

LECTURE 6

MINING WEB

CONTENT III

LEK HSIANG HUI

OUTLINE

Web Application Design

Access APIs using Python Packages

Access APIs without using Python Packages

Scraping using an actual browser/headless browser

RECAP: TECHNIQUES FOR WEB SCRAPING

The following are some of the techniques for doing web scraping:

- Extracting content from HTML source
- Extracting content using a HTML parser
- **Web Scraping using APIs**
- Scraping using an actual browser/headless browser

WEB APPLICATION DESIGN



Web
Application
Design

Access APIs
in Python

Scraping
using an
actual
browser

CLIENT/SERVER

The 2 main entities in a web environment are:

- **Client**
 - Web Browser, Mobile App, Wearables, etc
- **Server**
 - Web Server

WEB DEVELOPMENT

Even though Client/Server is still being used, the architecture of websites/web applications has evolved quite a bit

- Traditionally, users access the services through a web browser
- Now, users need not access the services through a web browser and the page/view can be dynamically generated on the client-side

TRADITIONAL WEB DEVELOPMENT

Traditional web development

- User requests a page from the server
- Server figures out what the user wants and generates the page dynamically into a HTML document before sending to user
- Client (web browser) displays the HTML document

MODERN WEB DEVELOPMENT

Modern web development

- Modern web development usually involves the use of **Application Programming Interface (API)**

Web API provides a mechanism for clients (browser, mobile app, etc) **to consume services or extend the capability of the site**

- Example:
Consume services: get the first 20 records, add a Weibo post, etc
Extend capability: new mobile app to access social media, etc

MODERN WEB DEVELOPMENT

Most web API uses a data exchange format for communication rather than sending HTML (resulting in lesser data transfer – faster)


- Data exchange formats: XML, **JSON** (more common, to be discussed later)

Some APIs are created and used only by the creator (private API**), where others are meant to be used for anyone (**public API**)**


PRIVATE API USAGE



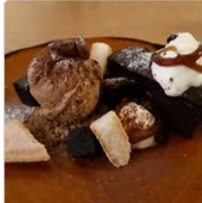
Porta **203 Reviews**
Clarke Quay · ~\$25/pax
1-for-1 Deals, Date Night, Breakfast & Brunch, Dinner with Drinks, Western, Euro...
BEYOND Hot 100 2019 Featured In 3 Guides



Sweet Potato and Kale Salad (Part Of 3 Course)
 Teo




[Porta] - Duck Ragout (\$24). The original version
 SG




Not-your-usual Tiramisu Tiramisu deconstructed
 Loong Wye

[See More >](#)


Shake Shack (Jewel Changi Airport) **90 Reviews**
Changi · ~\$15/pax
Western, Burgers, Newly Opened



Shake Shack! We finally got to try shake shack at
 Si Min




Shark attack Concrete Ice cream with lots of
 K



Shack Stack Burger Beef petty with shroom patty,
 K


[See More >](#)

Brine **120 Reviews**
Bugis · ~\$25/pax
1-for-1 Deals, Date Night, Breakfast & Brunch, Cafes & Coffee, Western
BEYOND Featured In 1

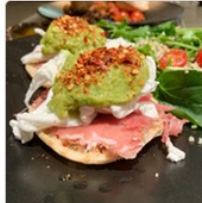


Banana Confit B...
confit, steamed b...
 Edith


The Coffee A **404 Reviews**
Orchard · ~\$30/pax
1-for-1 Deals, Breakfast
BEYOND Hot 100 2019



Fish Taco Crispy skin, mango salsa with fried
 Hsu



Confusing Eggs Benedict For some reason, the
 nic



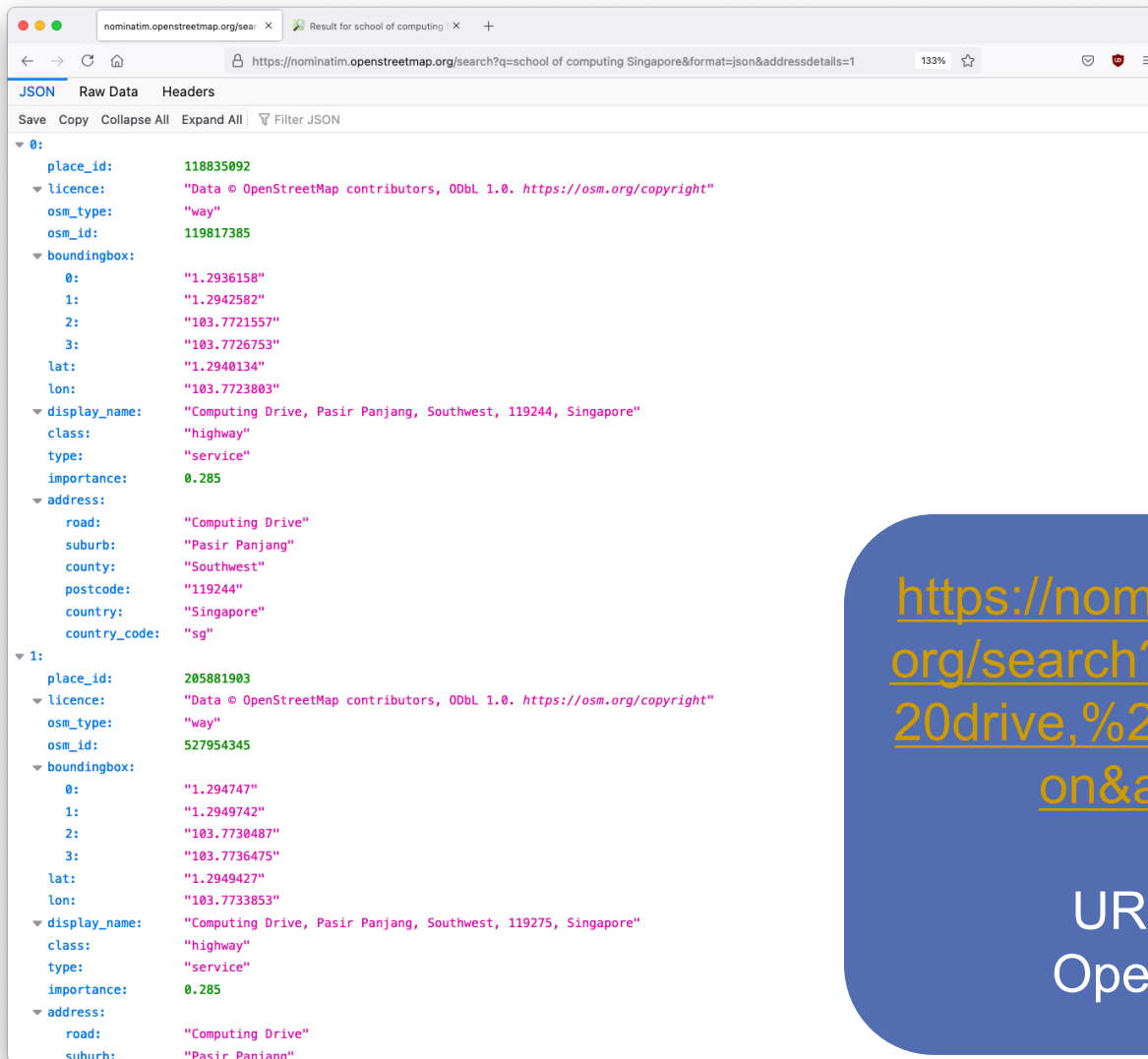
Quiet Brunch Place In Orchard Ordered the egg
 May

[See More >](#)

When the user clicks on LOAD MORE, the JavaScript will fetch the data using AJAX (Aynchronous JavaScript And XML) and populate new records

LOAD MORE ▾

PUBLIC API USAGE



The screenshot shows a web browser window with the URL `https://nominatim.openstreetmap.org/search?q=school of computing Singapore&format=json&addressdetails=1`. The page displays two JSON objects representing search results. The first object (index 0) has a `place_id` of 118835092 and a `display_name` of "Computing Drive, Pasir Panjang, Southwest, 119244, Singapore". The second object (index 1) has a `place_id` of 205881903 and a `display_name` of "Computing Drive, Pasir Panjang, Southwest, 119275, Singapore". Both objects include details like `licence`, `osm_type`, `osm_id`, `boundingbox`, `lat`, `lon`, `class`, `type`, `importance`, and `address`.

```
{
  "0": {
    "place_id": 118835092,
    "licence": "Data \u2122 OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
    "osm_type": "way",
    "osm_id": 119817385,
    "boundingbox": [
      1.2936158,
      1.2942582,
      103.7721557,
      103.7726753
    ],
    "lat": 1.2940134,
    "lon": 103.7723803,
    "display_name": "Computing Drive, Pasir Panjang, Southwest, 119244, Singapore",
    "class": "highway",
    "type": "service",
    "importance": 0.285,
    "address": {
      "road": "Computing Drive",
      "suburb": "Pasir Panjang",
      "county": "Southwest",
      "postcode": "119244",
      "country": "Singapore",
      "country_code": "sg"
    }
  },
  "1": {
    "place_id": 205881903,
    "licence": "Data \u2122 OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright",
    "osm_type": "way",
    "osm_id": 527954345,
    "boundingbox": [
      1.294747,
      1.2949742,
      103.7730487,
      103.7736475
    ],
    "lat": 1.2949427,
    "lon": 103.7733853,
    "display_name": "Computing Drive, Pasir Panjang, Southwest, 119275, Singapore",
    "class": "highway",
    "type": "service",
    "importance": 0.285,
    "address": {
      "road": "Computing Drive",
      "suburb": "Pasir Panjang"
    }
  }
}
```

<https://nominatim.openstreetmap.org/search?q=13%20computing%20drive,%20Singapore&format=json&addressdetails=1>

URL to access the
OpenStreetMap API

PUBLIC API USAGE

nomatim.openstreetmap.org/search Result for school of computing

https://nominatim.openstreetmap.org/ui/search.html?q=school+of+computing+Singapore 133%

Nominatim Search Reverse Search By ID About & Help

Simple Structured

school of computing Singapore Search

Advanced options


Data from API request (debug output)

Computing Drive, Pasir Panjang, Southwest, 119244, Singapore Service details

Computing Drive, Pasir Panjang, Southwest, 119275, Singapore Service

Search for more results

Data last updated: 1 minute ago (Details)



Addresses and postcodes are approximate

© OpenStreetMap contributors

nomatim.openstreetmap.org/search Result for school of computing

https://nominatim.openstreetmap.org/ui/search.html?q=school+of+computing+Singapore 133%

JSON Raw Data Headers

```
Save Copy Collapse All Expand All Filter: JSON
```

```
{
  "place_id": 118054902,
  "licence": "Data © OpenStreetMap contributors, OSM I.D. https://osm.org/copyright",
  "osm_type": "way",
  "osm_id": 119017385,
  "boundingbox": [
    0,
    1,
    2,
    3
  ],
  "lat": "1.2846234",
  "lon": "103.7733893",
  "display_name": "Computing Drive, Pasir Panjang, Southwest, 119244, Singapore",
  "class": "highway",
  "type": "service",
  "importance": 0.285,
  "address": {
    "road": "Computing Drive",
    "suburb": "Pasir Panjang",
    "county": "Southwest",
    "postcode": "119244",
    "country": "Singapore",
    "country_code": "sg"
  }
}
```

```
{
  "place_id": 289581003,
  "licence": "Data © OpenStreetMap contributors, OSM I.D. https://osm.org/copyright",
  "osm_type": "way",
  "osm_id": 527954345,
  "boundingbox": [
    0,
    1,
    2,
    3
  ],
  "lat": "1.2846234",
  "lon": "103.7733893",
  "display_name": "Computing Drive, Pasir Panjang, Southwest, 119275, Singapore",
  "class": "highway",
  "type": "service",
  "importance": 0.285,
  "address": {
    "road": "Computing Drive",
    "suburb": "Pasir Panjang"
  }
}
```

Public API can be used programmatically, and at the same time to power websites

EXTERNAL API USAGE

JSONRaw DataHeaders

SaveCopyCollapse AllExpand AllFilter JSON

▼ 0:

place_id:118835092

▼ licence:"Data © OpenStreetMap contributors, ODbL 1.0. [https://openstreetmap.org/help/faq-fair-use](#)"

osm_type:"way"

osm_id:119817385

▼ boundingbox:

0:"1.2936158"

1:"1.2942582"

2:"103.7721557"

3:"103.7726753"

lat:"1.2940134"

lon:"103.7723803"

▼ display_name:"Computing Drive, Pasir Panjang, Southwest, 119244, Singapore"

class:"highway"

type:"service"

importance:0.285

▼ address:

road:"Computing Drive"

suburb:"Pasir Panjang"

county:"Southwest"

postcode:"119244"

country:"Singapore"

country_code:"sg"

Example
data ex

It allows
represe
format to
st

Example of the JSON data exchange format

It allows any data to be represented in a text format together with its structure

