DS 200
Topic 2: Data Representation for Data Sciences

John Yen

Fall 2018

WHAT WE WILL LEARN TODAY?

- One dataset can be represented in multiple ways
- How to choose data representation for a DS project? PURPOSE.
- A simple representation of text: Bag of words
- Use Python and Jupyter Notebook for counting words in a Novel.

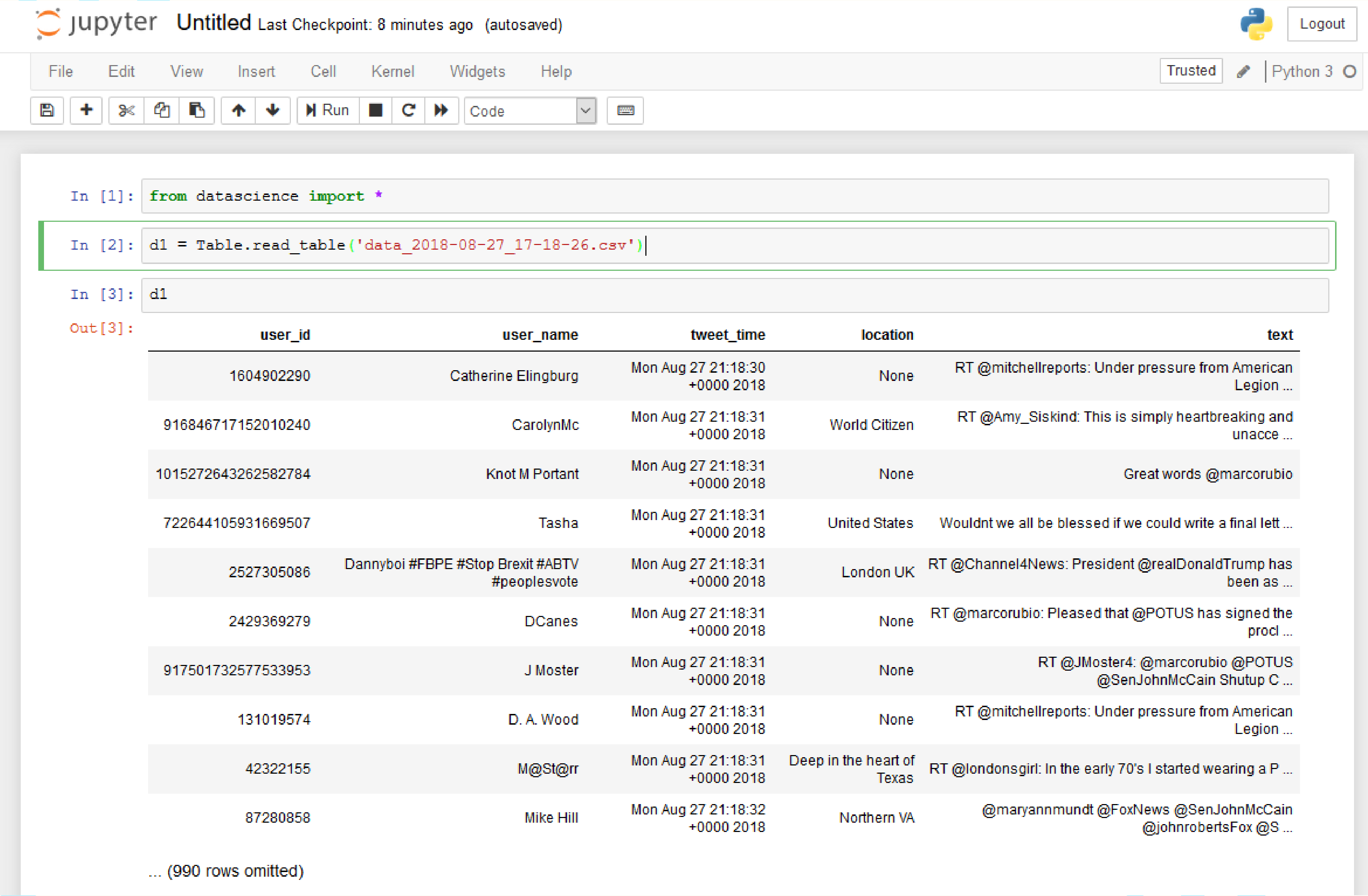
3 One Dataset Can be Represented in Multiple Ways

Question: What Types of data are included in a Twitter dataset?

