By the end of this activity, you will be able to:

1. Display the nested structure of a JSON file.
2. Extract data from a JSON file.

**Step 1.** **Open a terminal shell.** Open a terminal shell by clicking on the square black box on the top left of the screen.



Change into the JSON directory:

1

cd Downloads/big-data-2/json

Run *ls* to see the JSON file and scripts:

1

2

ls

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/OgDspgW0Eea8MA4ynun-pw_dabf299ecd3370d9540da32c3d5e5d06_ls.png?expiry=1574640000000&hmac=roqIi0PNFOnf9eHtpbvRSuGKVoTp3gKJXYgcV1CN4e8

**Step 2. Look at JSON file.** Let's look at the contents of the JSON file:

1

more twitter.json



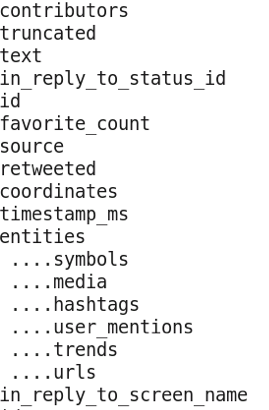
Press the spacebar to go down and *q*to quit more.

The contents of the file is difficult to understand since it is packed together.

**Step 3. View JSON schema.**We can view the schema of the JSON file by running *schema.py*:

1

./json\_schema.py twitter.json | more



The top-level fields are contributes, trucated, text, etc. Some fields have nested fields, such as entities, which contains symbols, media, hashtags, etc. If go you down (press spacebar), you will see multiple levels of nesting.

Enter *q* to quit more.

**Step 4. Extract values in JSON data.** We can extract individual values from fields within the JSON data by running print\_json.py:

1

./print\_json.py

The print\_json.py asks for the file name, tweet number, and path to extract. The path is the path to the field in the schema.

Let's look at the value for the *text* field in the 99th tweet. First, enter *twitter.json* for the filename:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/dXqD1AW9Eeahzgq54E2myw_d5f70055b92f89f52d085965a5602df6_filename-extract.png?expiry=1574640000000&hmac=ibwJXkiRBhBBAkY-LH8i66edKaq1XC-0cipeJbEqnNU

Next, enter *99* for the number:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/iYdEngW9Eea8MA4ynun-pw_5ff7a019a209e853f112ef85213c78ab_number-extract.png?expiry=1574640000000&hmac=WykY6tT5I3mKBrjrtT3hX1bV_jaGVRSBkxkFM8zVIeo

Next, enter *text* for the path:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/n5lLOAW9EeadmQogy2CpoQ_eb879b57424ec5d2fb156c333f474a7d_path-extract.png?expiry=1574640000000&hmac=6Fil5qzQLwEfx4SJVsiUEOuY0pjL1rpQ3G-lkwqTkPQ

The result is:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/2MQmrAW9EeadmQogy2CpoQ_0a7b1ecd00cef56aaf7863df8677aa0d_top-extract.png?expiry=1574640000000&hmac=9gYCvwVwXdwP1FnwV9Tr0FiS9uEhVrRX1VjuJ40n85M

Now let's find the value for *entities.hashtags* in the 99th tweet. The *hashtags* field is nested in the *entities* field, so we enter *entities/hashtags* for the path:

