DS 200 Lab2
Gather Real-Time Twitter Data Using Keywords and Hashtags
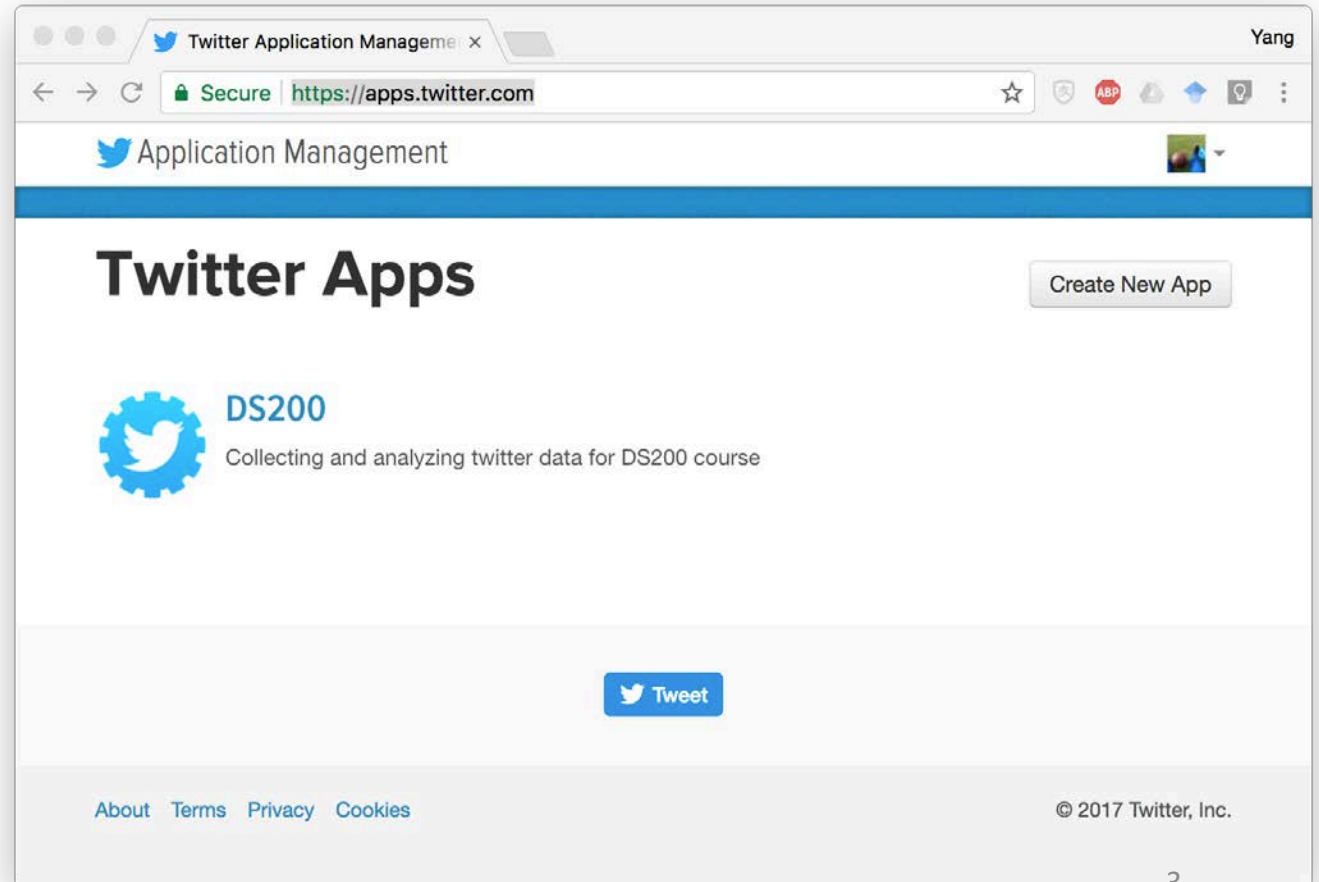
Aug 28, 2018

Instructor: John Yen

LA: Luewi Lei

Table of content

- Create a Twitter app (Slide 3 -- 9)

- Install tweepy package (Slide 10)

- Edit Python code for Gathering Tweets (Slide 11 -- 13)

- Gather tweets using the pipeline (Slide 14 -- 22)

- Load tweets in CSV file into Jupyter or Excel (Slide 23 -- 25)

Login In to Twitter Developer

- Go to https://apps.twitter.com/
- Login with your Twitter account

Apply for a Twitter developer account

Select a Twitter user profile to the developer account

1a. Use the Current Profile or
1b. Choose a Different Profile

2. Click "Continue"

Request access to a development account for personal use

- Choose a name for your development account (e.g., DS 200)

- Choose "United States" as the country of operation

- Click continue

developer.twitter.com/en/apply/usecase

Developer   Use cases   Products   Docs   More   Apply

User profile

✓ Account details

✓ Use case details

✓ Terms of service

✓ Email verification

What use case(s) are you interested in?

Select all that apply

☑ Academic research   ☑ Publish and curate Tweets

☐ Advertising   ☑ Student project / Learning to code

☑ Audience analysis
                                       ☑ Topic analysis
☑ Chatbots and automation
                                       ☑ Trend and event detection
☑ Consumer / end-user
experience   ☐ Other

☑ Engagement and customer
service

Describe in your own words what you are building

In English, please describe your product - the more detailed the response, the easier it is to review and approve. Be sure to answer the following:

- What is the purpose of your product or service?
- What will you deliver to your users/customers?
- How do you intend to analyze Tweets, Twitter users, or their content?
- How is Twitter data displayed to users of your end product or service (e.g. will Tweets and content be displayed at row level or in aggregate)?