A) Structured Data

B) Ill-structured Data

C) Both

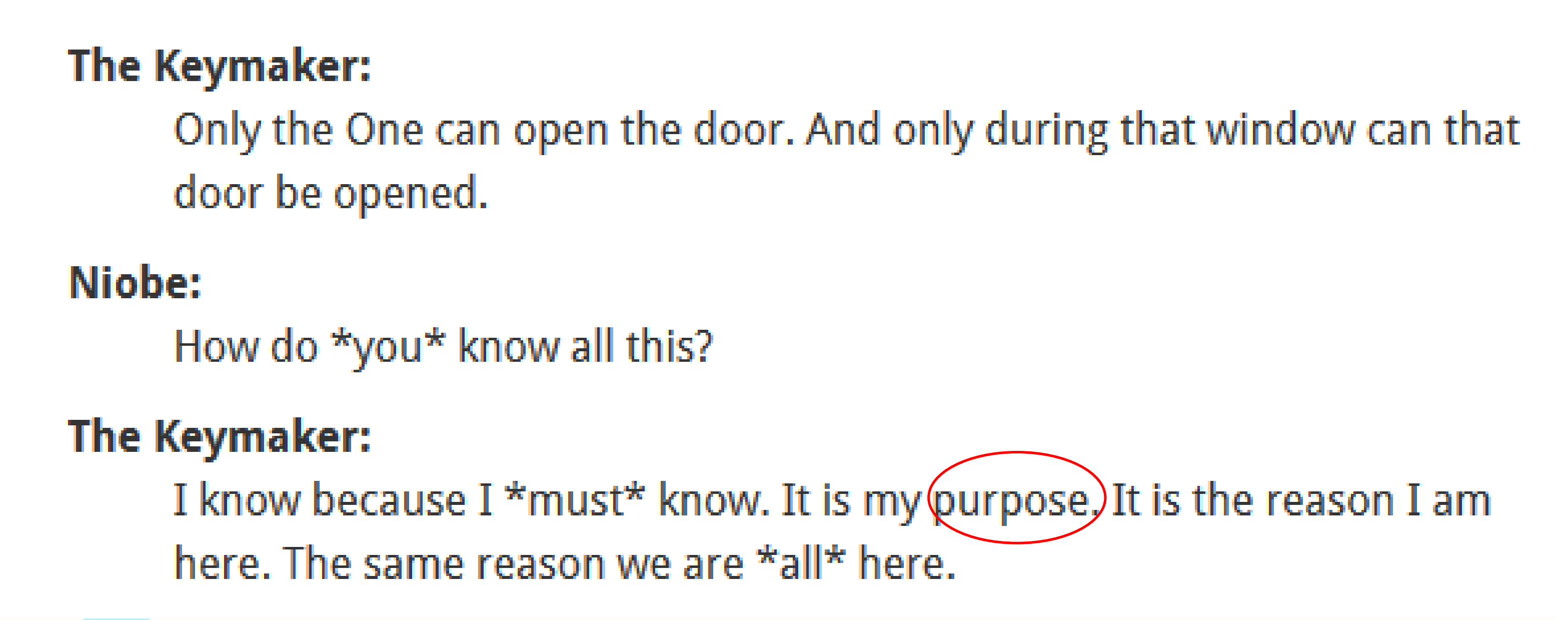


```
In [1]: from datascience import *

In [2]: d1 = Table.read_table('data_2018-08-27_17-18-26.csv')

In [3]: d1

Out[3]:
```

user_id	user_name	tweet_time	location	text
1604902290	Catherine Elingburg	Mon Aug 27 21:18:30 +0000 2018	None	RT @mitchellreports: Under pressure from American Legion ...
916846717152010240	CarolynMc	Mon Aug 27 21:18:31 +0000 2018	World Citizen	RT @Amy_Siskind: This is simply heartbreaking and unacce ...
1015272643262582784	Knot M Portant	Mon Aug 27 21:18:31 +0000 2018	None	Great words @marcorubio
722644105931669507	Tasha	Mon Aug 27 21:18:31 +0000 2018	United States	Wouldnt we all be blessed if we could write a final lett ...
2527305086	Dannyboi #FBPE #Stop Brexit #ABTV #peoplesvote	Mon Aug 27 21:18:31 +0000 2018	London UK	RT @Channel4News: President @realDonaldTrump has been as ...
2429369279	DCanes	Mon Aug 27 21:18:31 +0000 2018	None	RT @marcorubio: Pleased that @POTUS has signed the procl ...
917501732577533953	J Moster	Mon Aug 27 21:18:31 +0000 2018	None	RT @JMoster4: @marcorubio @POTUS @SenJohnMcCain Shutup C ...
131019574	D. A. Wood	Mon Aug 27 21:18:31 +0000 2018	None	RT @mitchellreports: Under pressure from American Legion ...
42322155	M@St@rr	Mon Aug 27 21:18:31 +0000 2018	Deep in the heart of Texas	RT @londonsgirl: In the early 70's I started wearing a P ...
87280858	Mike Hill	Mon Aug 27 21:18:32 +0000 2018	Northern VA	@maryannmundt @FoxNews @SenJohnMcCain @johnrobertsFox @S ...

... (990 rows omitted)

5 How to Choose Data Representation for a DS Project?



The Keymaker:

Only the One can open the door. And only during that window can that door be opened.

Niobe:

How do *you* know all this?

The Keymaker:

I know because I *must* know. It is my purpose. It is the reason I am here. The same reason we are *all* here.

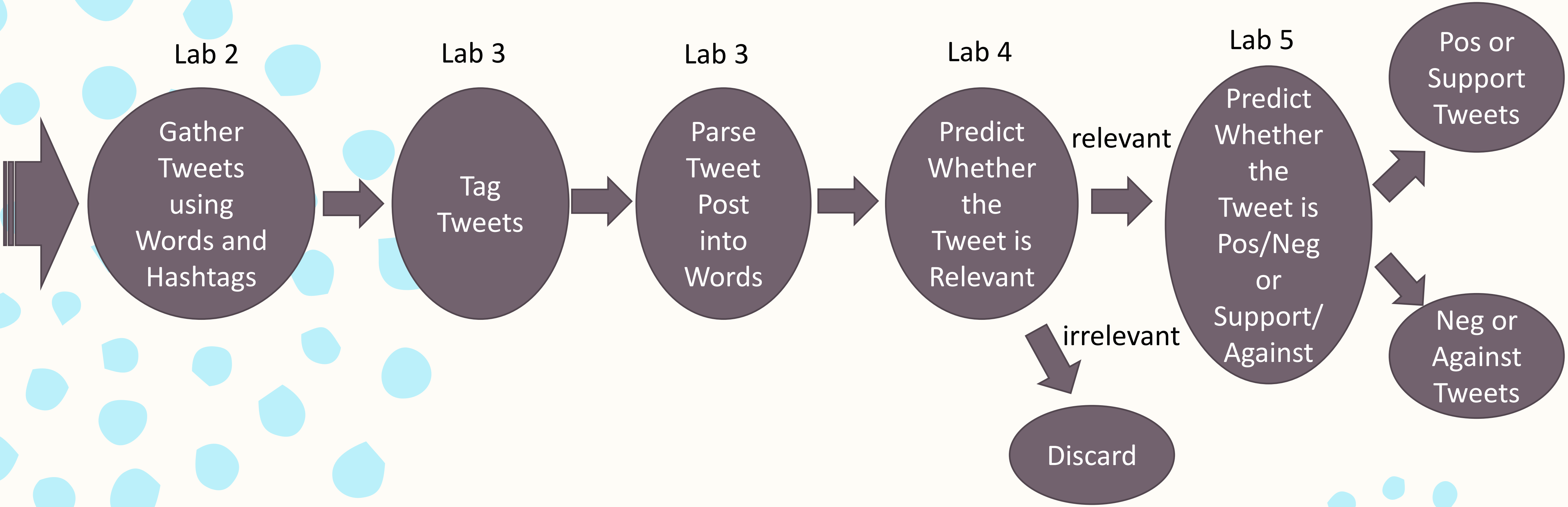
6 You are a member of a DS team to analyze the spreading of a viral tweet

Q: What information do you need to conduct this data analysis?

