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.

amazon
indeed
one search. all jobs.
Dice®

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:

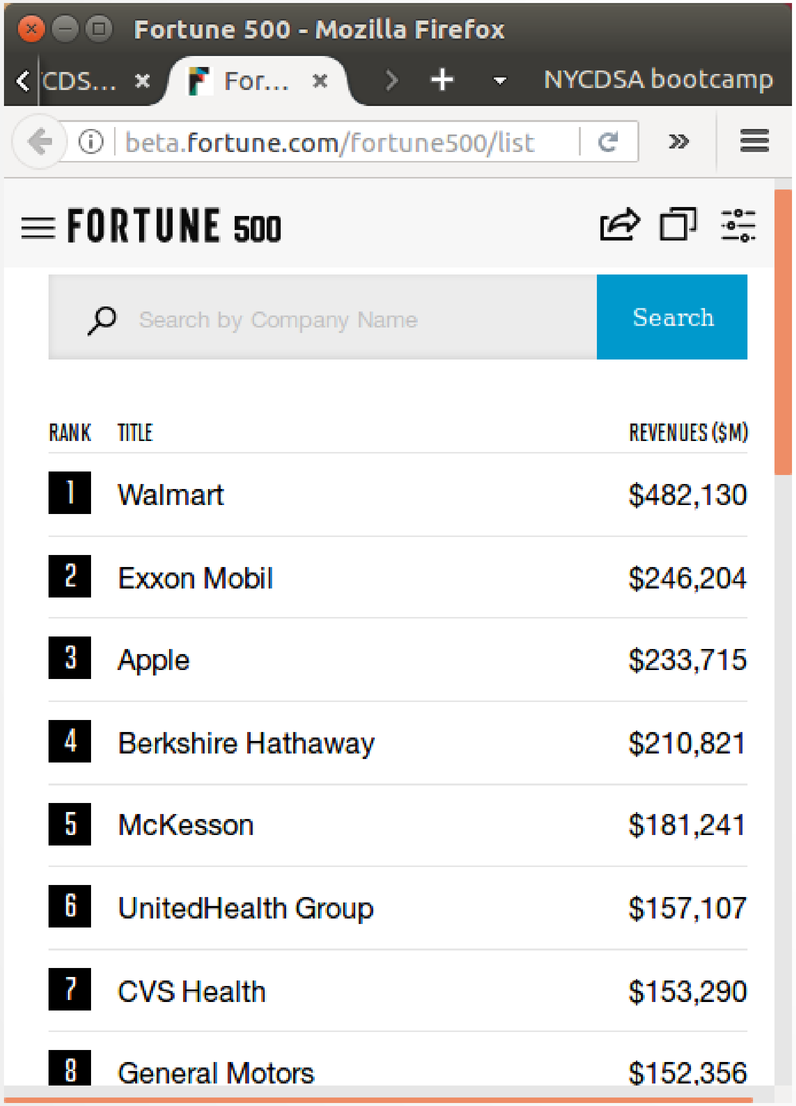
- Can obtain great amounts of data using a single spider
- Many more companies available
- Often, an API will be available

Main disadvantage:

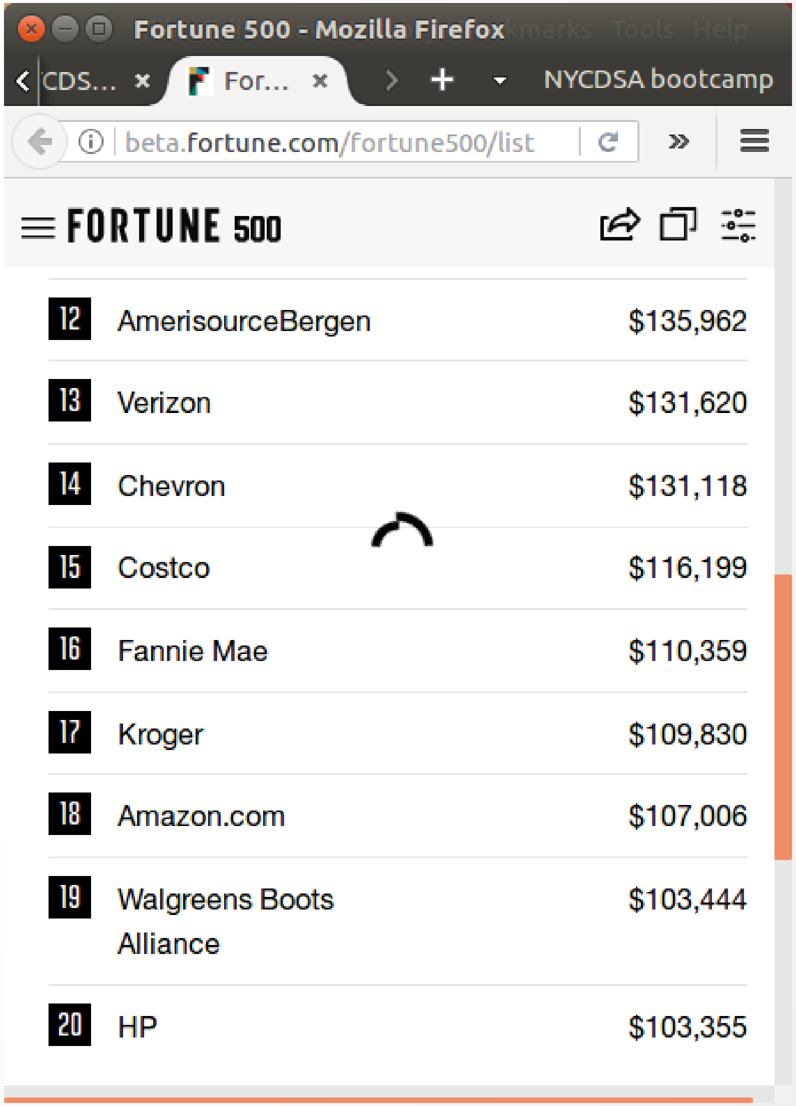
- Dirty data

First Approach Challenges

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



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
CDS…
For…
NYCDSA bootcamp
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.

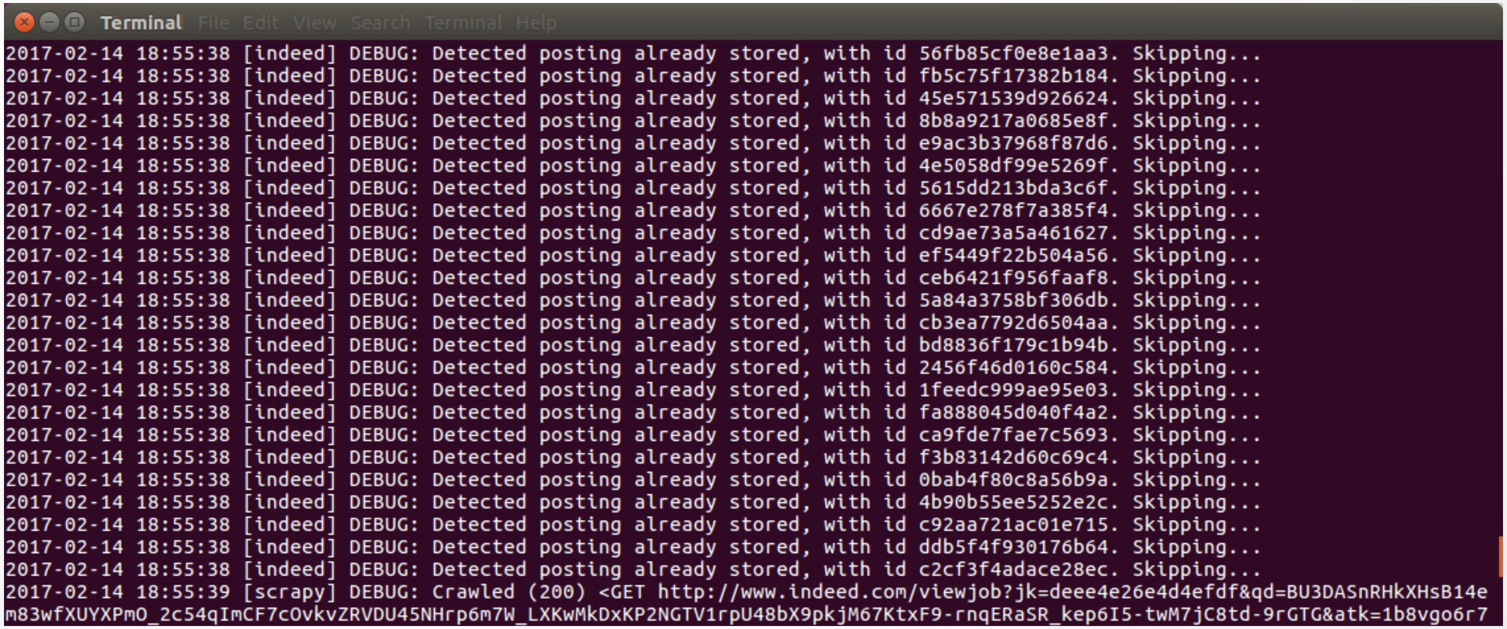
- Apple and Amazon had a similar obstacle.