The purpose of this twitter student project is to build a predictive model for classifying tweets (gathered based on keywords and hashtags) into two or more pre-defined classes (e.g., sentiment analysis regarding a person, an institution; support or against an opinion/policy), applying the model to tweets gathered later to assess accuracy of the model, trend or changes of the sentiment/opinion.

Will your product, service, or analysis make Twitter content or derived information available to a government entity?

In general, schools, colleges, or universities do *not* fall under this category.

◉ No

○ Yes

Read and agree to the Terms of Service

8

You will receive a confirmation email on your e-mail address associated with the account.

After you click the confirmation link on your email, you can create a Twitter App

Click to Create a Twitter App

Welcome!

Congratulations! You have successfully created a new Twitter developer account. With this account, you now have access to new APIs, app management, and tools to facilitate and support development.

Below are a few steps to help you create an app and to get up and running with the new premium APIs.

If you're planning to use our standard APIs instead of our premium APIs, simply follow the steps below to create an app, then refer to the "Getting started" guide in our documentation for next steps.

🐦 **Helpful tools**

Dive into the docs ⌄

View API usage ⌄

Have a question? ⌄

Looking for something else? ⌄

Get started

Create an app ⌃

To use an API, we require you create an app as part of our OAuth authorization scheme. Visit the Apps page of this developer portal to create one. Then, return to this page to complete the next step.

Set up a dev environment ⌃

To begin using the new Premium APIs, you need to set up one or more dev environments for the endpoint — and connect it to an app. Dev environments can be used to isolate usage, rules, rate limits, and more. If you are planning to use our standard APIs, you can skip this step.

10

Obtain An API key to create a Twitter App
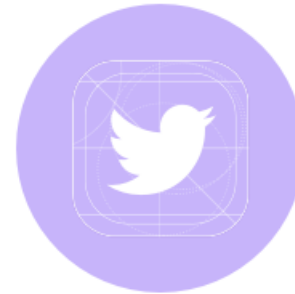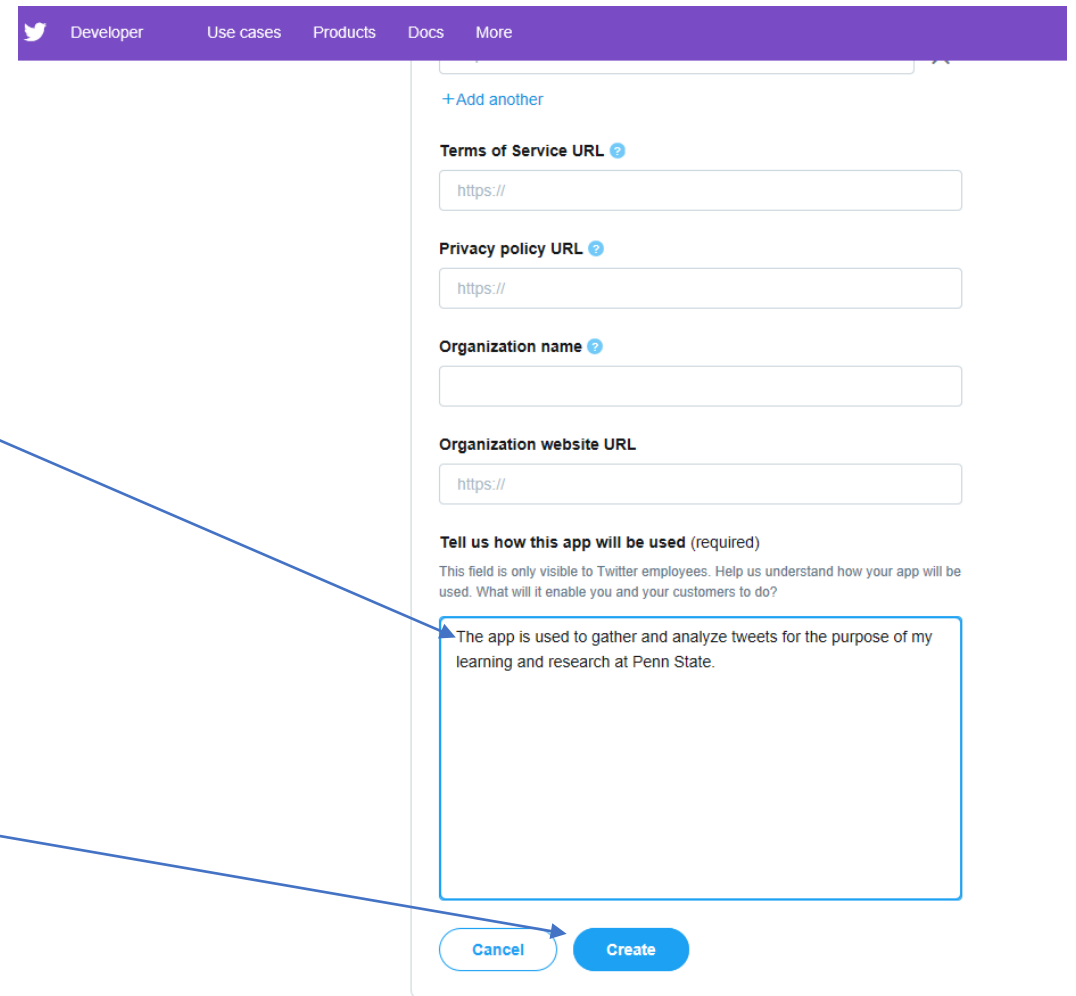
Create a Twitter app

- Fill in all the necessary information

- "App name", "Application description", "Website URL", and "Tell us how this App will be used" are required

- "Website" could be your Facebook homepage, your personal home page, etc.