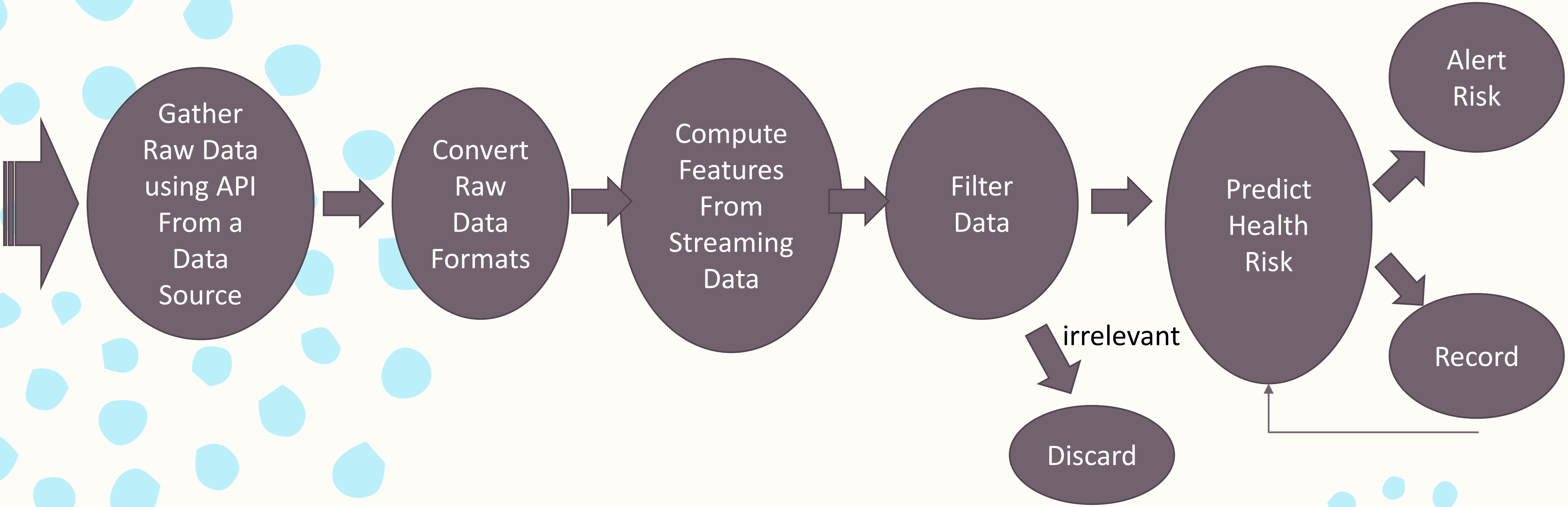
7 How to represent paper for the purpose of retrieving them by words?

8 How to represent web pages for the purpose of search engine?

9 Labs for Your Streaming Twitter Data Analytics Pipeline


Lab 2
Gather Tweets using Words and Hashtags
Lab 3
Tag Tweets
Lab 3
Parse Tweet Post into Words
Lab 4
Predict Whether the Tweet is Relevant
relevant
irrelevant
Discard
Lab 5
Predict Whether the Tweet is Pos/Neg or Support/ Against
Pos or Support Tweets
Neg or Against Tweets

10 Another Exemplar Streaming Data Analytics Pipeline



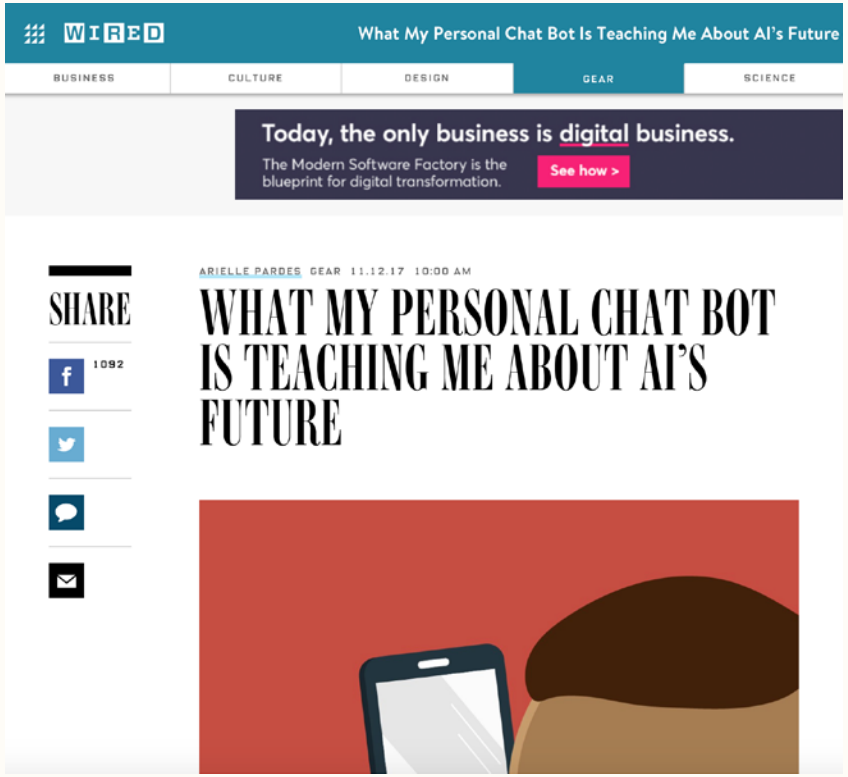
11 Parse Strings into Words

- An important step in all data science projects involving text data, including
- Parsing new Web pages for the preparation of Google search engine
- Parsing journal papers to identify its authors, references, etc. (Google Scholar, CiteSeer)
- Parsing tweets for the preparation of sentiment analysis, opinion mining, viral analysis, social influence analysis
- Parsing questions and conversations with customers for “chatbot”