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
  - Amazon uses the same url for html and json responses
- Results:
  - List of companies: 500 entries with all sorts of key indicators
  - Posts by company: Amazon (727/5862), Apple (84/877), Facebook (44/79)

Second Approach Challenges

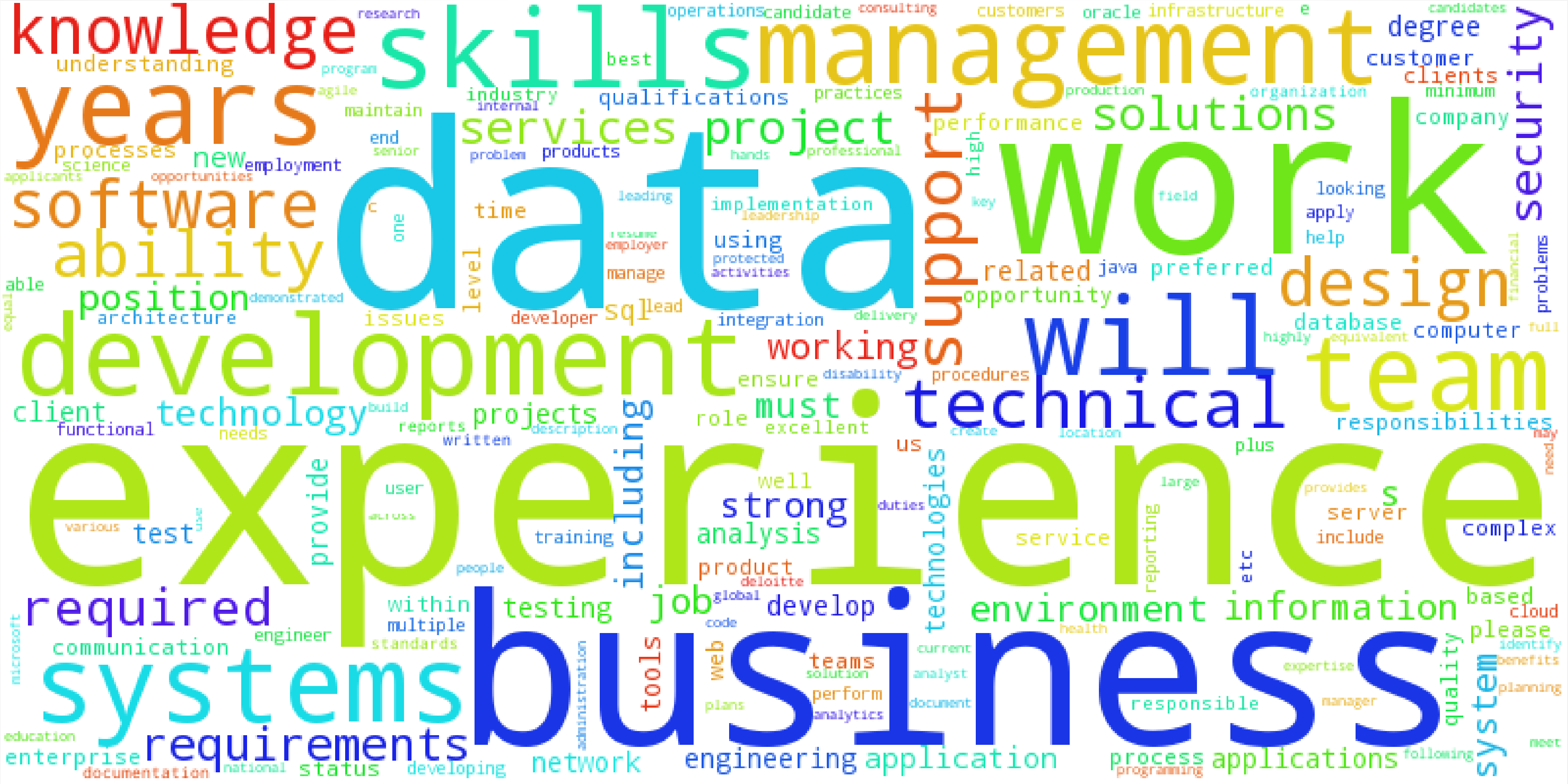
- Scraped Indeed and Dice
- Both provide an API to search positions, but one must scrape their page for each individual position.
- A lot of duplicate results within the same search.



```
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 56fb85cf0e8e1aa3. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id fb5c75f17382b184. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 45e571539d926624. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 8b8a9217a0685e8f. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id e9ac3b37968f87d6. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 4e5058df99e5269f. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 5615dd213bda3c6f. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 6667e278f7a385f4. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id cd9ae73a5a461627. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id ef5449f22b504a56. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id ceb6421f956faaf8. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 5a84a3758bf306db. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id cb3ea7792d6504aa. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id bd8836f179c1b94b. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 2456f46d0160c584. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 1feedc999ae95e03. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id fa888045d040f4a2. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id ca9fde7fae7c5693. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id f3b83142d60c69c4. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 0bab4f80c8a56b9a. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id 4b90b55ee5252e2c. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id c92aa721ac01e715. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id ddb5f4f930176b64. Skipping...
2017-02-14 18:55:38 [indeed] DEBUG: Detected posting already stored, with id c2cf3f4adace28ec. Skipping...
2017-02-14 18:55:39 [scrapy] DEBUG: Crawled (200) <GET http://www.indeed.com/viewjob?jk=deee4e26e4d4efdf&qd=BU3DASnRHkXHsB14e
m83wfXUYXPmO_2c54qImCF7cOvkvZRVDU45NHrp6m7W_LXKwMkDxKP2NGTV1rpU48bX9pkjM67KtxF9-rnqERaSR_kep6I5-twM7jC8td-9rGTG&atk=1b8vgo6r7
```

