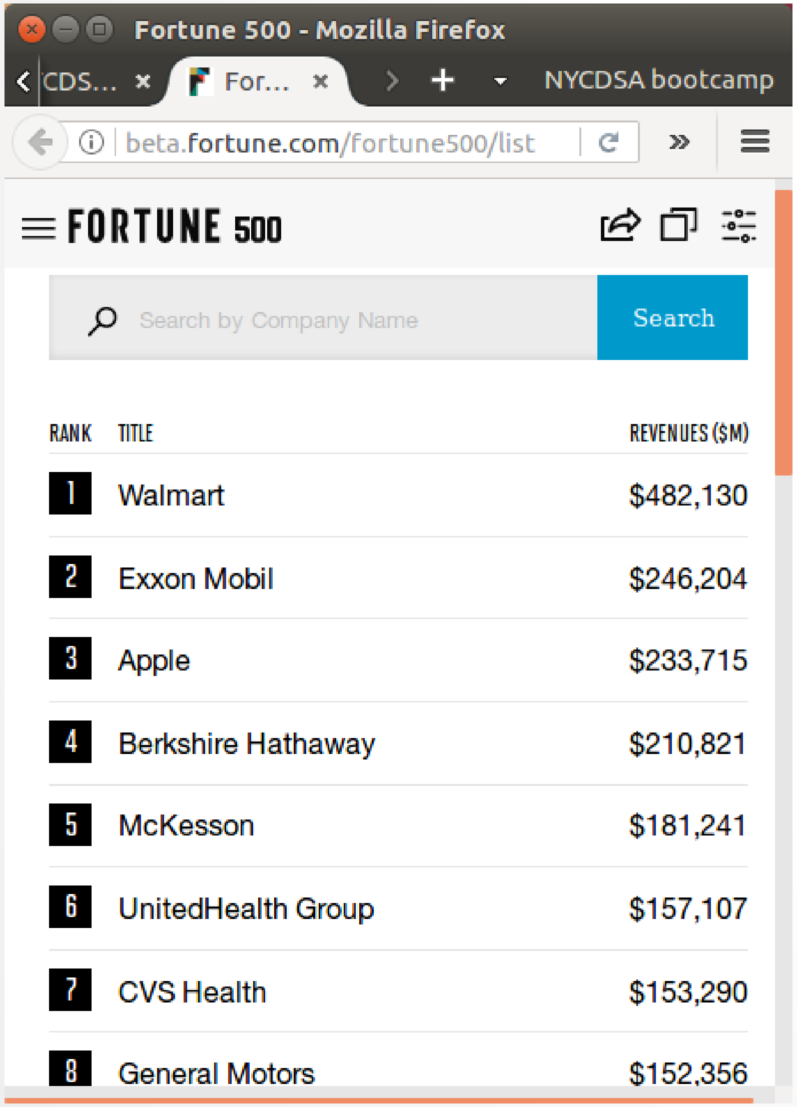
  – Can obtain great amounts of data using a single spider
  – Many more companies available
  – Often, an API will be available

  Main disadvantage:

  – Dirty data

First Approach Challenges

- Get list of top companies: http://beta.fortune.com/fortune500/list



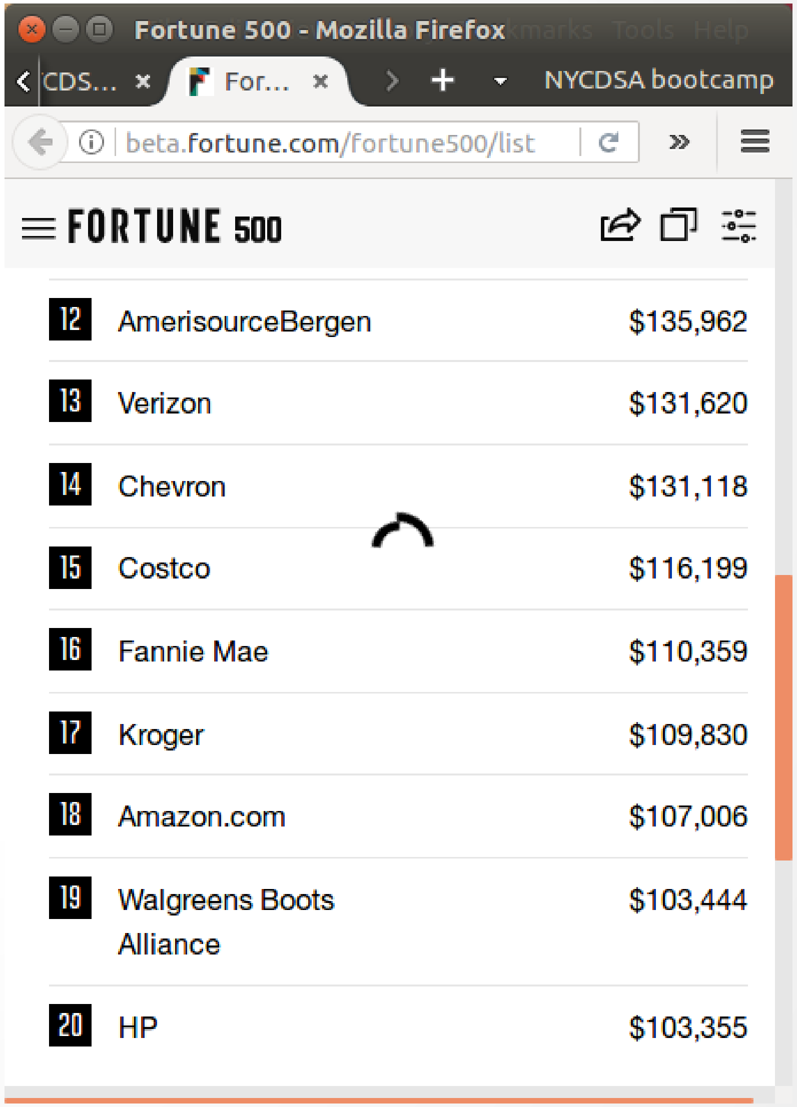
Fortune 500 - Mozilla Firefox

beta.fortune.com/fortune500/list

FORTUNE 500

Search by Company Name | Search

RANK	TITLE	REVENUES ($M)
1	Walmart	$482,130
2	Exxon Mobil	$246,204
3	Apple	$233,715
4	Berkshire Hathaway	$210,821
5	McKesson	$181,241
6	UnitedHealth Group	$157,107
7	CVS Health	$153,290
8	General Motors	$152,356



Fortune 500 - Mozilla Firefox

beta.fortune.com/fortune500/list

FORTUNE 500

12	AmerisourceBergen	$135,962
13	Verizon	$131,620
14	Chevron	$131,118
15	Costco	$116,199
16	Fannie Mae	$110,359
17	Kroger	$109,830
18	Amazon.com	$107,006
19	Walgreens Boots Alliance	$103,444
20	HP	$103,355

First Approach Challenges

- List starts with 20 rows.
- Dynamically expands as the user scrolls down.
- Not very scrapy-friendly.

  Solution:

  > Listen for AJAX (Asynchronous Javascript And XML) requests and identify the ones in which the important data is transmitted. Then replicate those within scrapy.

- Apple and Amazon had a similar obstacle.

First Approach Challenges

- Other challenges:
  - Add new job posts incrementally (not scraping old jobs again),
  - Take advantage of a common structure among posts
  - Location filter
- Results:
  - List of companies: 500 entries with all sorts of key indicators
  - Posts by company: Amazon (727), Apple (84), Facebook (44)

Second Approach Challenges

- Scraped Indeed and Dice
- Both provide an API to search positions, but one must scrape their page for each individual position.
- A lot of duplicate results within the same search.