12 A Chatbot as a “Friend”

- Replika

- Siri (Apple iPhone/iPad)
- Alexa (Amazon Echo)



What My Personal Chat Bot Is Teaching Me About AI’s Future

ARIELLE PARDES GEAR 11.12.17 10:00 AM

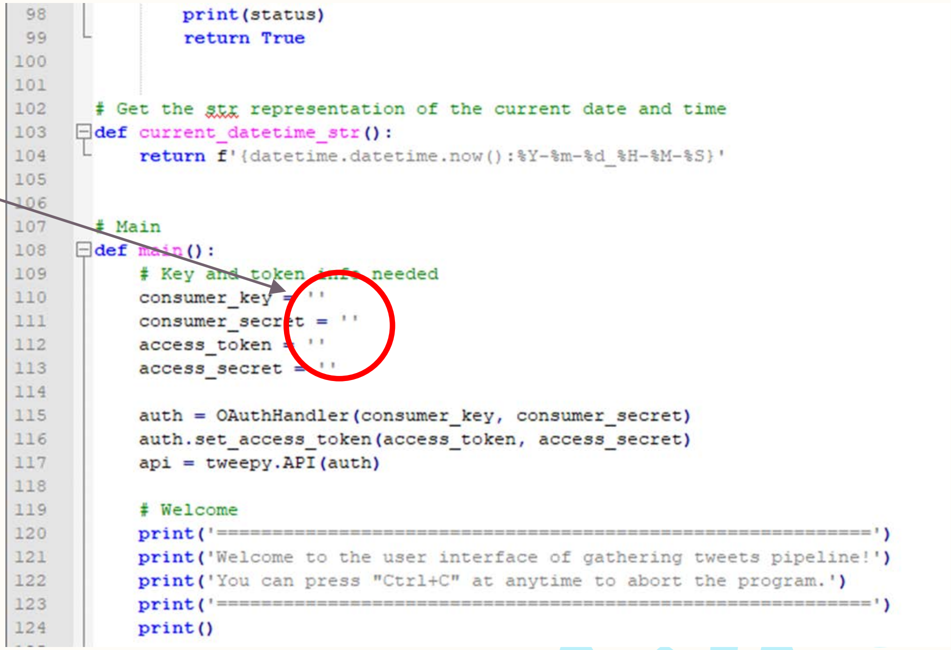
WHAT MY PERSONAL CHAT BOT IS TEACHING ME ABOUT AI’S FUTURE

13 Variable and Assignment in Python

- Variable: Refers to a ‘place’ in the memory for program to store and update data
- Data has different types: integer, floating point, string, list, etc.
- Assignment statement assigns a data to a variable
- Ex:
- x = 93 (integer)
- x = 93.5 (floating point)
- x = ‘I have a dream.’ (string)
- x = [‘I’, ‘have’, ‘a’, ‘dream.’]. (list of strings)

14 String

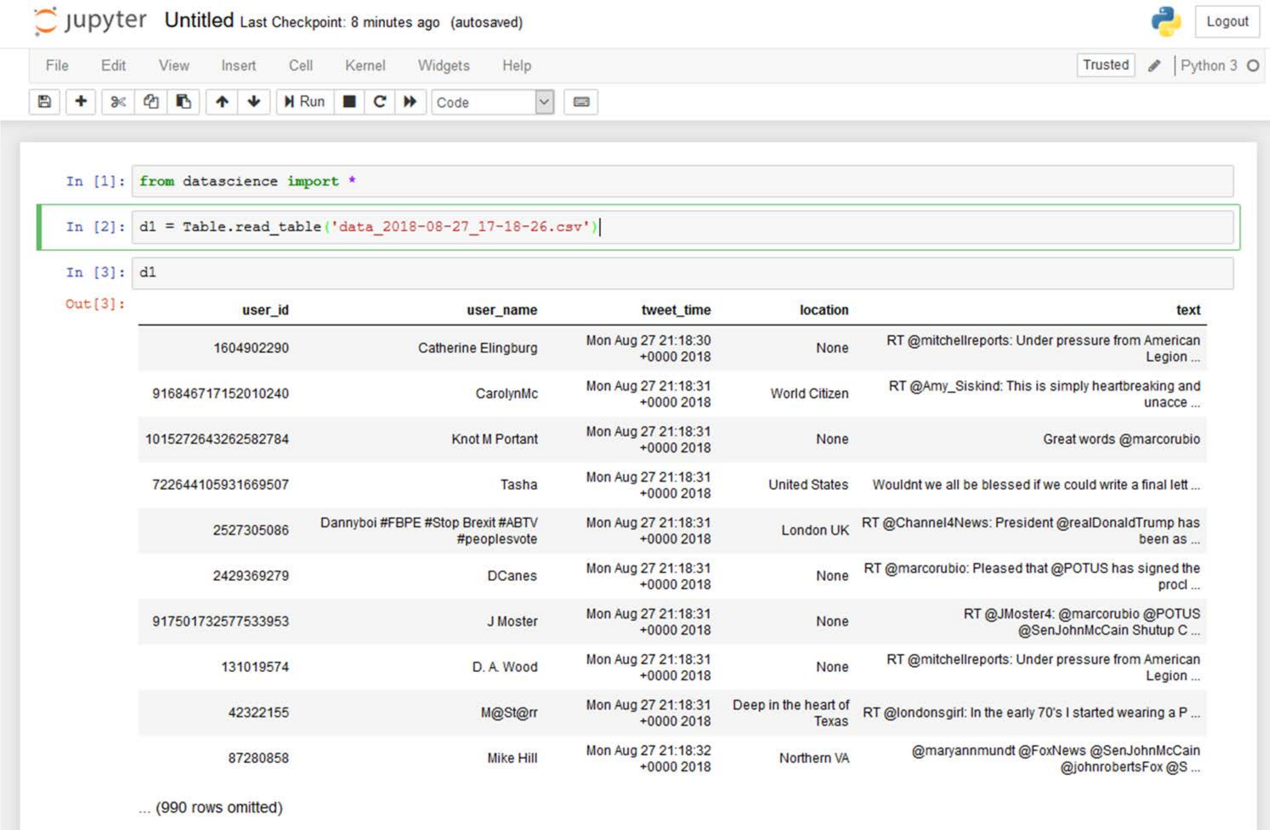
- In Python, strings are inside a pair of single quotes.
- The API keys you used to obtain Twitter approval for gathering tweets are strings.
- A string can contain "space", punctuation marks, etc.
- The



```
        print(status)
        return True


# Get the str representation of the current date and time
def current_datetime_str():
    return f'{datetime.datetime.now():%Y-%m-%d_%H-%M-%S}'


# Main
def main():
    # Key and token info needed
    consumer_key = ''
    consumer_secret = ''
    access_token = ''
    access_secret = ''

    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_secret)
    api = tweepy.API(auth)

    # Welcome
    print('===========================================================')
    print('Welcome to the user interface of gathering tweets pipeline!')
    print('You can press "Ctrl+C" at anytime to abort the program.')
    print('===========================================================')
    print()
```

15 What components of a tweet is a string?

a) user_id
b) User_name
c) Location
d) Text (tweet content)
e) All of above



```
In [1]: from datascience import *

In [2]: d1 = Table.read_table('data_2018-08-27_17-18-26.csv')

In [3]: d1

Out[3]:
```

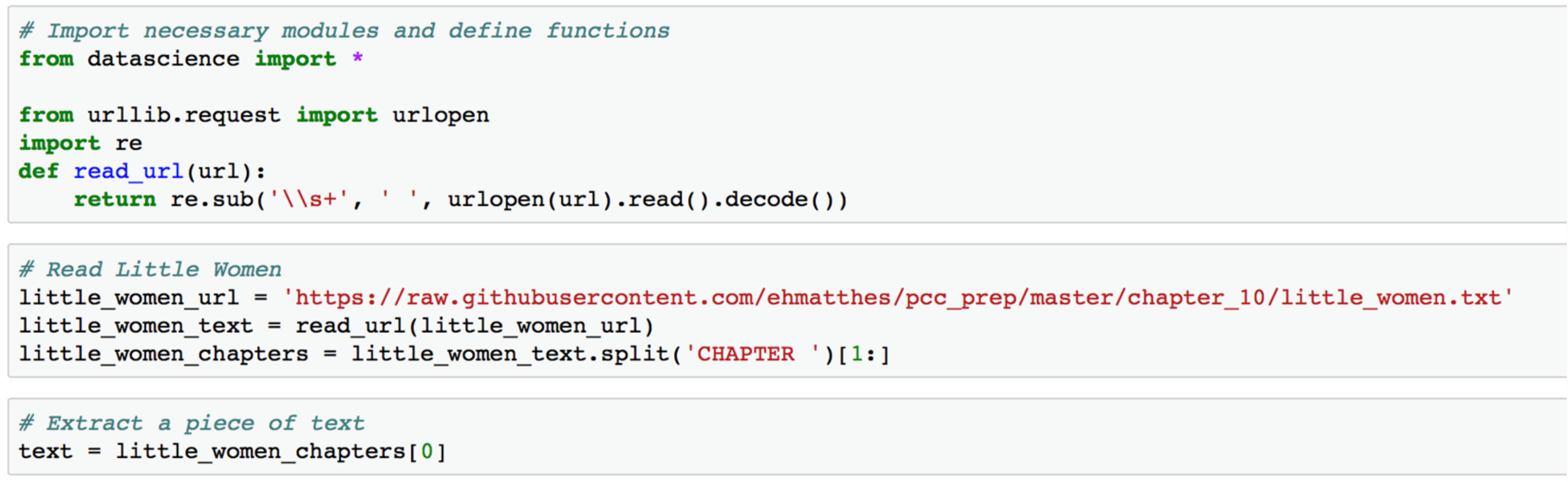
user_id	user_name	tweet_time	location	text
1604902290	Catherine Elingburg	Mon Aug 27 21:18:30 +0000 2018	None	RT @mitchellreports: Under pressure from American Legion ...
916846717152010240	CarolynMc	Mon Aug 27 21:18:31 +0000 2018	World Citizen	RT @Amy_Siskind: This is simply heartbreaking and unacce ...
1015272643262582784	Knot M Portant	Mon Aug 27 21:18:31 +0000 2018	None	Great words @marcorubio
722644105931669507	Tasha	Mon Aug 27 21:18:31 +0000 2018	United States	Wouldnt we all be blessed if we could write a final lett ...
2527305086	Dannyboi #FBPE #Stop Brexit #ABTV #peoplesvote	Mon Aug 27 21:18:31 +0000 2018	London UK	RT @Channel4News: President @realDonaldTrump has been as ...
2429369279	DCanes	Mon Aug 27 21:18:31 +0000 2018	None	RT @marcorubio: Pleased that @POTUS has signed the procl ...
917501732577533953	J Moster	Mon Aug 27 21:18:31 +0000 2018	None	RT @JMoster4: @marcorubio @POTUS @SenJohnMcCain Shutup C ...
131019574	D. A. Wood	Mon Aug 27 21:18:31 +0000 2018	None	RT @mitchellreports: Under pressure from American Legion ...
42322155	M@St@rr	Mon Aug 27 21:18:31 +0000 2018	Deep in the heart of Texas	RT @londonsgirl: In the early 70's I started wearing a P ...
87280858	Mike Hill	Mon Aug 27 21:18:32 +0000 2018	Northern VA	@maryannmundt @FoxNews @SenJohnMcCain @johnrobertsFox @S ...

... (990 rows omitted)

16 Split: A Method to Parse a String into a A List of Substrings

- Split is a method that is applied to a string, returns a list of substrings.
- It takes a parameter, which is the delimiter used to separate a string into a list of substrings.
- Ex:
- x = ‘I have a dream.’
- y = x.split(‘ ‘)
- The delimiter is the space (‘ ‘) character.
- X.split is the syntax of invoking a Method on a STRING object (X)
- The list of words obtained by splitting the string x (using space as the delimiter) is assigned to y as its value.
- Type the two assignments above to Jupyter Notebook. What is the value of print(Y) ?

17 Parse the text of “Little Women”

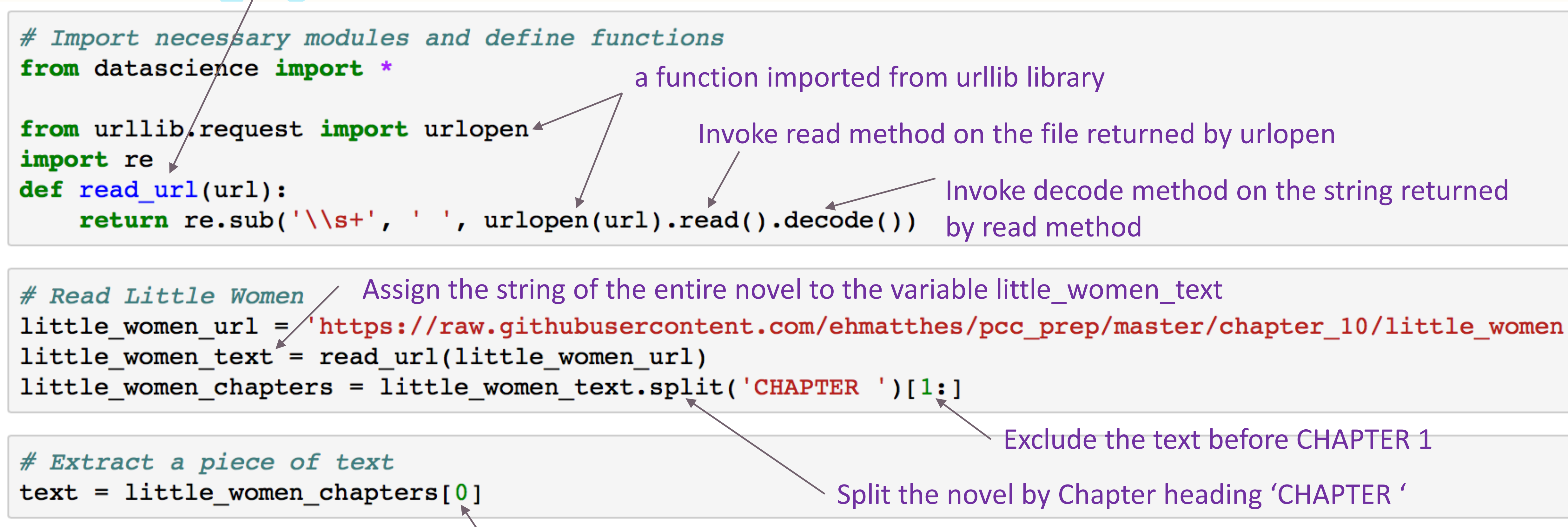


```
# Import necessary modules and define functions
from datascience import *

from urllib.request import urlopen
import re
def read_url(url):
    return re.sub('\\s+', ' ', urlopen(url).read().decode())
```

```
# Read Little Women
little_women_url = 'https://raw.githubusercontent.com/ehmatthes/pcc_prep/master/chapter_10/little_women.txt'
little_women_text = read_url(little_women_url)
little_women_chapters = little_women_text.split('CHAPTER ')[1:]
```

```
# Extract a piece of text
text = little_women_chapters[0]
```

Define a function that takes url as input parameter



```
# Import necessary modules and define functions
from datascience import *

from urllib.request import urlopen
import re
def read_url(url):
    return re.sub('\\s+', ' ', urlopen(url).read().decode())
```

```
# Read Little Women
little_women_url = 'https://raw.githubusercontent.com/ehmatthes/pcc_prep/master/chapter_10/little_women
little_women_text = read_url(little_women_url)
little_women_chapters = little_women_text.split('CHAPTER ')[1:]
```

```
# Extract a piece of text
text = little_women_chapters[0]
```

a function imported from urllib library

Invoke read method on the file returned by urlopen

Invoke decode method on the string returned by read method

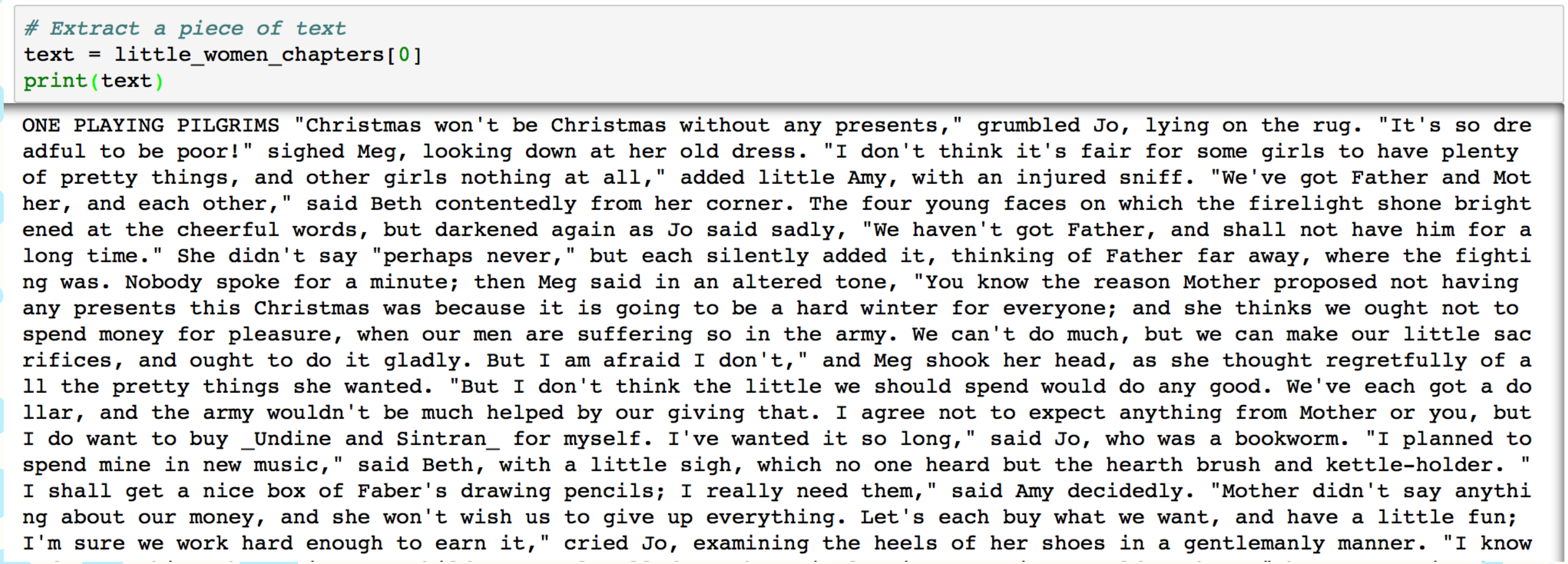
Assign the string of the entire novel to the variable little_women_text