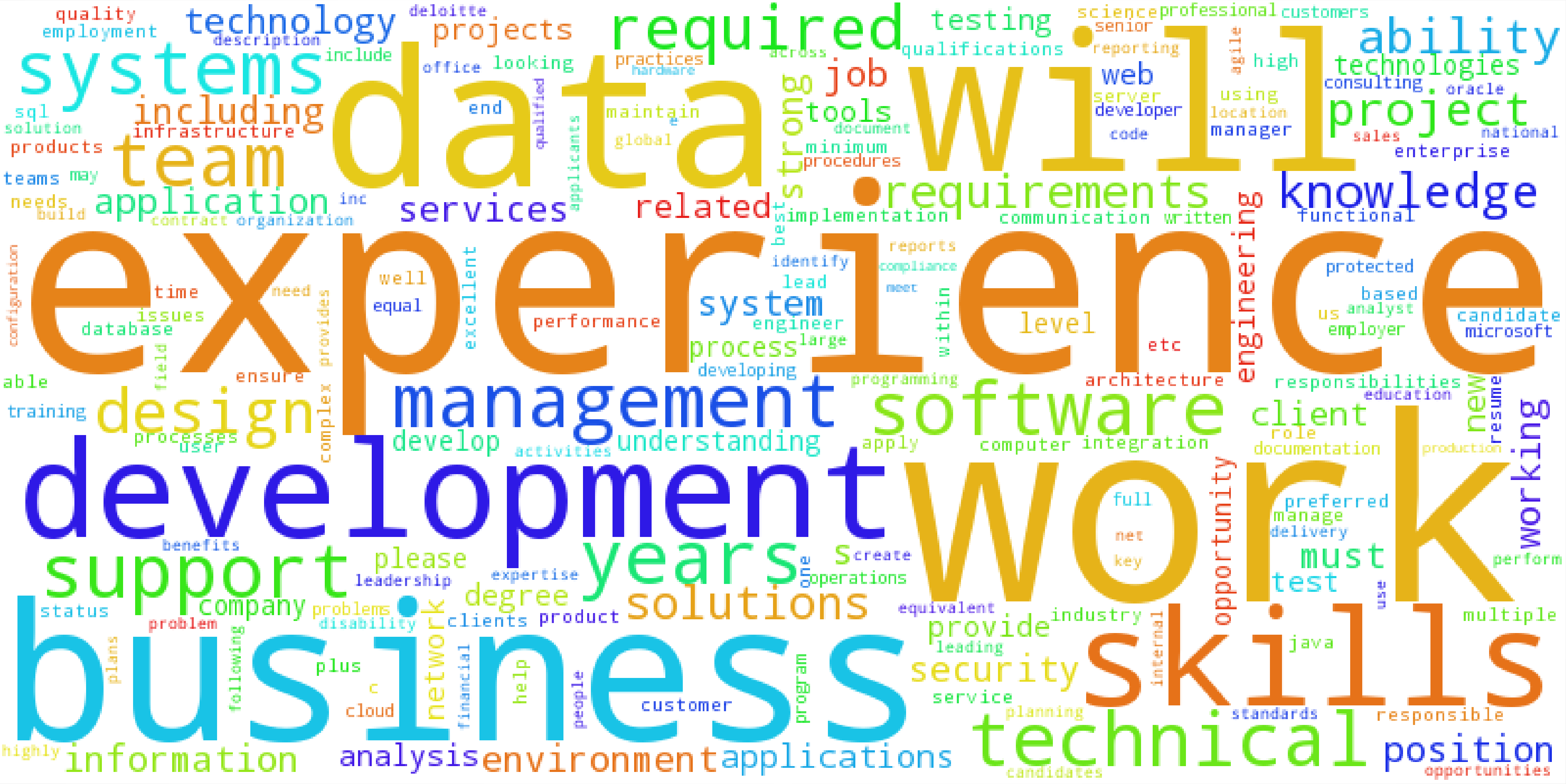
Indeed and Dice data

From “data scientist” search query



Indeed and Dice data

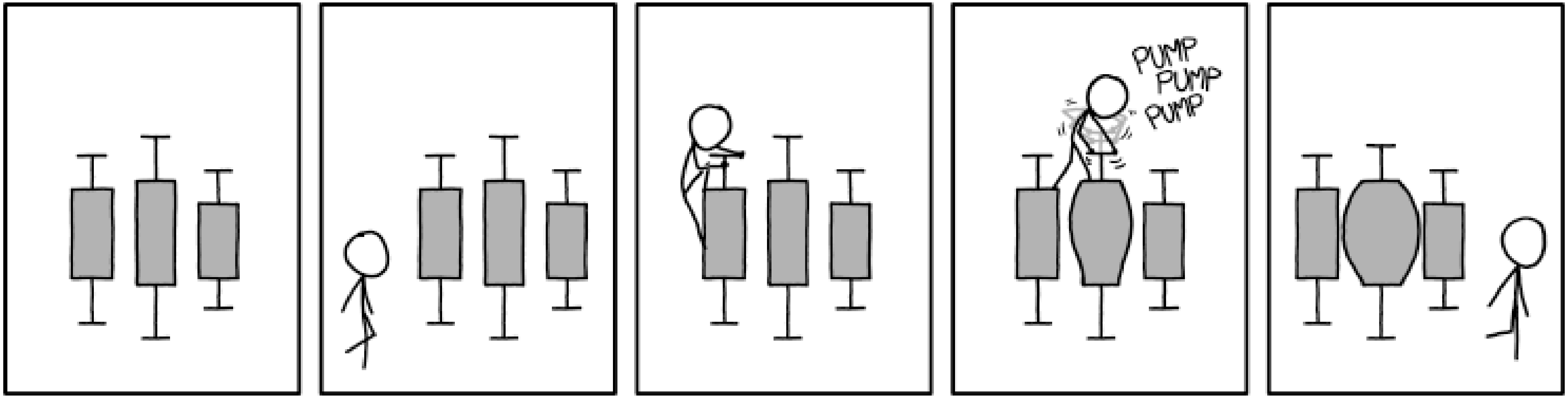
Also using “data engineer” and “business analyst”



Moving forward

- Use even more search queries
- Possibly scrape completely unrelated results – for classification purposes
- Be able to classify sentences as either part of a description or as part of requirements
- Proper classification of job postings (into several dimensions)

Thank you!



Credit: xkcd