- Scroll down and click on "Create"

12

Review Developer Terms

1. Read a summary of the Developer Terms

2. Click "Create"

Privacy policy URL

Review our Developer Terms ✕

As a reminder, you have agreed to our Developer Agreement and Policy. Please be mindful of the following restricted use cases

⊗ **Sensitive Information**

Be careful about using Twitter data to derive or infer potentially sensitive characteristics about Twitter users (ie. health, political or religious affiliation, ethnic origin, sexual orientation, and more).

⊗ **Government use and surveillance**

We prohibit the use of Twitter Data and Twitter APIs by any entity for surveillance purposes. Period.

⊗ **Automation**

If your app is will be used to post Tweets, follow accounts, or send Direct Messages, you should carefully review the Automation Rules to ensure you comply with our guidelines and never perform bulk, aggressive, or spammy actions.

Cancel   Create

Create Keys and Tokens for Tweets Gathering

You will see a web page for your twitter App, showing the App name, description, and URL you provided in the previous step.

Click
"Keys and tokens"

You will see your Consumer API keys

API key
API secret key

Copy these keys for later use.
Don't share with anyone.

Click "Create" to create Access token and Access token secret

Access token & Access token secret

Obtain Twitter Keys for Tweets Gathering -- Done

- Now you have four long keys for the following:
- ***API key***
- ***API secret key***
- ***Access token***
- ***Access token secret***

- These are what you need for using Twitter API (Application Program Interface) to gather tweets

Install *tweepy* package

- Open Command Line (Windows) or Terminal (Mac/Linux)

- Type: **pip install tweepy**

- Use pip3 in case you also have python2 in your system

- For Windows, go to next page
- For Mac, skip next page, go to slide 20.

Copy the Python code for Gathering Tweets (Windows)

- Download the ***gather_tweets_pipeline.py*** file from Canvas.

- Open **Command Line** in Windows

- Create a directory for the lab using "mkdir" command
  - E.g., "mkdir C:\Users\YourName\MyDocument\DS200lab3" for Windows
- Enter the directory where you download *get_tweets_pipeline.py*
  - E.g., "cd C:\Users\YourName\Downloads" for Windows

- Copy the file to your lab3 directory
  - E.g., "copy get_tweets_pipeline.py C:\Users\YourName\MyDocument\DS200lab3" for Windows

Copy the Python code for Gathering Tweets (Mac/Linux)

- Download the ***gather_tweets_pipeline.py*** file from Canvas.

- Open **Terminal** in Mac/Linux
- Create a directory for the lab using "mkdir" command
  - E.g., "mkdir ~/DS200lab3" for Mac/Unix   ("~" refers to your home directory)

- Enter the directory where you download *get_tweets_pipeline.py*
  - E.g., "cd ~/Downloads" for Mac/Linux

- Copy the file to your lab3 directory
  - E.g., "cp get_tweets_pipeline.py ~/DS200lab3" for Mac/Linux

Download one of the text editors

We need a text editor suitable for editing code.

Text editor possibilities:
- Windows: Notepad++, sublime text, Atom
- Mac/Linux: Atom, sbulime text, gedit, vim, etc.

Download Notepad++:
https://notepad-plus-plus.org/download/v7.5.8.html
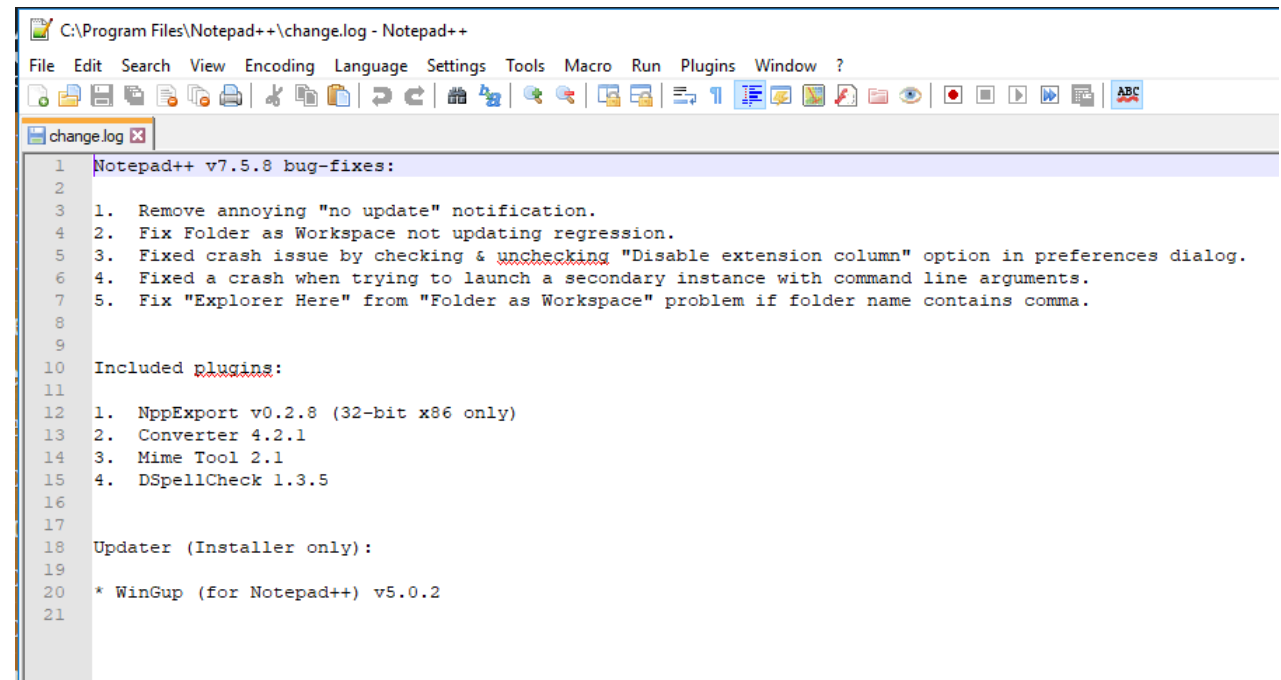
Download Atom:
http://atom.io
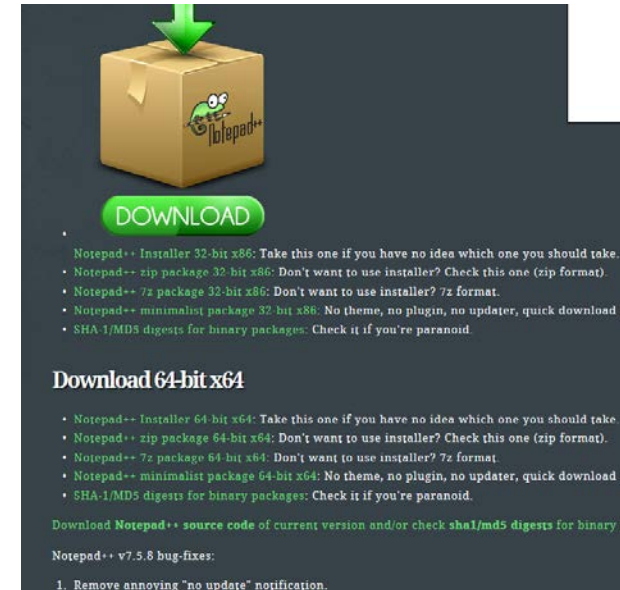
Edit the Python code using Text Editor

- Open *get_tweets_pipeline.py* file using one of the text editors (NOT WORD)

- For Notepad++, select "File", then "Open", find the get_tweets_pipeline.py file

Add API keys to gater_tweets_pipeline.py

- Scroll down to the middle of the file, you will find four assignments for the API keys we gathered earlier.

- Enter API key,
  API secret key,
  Access token,
  Access token secret
  between the two quotes

- They are strings assigned to the corresponding variables.

- Save the file
  get_tweets_pipeline.py
  (If you don't save, you will loose the changes you made above)

```python
 98                 print(status)
 99                 return True
100
101
102     # Get the str representation of the current date and time
103     def current_datetime_str():
104         return f'{datetime.datetime.now():%Y-%m-%d_%H-%M-%S}'
105
106
107     # Main
108     def main():
109         # Key and token info needed
110         consumer_key = ''
111         consumer_secret = ''
112         access_token = ''
113         access_secret = ''
114
115         auth = OAuthHandler(consumer_key, consumer_secret)
116         auth.set_access_token(access_token, access_secret)
117         api = tweepy.API(auth)
118
119         # Welcome
120         print('===========================================================')
121         print('Welcome to the user interface of gathering tweets pipeline!')
122         print('You can press "Ctrl+C" at anytime to abort the program.')
123         print('===========================================================')
124         print()
```

How to choose keywords and hashtags?

1. Log in to Twitter.com, use the search box on the upper right.
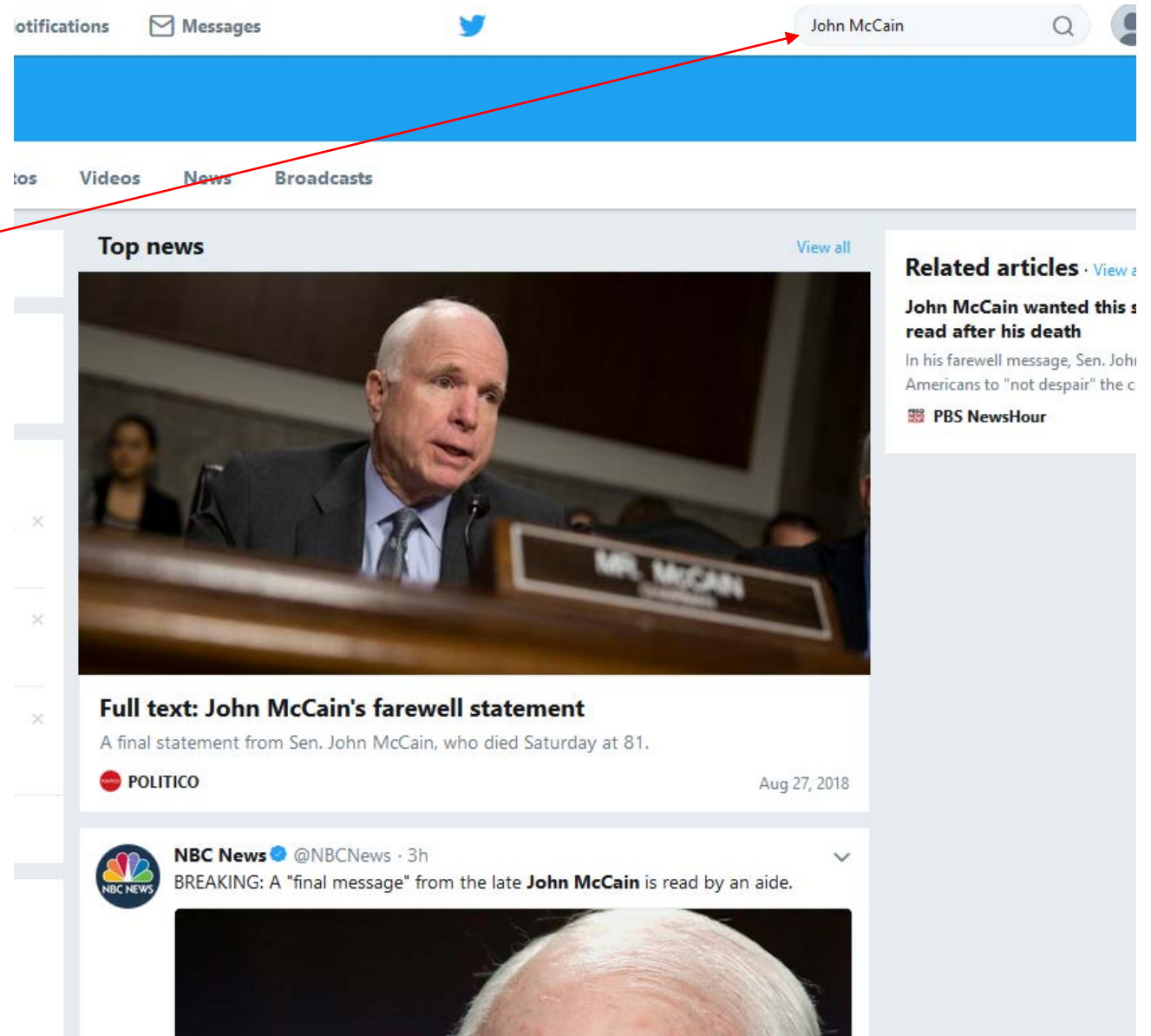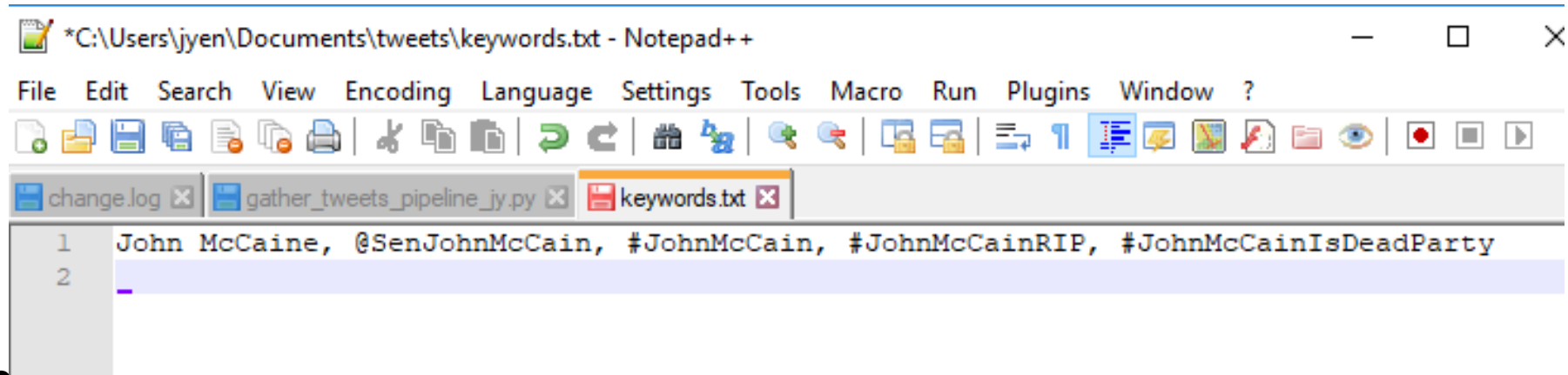
Scroll down to find hashtags, twitterIDs that occur frequently

- Example: @SenJohnMcCain showed up in a few tweets

Create a new file of keywords using a text editor

- Create this file using a Text Editor (File -> New)
- How a key word file (using Notepad++) looks like

- Hashtags need to be preceeded with #
- TweeterIDs need to be preceeded with @
- Multiple key words and hashtags are allowed
- Multiple lines are allowed
- Save this keyword file as "keywords.txt" in the same directory as get_tweets_pipeline.py

Start the pipeline for gathering tweets

- Open **Command Line** in Windows, or **Terminal** in Mac/Linux

- Enter the directory (e.g., DS200lab) to which you copied the downloaded files *get_tweets_pipeline.py*
  - E.g., "cd C:\Users\MyDocument\DS200lab" for Windows
  - E.g., "cd ~/DS200lab" for Mac/Linux

- Invoke get_tweets_pipeline.py to start the "pipeline" for gathering tweets

python gather_tweets_pipeline.py (For Windows)
python3 gather_tweets_pipeline.py (For MacOS/Linux)

Step 1

- Choose the way to specify your key words

```
C:\Users\jyen\Documents\tweets>python gather_tweets_pipeline.py
============================================================
Welcome to the user interface of gathering tweets pipeline!
You can press "Ctrl+C" at anytime to abort the program.
============================================================


How do you want to specify your key words?
Type "manual" or "file" >>> file
```

- Recommend to enter keywords through a file
- Benefit of using a file: Reuse the keywords and hashtags each time you gather tweets facilitate comparing tweets gathered in different time.

Step 1.1

- Type the file name to the prompt "Type your file name >>>"


```
Command Prompt - python gather_tweets_pipeline.py                                    —    □    ✕

C:\Users\jyen\Documents\tweets>copy gather_tweets_pipeline_jy.py gather_tweets_pipeline.py
        1 file(s) copied.

C:\Users\jyen\Documents\tweets>python gather_tweets_pipeline.py
=================================================================
Welcome to the user interface of gathering tweets pipeline!
You can press "Ctrl+C" at anytime to abort the program.
=================================================================


How do you want to specify your key words?
Type "manual" or "file" >>> keyword
"keyword" is an invalid input! Please try again.

Type "manual" or "file" >>> file
=================================================================
Please input the file name that contains your key words.
Notes:
    The file should contain key words in one or multiple lines, and multiple key words should be separated by *COMMA*.

        For example: NBA, basketball, Lebron James
    If the file is under the current directory, you can directly type the file name, e.g., "keywords.txt".
    If the file is in another directory, please type the full file name, e.g., "C:\Downloads\keywords.txt" (for Window
s), or "/Users/xy/Downloads/keywords.txt" (for MacOS/Linux).
Type your file name >>> keywords.txt
```

Step 1.2

- Alternatively, you can input key words manually

```
~/GitHub/tweepy-exp(master) » python3 gather_tweets_pipeline.py
================================================================
Welcome to the user interface of gathering tweets pipeline!
You can press "Ctrl+C" at anytime to abort the program.
================================================================


How do you want to specify your key words?
Type "manual" or "file" >>> manual
================================================================
Please input your key words (separated by comma), and hit <ENTER> when done.
Type the key words >>> █
```

- Type in your key words and/or hashtags, separated by comma

Step 1.2

- Manual input of key words and hashtags


```
~/GitHub/tweepy-exp(master) » python3 gather_tweets_pipeline.py
================================================================
Welcome to the user interface of gathering tweets pipeline!
You can press "Ctrl+C" at anytime to abort the program.
================================================================


How do you want to specify your key words?
Type "manual" or "file" >>> manual
================================================================
Please input your key words (separated by comma), and hit <ENTER> when done.
Type the key words >>> Lebron James, NBA, basketball

3 unique key words being used:  ['basketball', 'NBA', 'Lebron James']
================================================================
How many tweets do you want to gather?
Input an integer number, or just hit <ENTER> to use the default number 300.
Input an integer >>> 
```
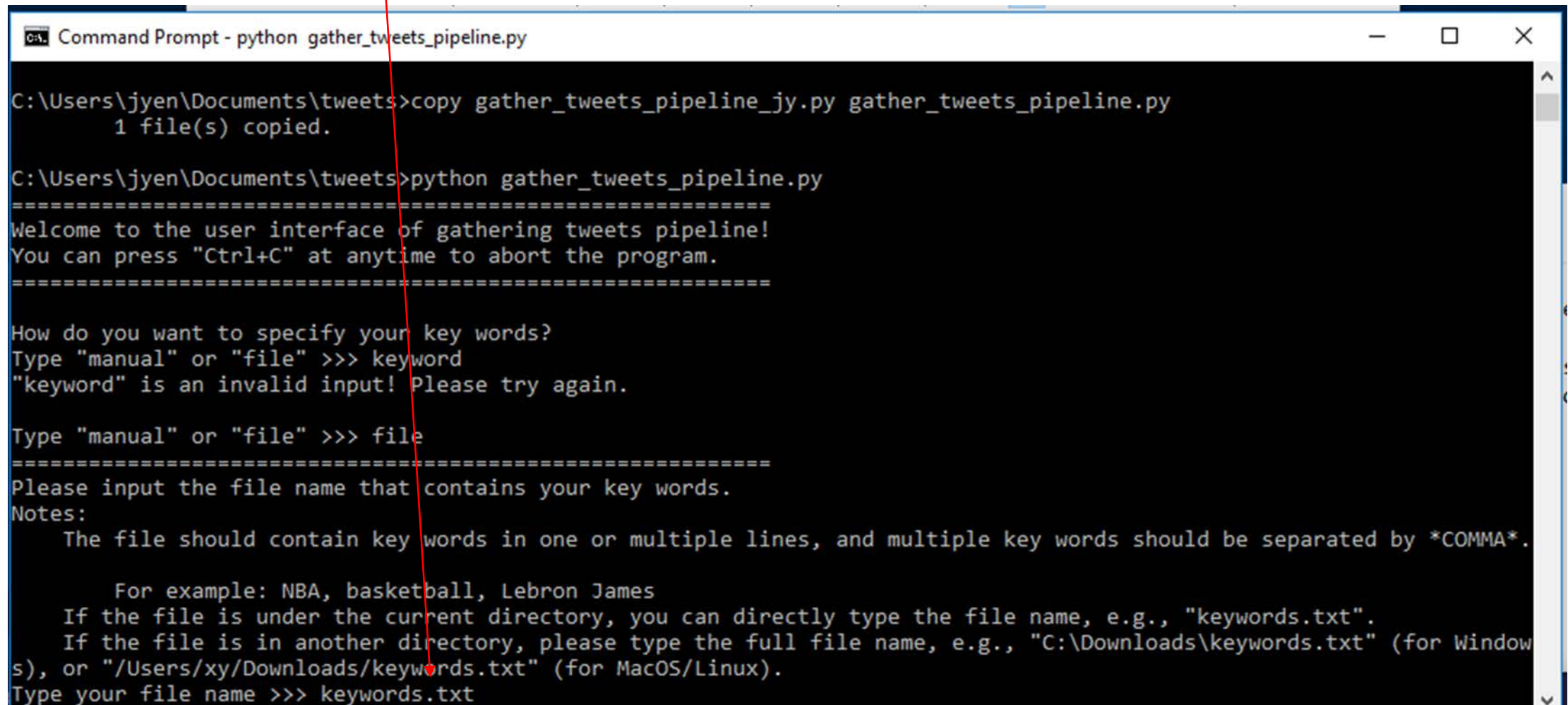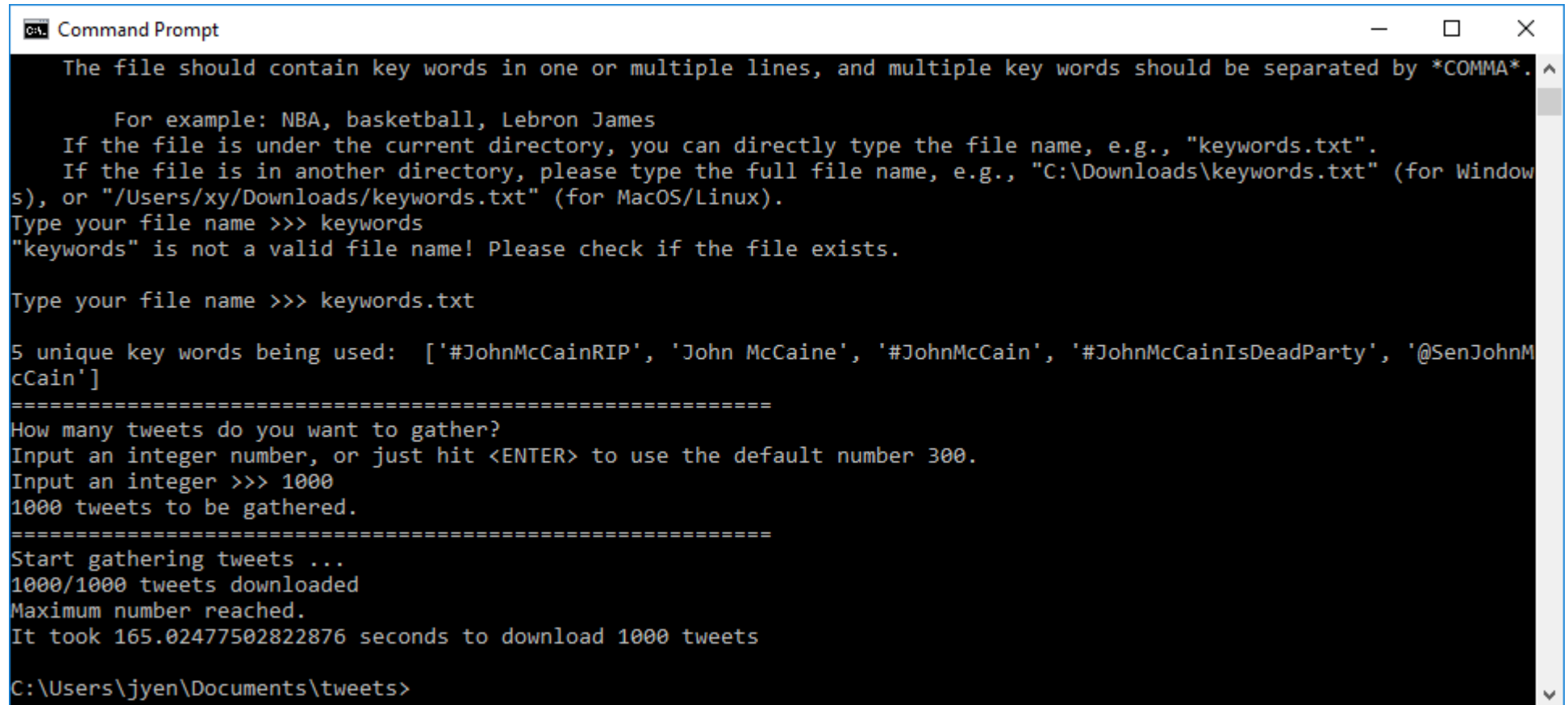
Step 2

- Decide the number of tweets you want to gather this time (recommend: 1000)


```
The file should contain key words in one or multiple lines, and multiple key words should be separated by *COMMA*.

        For example: NBA, basketball, Lebron James
    If the file is under the current directory, you can directly type the file name, e.g., "keywords.txt".
    If the file is in another directory, please type the full file name, e.g., "C:\Downloads\keywords.txt" (for Window
s), or "/Users/xy/Downloads/keywords.txt" (for MacOS/Linux).
Type your file name >>> keywords
"keywords" is not a valid file name! Please check if the file exists.

Type your file name >>> keywords.txt

5 unique key words being used:  ['#JohnMcCainRIP', 'John McCaine', '#JohnMcCain', '#JohnMcCainIsDeadParty', '@SenJohnM
cCain']
================================================================
How many tweets do you want to gather?
Input an integer number, or just hit <ENTER> to use the default number 300.
Input an integer >>> 1000
1000 tweets to be gathered.
================================================================
Start gathering tweets ...
1000/1000 tweets downloaded
Maximum number reached.
It took 165.02477502822876 seconds to download 1000 tweets

C:\Users\jyen\Documents\tweets>
```

- You can repeat step 1 and 2 to gather additional tweets.

Step 3

- Wait until all tweets are gathered
- If you do not see the command prompt for a while (which means the number of tweets gathered has not reached the maximum you specified), you can choose to press "Ctrl+C" to interrupt. You will still be able to see the tweets gathered so far.