Exclude the text before CHAPTER 1

Split the novel by Chapter heading ‘CHAPTER ‘

Text of the first chapter

19 Load text of one chapter



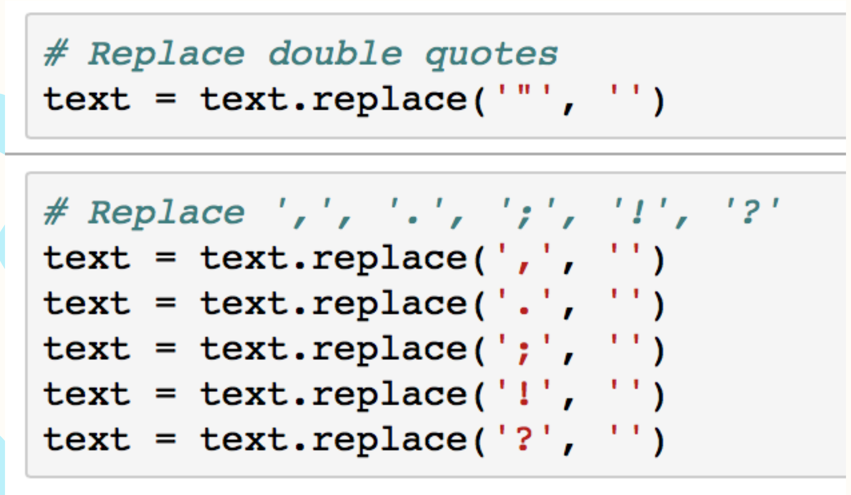
```
# Extract a piece of text
text = little_women_chapters[0]
print(text)
```

```
ONE PLAYING PILGRIMS "Christmas won't be Christmas without any presents," grumbled Jo, lying on the rug. "It's so dre
adful to be poor!" sighed Meg, looking down at her old dress. "I don't think it's fair for some girls to have plenty
of pretty things, and other girls nothing at all," added little Amy, with an injured sniff. "We've got Father and Mot
her, and each other," said Beth contentedly from her corner. The four young faces on which the firelight shone bright
ened at the cheerful words, but darkened again as Jo said sadly, "We haven't got Father, and shall not have him for a
long time." She didn't say "perhaps never," but each silently added it, thinking of Father far away, where the fighti
ng was. Nobody spoke for a minute; then Meg said in an altered tone, "You know the reason Mother proposed not having
any presents this Christmas was because it is going to be a hard winter for everyone; and she thinks we ought not to
spend money for pleasure, when our men are suffering so in the army. We can't do much, but we can make our little sac
rifices, and ought to do it gladly. But I am afraid I don't," and Meg shook her head, as she thought regretfully of a
ll the pretty things she wanted. "But I don't think the little we should spend would do any good. We've each got a do
llar, and the army wouldn't be much helped by our giving that. I agree not to expect anything from Mother or you, but
I do want to buy _Undine and Sintran_ for myself. I've wanted it so long," said Jo, who was a bookworm. "I planned to
spend mine in new music," said Beth, with a little sigh, which no one heard but the hearth brush and kettle-holder. "
I shall get a nice box of Faber's drawing pencils; I really need them," said Amy decidedly. "Mother didn't say anythi
ng about our money, and she won't wish us to give up everything. Let's each buy what we want, and have a little fun;
I'm sure we work hard enough to earn it," cried Jo, examining the heels of her shoes in a gentlemanly manner. "I know
```

Preprocessing: Remove punctuations

– Preprocessing of text remove things not needed for later processing. Ex: Punctuations are not needed for counting frequency of words.