ACCESS APIS IN PYTHON PACKAGES



Web
Application
Design

Access APIs
in Python

Scraping
using an
actual
browser

ACCESSING API USING PYTHON PACKAGES

Before you try to work with an API manually, could check whether there is a package to access the API

We will use the **OSMPythonTools** package as an example

- OSMPythonTools allow us to access the OpenStreetMap services
- <https://wiki.openstreetmap.org/wiki/OSMPythonTools>

ACCESSING API MANUALLY

Not all site has a package that integrates with the API

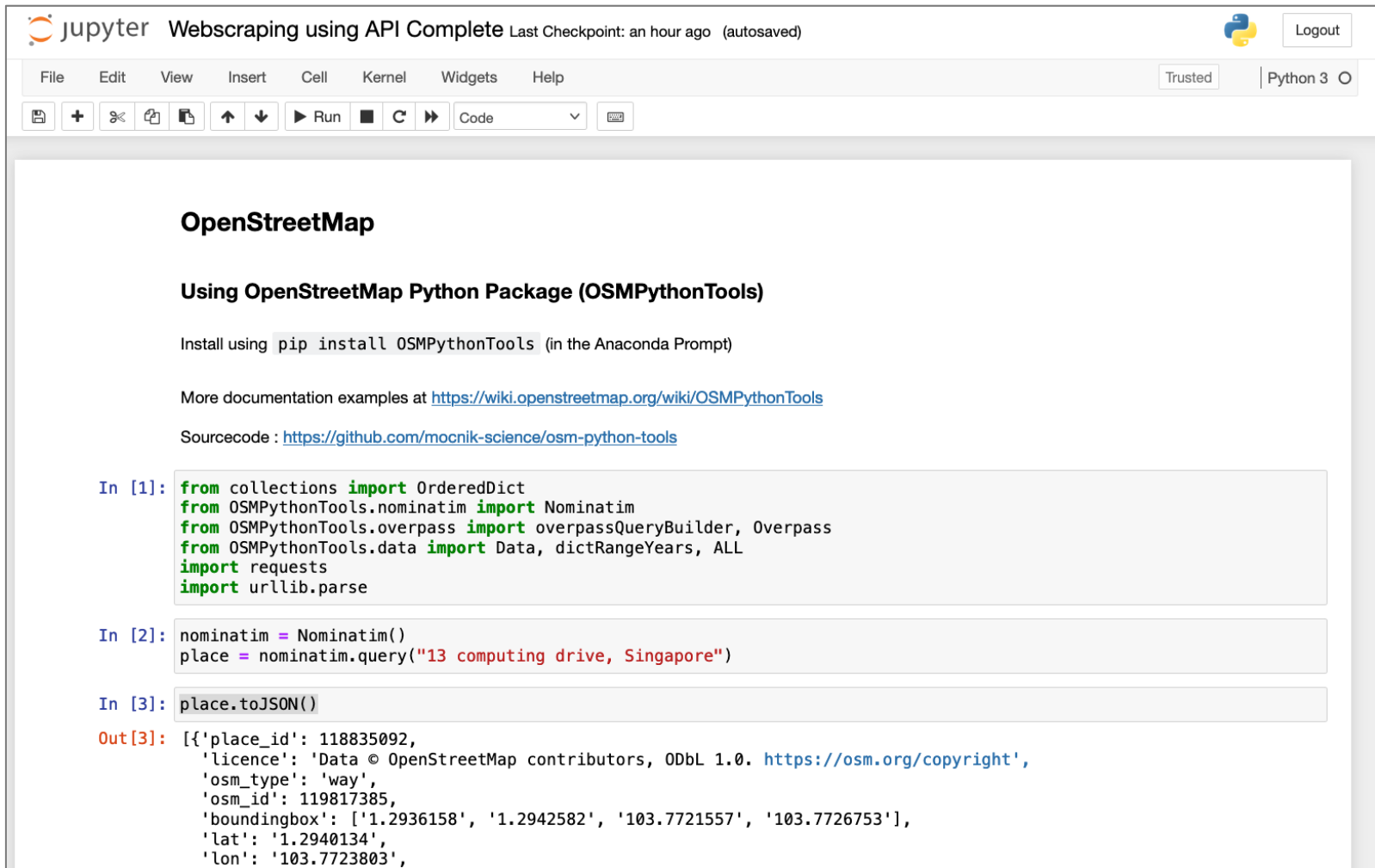
- Furthermore, each time when an API changes, it might potentially break the packages that using with the API

Better to learn how to access APIs manually without packages

- Need to learn how to work with JSON output

HANDS-ON: WEB SCRAPING USING API

Download and access:
[Webscrapping using API.ipynb](#)



The screenshot shows a Jupyter Notebook interface with the title "Webscrapping using API Complete". The notebook contains the following content:

OpenStreetMap

Using OpenStreetMap Python Package (OSMPythonTools)

Install using `pip install OSMPythonTools` (in the Anaconda Prompt)

More documentation examples at <https://wiki.openstreetmap.org/wiki/OSMPythonTools>

Sourcecode : <https://github.com/mocnik-science/osm-python-tools>

```
In [1]: from collections import OrderedDict
        from OSMPythonTools.nominatim import Nominatim
        from OSMPythonTools.overpass import overpassQueryBuilder, Overpass
        from OSMPythonTools.data import Data, dictRangeYears, ALL
        import requests
        import urllib.parse

In [2]: nominatim = Nominatim()
        place = nominatim.query("13 computing drive, Singapore")

In [3]: place.toJSON()

Out[3]: [{'place_id': 118835092,
          'licence': 'Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright',
          'osm_type': 'way',
          'osm_id': 119817385,
          'boundingbox': ['1.2936158', '1.2942582', '103.7721557', '103.7726753'],
          'lat': '1.2940134',
          'lon': '103.7723803',
```

CONCLUDING NOTES ON API

Should use API as far as possible instead of other approaches

- Lesser data payload → faster

****But take note that most API adopts rate limiting**

- Should not violate the rate limits if not you might be banned!
- Protip: If the service you are accessing requires you to authenticate, use another dummy account

SCRAPING USING AN ACTUAL BROWSER



Web
Application
Design

Access APIs
in Python

Scraping
using an
actual
browser

RECAP: TECHNIQUES FOR WEB SCRAPING

The following are some of the techniques for doing web scraping:

- Extracting content from HTML source
- Extracting content using a HTML parser
- Web Scraping using APIs
- **Scraping using an actual browser/headless browser**

If the site loads contents dynamically and does not offer an API, you will have to use the last technique

SCRAPING USING ACTUAL BROWSER/HEADLESS BROWSER

This is the most powerful technique

- *“Anything that you see on the browser can be scraped”*

Not all site offers an API and sometimes the page is rendered dynamically using JavaScript, by looking at the HTML source code, you will not be able to extract the data

- They probably did not intend you to do web scraping on the site!

Headless browser = browser without a GUI

DISCLAIMER

It is a gray area whether you are allowed to do web scraping

- Most site would not allow you to do that
- But everyone is doing it anyways

Some things are outrightly illegal such as selling data that you do not own, etc

This section is meant for educational purposes, you are responsible for your own actions 😊

predictive analytics_百度搜索

https://www.baidu.com/s?ie=utf-8&f=8&rsv_bp=1&rsv_idx=1&tn=baidu&wd=predictive+analytics&fenlei=256&rsv_pq=fabfd20b0002c 133% ☆

Baidu 百度

predictive analytics

百度一下

百度首页 设置 登录

Q 网页 资讯 视频 图片 知道 文库 贴吧 地图 采购 更多

百度为您找到相关结果约5,550,000个

搜索工具


[Predictive Analytics Definition](#)

查看此网页的中文翻译, 请点击 [翻译此页](#)

2021年5月5日 **Predictive analytics** is the use of statistics and modeling techniques to determine future performance based on current and historical data. **Predictive...**


[www.investopedia.com/terms/p/p... 百度快照](#)

[商业预测分析\(Predictive Analytics\) - 简书](#)

 2016年4月18日 商业预测分析(Predictive Analytics) 人的行为是有模式的,比如购买行为。模式通常难以打破,但人生非不变。人生总有特殊时刻,这时模式容易发生改变,商家的机会就在这个时刻。所以预测...

[简书社区 百度快照](#)

[为什么预测分析\(Predictive Analytics\)是人力资本管理改变...](#)

 2018年9月28日 最后,预测性人力资源分析将创造有意义和令人满意的员工体验,带来真正的长期利益和关系建设。 以上为AI翻译,观点仅供参考。 原文链接:Why **Predictive Analytics** is a Game Changer fo...