- Three files will be generated in the same directory for each run
  - raw_YYYY-MM-DD-HH-MM-SS.json   (raw tweets from Twitter, JSON format)
  - data_YYYY-MM-DD-HH-MM-SS.csv   (selected fields from raw tweets, CSV format)
  - text_YYYY-MM-DD-HH-MM-SS.csv   (filtered tweets, where special characters in the body of tweets are removed)
  - YYYY-MM-DD is the date you gathered tweets
  - HH-MM-SS is the time your tweets files were generated

Load the CSV twitter file into Jupyter Notebook

Load the CSV file of the filtered tweets in Jupyter Notebook

Jupyter **Untitled** Last Checkpoint: 9 minutes ago (autosaved)

File   Edit   View   Insert   Cell   Kernel   Widgets   Help

▶ Run   ■   C   ▶▶   Code

In [4]: `d2 = Table.read_table('text_2018-08-27_17-18-26.csv')`
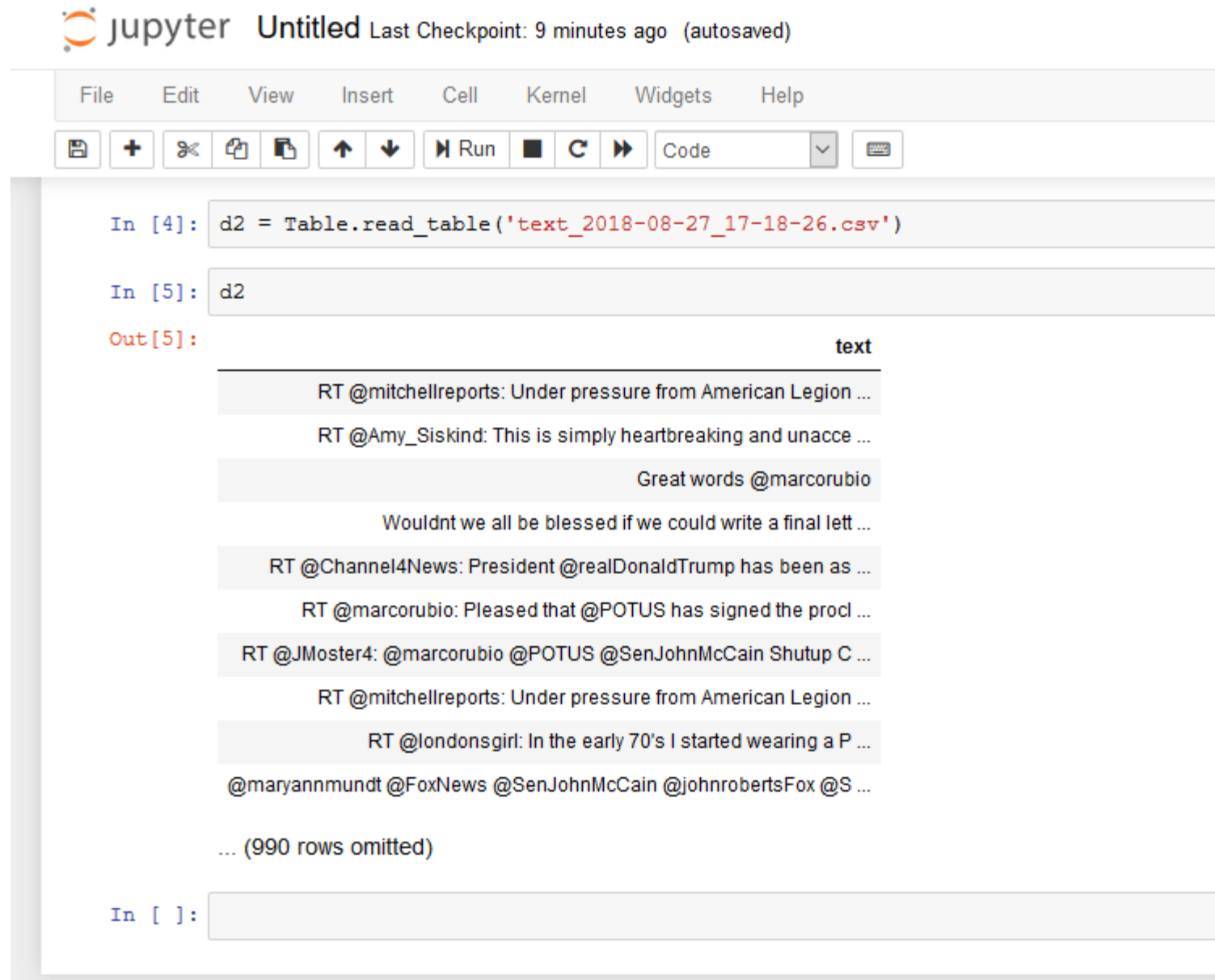
In [5]: `d2`

Out[5]:

text
RT @mitchellreports: Under pressure from American Legion ...
RT @Amy_Siskind: This is simply heartbreaking and unacce ...
Great words @marcorubio
Wouldnt we all be blessed if we could write a final lett ...
RT @Channel4News: President @realDonaldTrump has been as ...
RT @marcorubio: Pleased that @POTUS has signed the procl ...
RT @JMoster4: @marcorubio @POTUS @SenJohnMcCain Shutup C ...
RT @mitchellreports: Under pressure from American Legion ...
RT @londonsgirl: In the early 70's I started wearing a P ...
@maryannmundt @FoxNews @SenJohnMcCain @johnrobertsFox @S ...

... (990 rows omitted)

In [ ]:

Lab 2 Submission

- Submit two CSV files containing tweets you gathered (named "data_YYYY-MM-DD-HH-MM-SS.csv" and "text_YYYY-MM-DD-HH-MM-SS.csv")
- Submit a screenshot of displaying the first ten tweets of these two files in Jupyter notebook.