– Remove punctuations, including **“ ” , . ; ! ?**



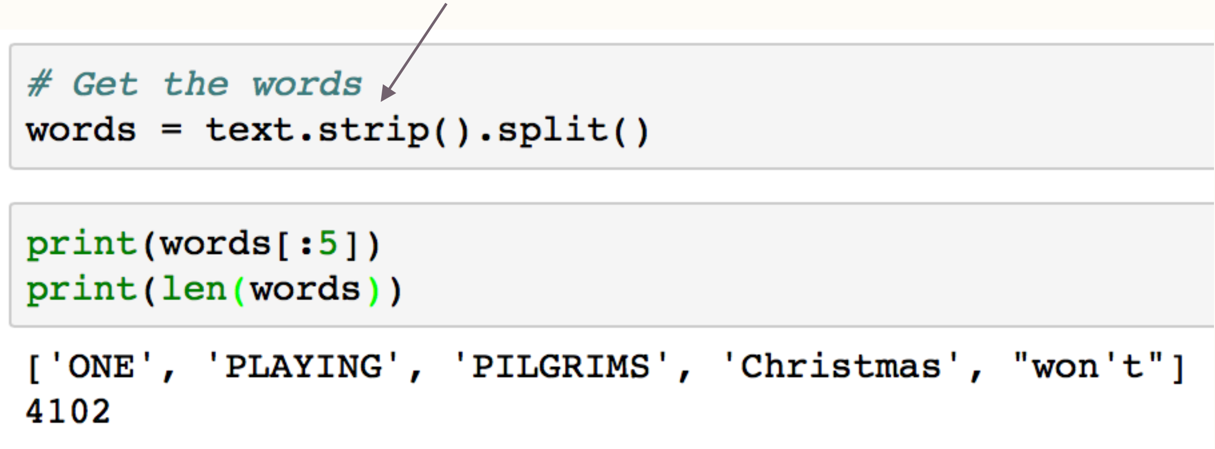
```
# Replace double quotes
text = text.replace('"', '')
```

```
# Replace ',', '.', ';', '!', '?'
text = text.replace(',', '')
text = text.replace('.', '')
text = text.replace(';', '')
text = text.replace('!', '')
text = text.replace('?', '')
```

Here, we use ***replace*** function

21 Extracting words in a chapter

Remove punctuations (comma ”,”, period “.”) immediately following a word.



```
# Get the words
words = text.strip().split()
```

```
print(words[:5])
print(len(words))
```

```
['ONE', 'PLAYING', 'PILGRIMS', 'Christmas', "won't"]
4102
```

22 Homework Exercise

- Save the Jupyter Notebook
- Reload the Jupyter Notebook, modify the chapter index in the assignment to extract text of a different chapter.
- text = little_women_chapters[0]
- Execute the modified Jupyter Notebook to find out total number of words in a chapter of your choice.

Use Vlab to run Jupyter Notebook & Python

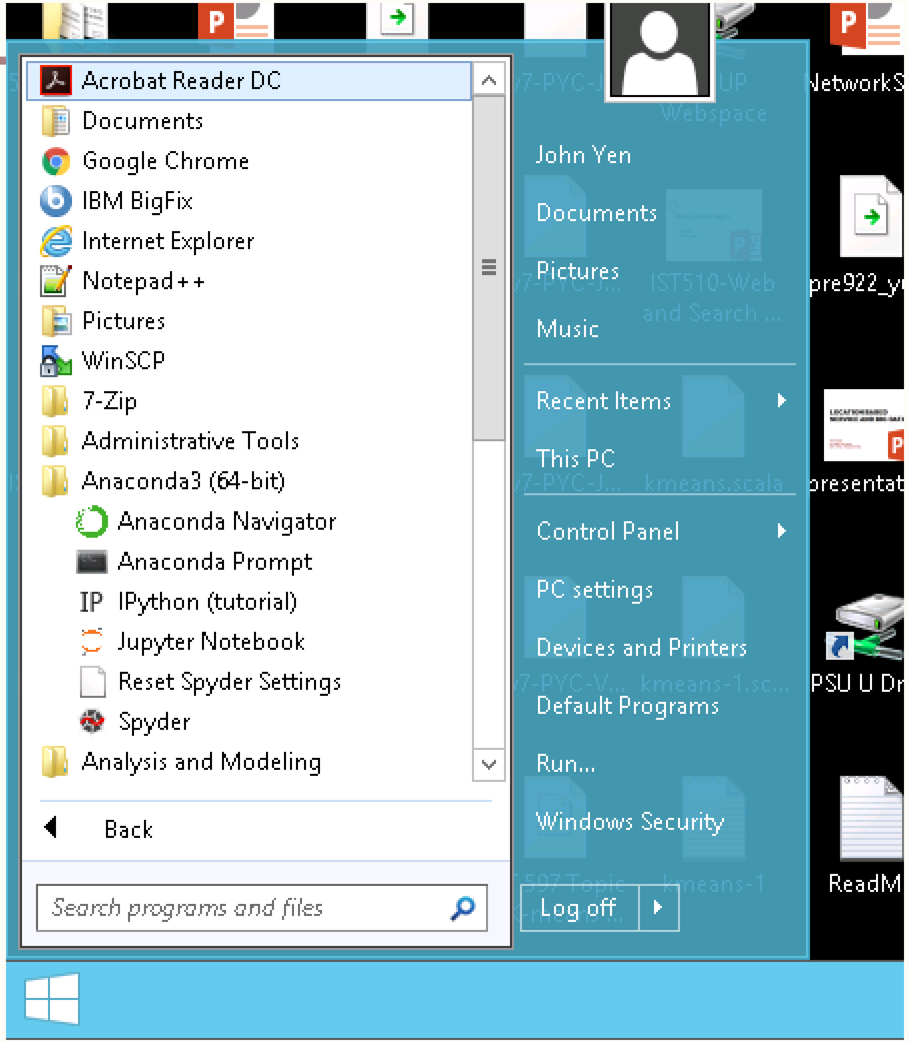
- Use a browser to go to http://bit.ly/psuds200
- Sign in using your PSU access ID and password

You will see a Window-like interface (for the Vlab)

- Click on the Window icon on the lower left, select "All Programs", select "Anaconda3 (64-bit)", select Jupyter Notebook

- Click New → Python 3

- Type import datascience then click Run to test whether you can import the datascience module.


Acrobat Reader DC
Documents
Google Chrome
IBM BigFix
Internet Explorer
Notepad++
Pictures
WinSCP
7-Zip
Administrative Tools
Anaconda3 (64-bit)
Anaconda Navigator
Anaconda Prompt
IPython (tutorial)
Jupyter Notebook
Reset Spyder Settings
Spyder
Analysis and Modeling
Back
Search programs and files
John Yen
Documents
Pictures
Music
Recent Items
This PC
Control Panel
PC settings
Devices and Printers
Default Programs
Run...
Windows Security
Log off
NetworkS
pre922_y
presentat
PSU U Dr
ReadM