[知乎 百度快照](#)

[Predictive Analytics | IBM](#)

2021年1月21日 Analyze data and build **analytics** models to **predict** future outcomes. L... s and opportunities for your business.

[www.ibm.com/analytics/predicti... 保障 百度快照 - 翻译此页](#)

[Predictive Analytics | 及时分析数据,采取行动 | Micro F...](#)

Transform volumes of high-growth disparate data into accurate and actionable insights **predictive analytics** at scale.

[www.microfocus.com/trend/predi... 百度快照](#)

其他人还在搜

[diagnostic analytics](#) [tableau数据可视化](#) [google analytics是什么意思](#)

[data mapping是什么意思](#) [connectomics](#) [collection strategy](#) [initial load](#)

百度热搜

换一换

- 1 全国新冠疫苗接种剂次超9亿 **热**
- 2 深圳机场一员工确诊 密接者87人
- 3 中方回应“G7公报对中国横加指责”
- 4 外交部说中俄合作上不封顶下接地气 **新**
- 5 神舟十二号载人飞行任务标识发布
- 6 江西专升本作弊:多名大学教师被刑拘
- 7 河南货车侧翻致8死11伤:21人被追责
- 8 中国第一股民“杨百万”去世
- 9 世界最大家庭户主在印度去世
- 10 TFBOYS将解散系谣言

Suppose we want to scrap data off the Baidu search engine

predictive analytics_百度搜索

https://www.baidu.com/s?ie=utf-8&f=8&rsv_bp=1&rsv_idx=1&tn=baidu&wd=predictive analytics&fenlei=256&rsv_pq=fabfd20b0002c 133% ☆

Baidu 百度

predictive analytics 百度一下

百度首页 设置 登录

Q 网页 资讯 视频 图片 知道 文库 贴吧 地图 采购 更多

百度为您找到相关结果约5,550,000个 搜索工具

Predictive Analytics Definition

查看此网页的中文翻译, 请点击 翻译此页

2021年5月5日 **Predictive analytics** is the use of statistics and modeling techniques to determine future performance based on current and historical data. **Predictive...**

www.investopedia.com/terms/p/p... 百度快照

商业预测分析(Predictive Analytics) - 简书

2016年4月18日 商业预测分析(Predictive Analytics) 人的行为是有模式的,比如购买行为。模式通常难以打破,但人生非不变。人生总有特殊时刻,这时模式容易发生改变,商家的机会就在这个时刻。所以预测...

简书社区 百度快照

为什么预测分析(Predictive Analytics)是人力资本管理改变...

2018年9月28日 最后,预测性人力资源分析将创造有意义和令人满意的员工体验,带来真正的长期利益和关系建设。以上为AI翻译,观点仅供参考。原文链接:Why **Predictive Analytics** is a Game Changer fo...

知乎 百度快照

Predictive Analytics | IBM

2021年1月21日 Analyze data and build **analytics** models to **predict** future outcomes. s and opportunities for your business.

www.ibm.com/analytics/predicti... 保障 百度快照 - 翻译此页

Predictive Analytics | 及时分析数据,采取行动 | Micro F...

Transform volumes of high-growth disparate data into accurate and actionable insight **predictive analytics** at scale.

www.microfocus.com/trend/predi... 百度快照

其他人还在搜

diagnostic analytics tableau数据可视化 google analytics是什么意思 data mapping是什么意思 connectomics collection strategy initial load

百度热搜 换一换

- 1 全国新冠疫苗接种剂次超9亿 热
- 2 深圳机场一员工确诊 密接者87人
- 3 中方回应“G7公报对中国横加指责”
- 4 外交部说中俄合作上不封顶下接地气 新
- 5 神舟十二号载人飞行任务标识发布
- 6 江西专升本作弊:多名大学教师被刑拘
- 7 河南货车侧翻致8死11伤:21人被追责
- 8 中国第一股民“杨百万”去世
- 9 世界最大家庭户主在印度去世
- 10 TFBOYS将解散系谣言

The page is loaded based on what the user enters in the search box

SCRAPING USING ACTUAL BROWSER/HEADLESS BROWSER

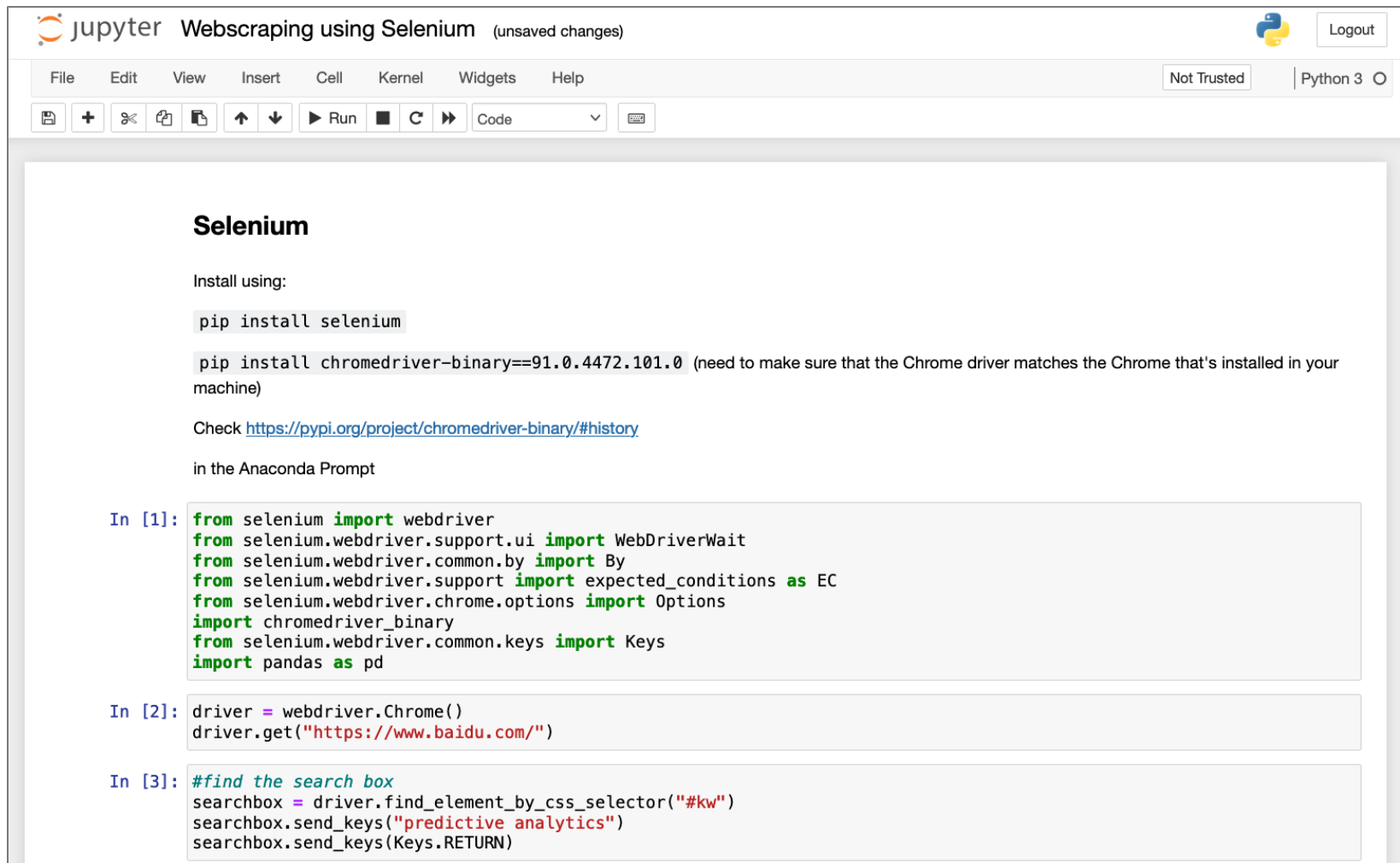
Idea:

- Load an instance of a browser
- Navigate the browser instance to the starting page
- Programmatically perform certain actions on the page
- The page is rendered accordingly on the browser instance
- Perform web scraping on the browser instance

To do this in Python, we can make use of the **selenium** package

HANDS-ON: WEB SCRAPING USING BROWSER

Download and access:
[Webscrapping using Selenium.ipynb](#)



The screenshot shows a Jupyter Notebook interface with the title 'Webscrapping using Selenium (unsaved changes)'. The top bar includes a menu (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), a 'Not Trusted' status, and 'Python 3'. Below the menu is a toolbar with icons for saving, adding cells, deleting, copying, pasting, undo, redo, and running code. The notebook content is as follows:

Selenium

Install using:

```
pip install selenium
```

`pip install chromedriver-binary==91.0.4472.101.0` (need to make sure that the Chrome driver matches the Chrome that's installed in your machine)

Check <https://pypi.org/project/chromedriver-binary/#history> in the Anaconda Prompt

```
In [1]: from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.common.by import By
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.chrome.options import Options
import chromedriver_binary
from selenium.webdriver.common.keys import Keys
import pandas as pd

In [2]: driver = webdriver.Chrome()
driver.get("https://www.baidu.com/")

In [3]: #find the search box
searchbox = driver.find_element_by_css_selector("#kw")
searchbox.send_keys("predictive analytics")
searchbox.send_keys(Keys.RETURN)
```

WEB SCRAPING USING BROWSER INSTANCE

Advantages:

- Pretty much can handle any website
- Simulate as if it were someone doing actual web-surfing but we can now programmatically grab contents from the page

Disadvantages:

- Slow (would want to open up an actual browser, load css, images, etc)
 - Can try disable loading of images

CONCLUDING NOTES ON USING SELENIUM

There are other technologies that runs much faster

- PhantomJS, Puppeteer (based on NodeJS)
- In my opinion, web scraper using NodeJS is much more natural (since NodeJS is Javascript (JS) and logic on the web is powered by JS)

Don't adopt Selenium as the first choice

- Generally, this approach of web scraping is slow (even if using NodeJS)
- Public API > Private API > HTML Parsing > Selenium

CONCLUDING NOTES ON USING SELENIUM

Practically speaking...

- Don't have to code every action, some actions which are difficult to code, you could manually do it on the browser
- Being good at Javascript/CSS selectors is probably more important than knowing how to code in Selenium
- Possible to do web scraping directly on the browser!

predictive analytics_百度搜索

baidu.com/s?ie=utf-8&f=8&rsv_bp=1&rsv_idx=1&tn=baidu&wd=predictive%20analytics&fenlei=256&rsv_pq=a046147f000cec43&...

Baidu 百度

predictive analytics

网页 资讯 视频 图片 知道 文库 贴吧

百度为您找到相关结果约5,550,000个

[Predictive Analytics Definition](#)

查看此网页的中文翻译, 请点击 翻译此页

2021年5月5日 Predictive analytics is the use of statistics and modeling techniques to forecast future performance based on current and historical data. Predictive...

[www.investopedia.com/terms/p/p... 百度快照](#)

[商业预测分析\(Predictive Analytics\) - 简书](#)

2016年4月18日 商业预测分析(Predictive Analytics) 人的行为模式通常难以打破,但人生非不变。人生... 这时模式容易发生改变,商家的机会就在这个时刻。所以预测...

简书社区 百度快照

7 钟薛高雪糕最贵一支66元 421

8 神舟十二号载人飞行任务标识发布 41

Elements Console Sources Network Performance Memory >> 7 Issues: 7

```
> links = $$("#content_left .result h3 a")
< ▶ (9) [a, a, a, a, a, a, a, a, a]
> links.map((e) => e.textContent)
< (9) ["Predictive Analytics Definition", "商业预测分析(Predictive Analytics) - 简书", "为什么预测分析(Predictive Analytics)是人力资本管理改变...", "Predictive Analytics | IBM", "Predictive Analytics | 及时分析数据,采取行动 | Micro F...", "Predictive Analytics | Oil & Gas | McKinsey & Company", "Predictive Analytics - an overview | ScienceDirect Top...", "Predictive Analytics (豆瓣)", "Predictive Analytics: What it is and why it matters | ..."]
> JSON.stringify(links.map((e) => e.textContent))
< "[\"Predictive Analytics Definition\", \"商业预测分析(Predictive Analytics) - 简书\", \"为什么预测分析(Predictive Analytics)是人力资本管理改变...\", \"Predictive Analytics | IBM\", \"Predictive Analytics | 及时分析数据,采取行动 | Micro F...\", \"Predictive Analytics | Oil & Gas | McKinsey & Company\", \"Predictive Analytics - an overview | ScienceDirect Top...\", \"Predictive Analytics (豆瓣)\", \"Predictive Analytics: What it is and why it matters | ...\"]"
>
```

Possible to do web scraping directly from browser

Copy and paste JSON string

SUMMARY

Web Application Design

Access APIs in Python

- Using packages and accessing using URL

Scraping using an actual browser/headless browser

WHAT'S NEXT?

Recommender System