Python數據擷取與分析暨ＡＩ人工智慧應用

Louis

Outline

* Python資料處理與儲存
* Python網頁 資料擷取與轉換

python™

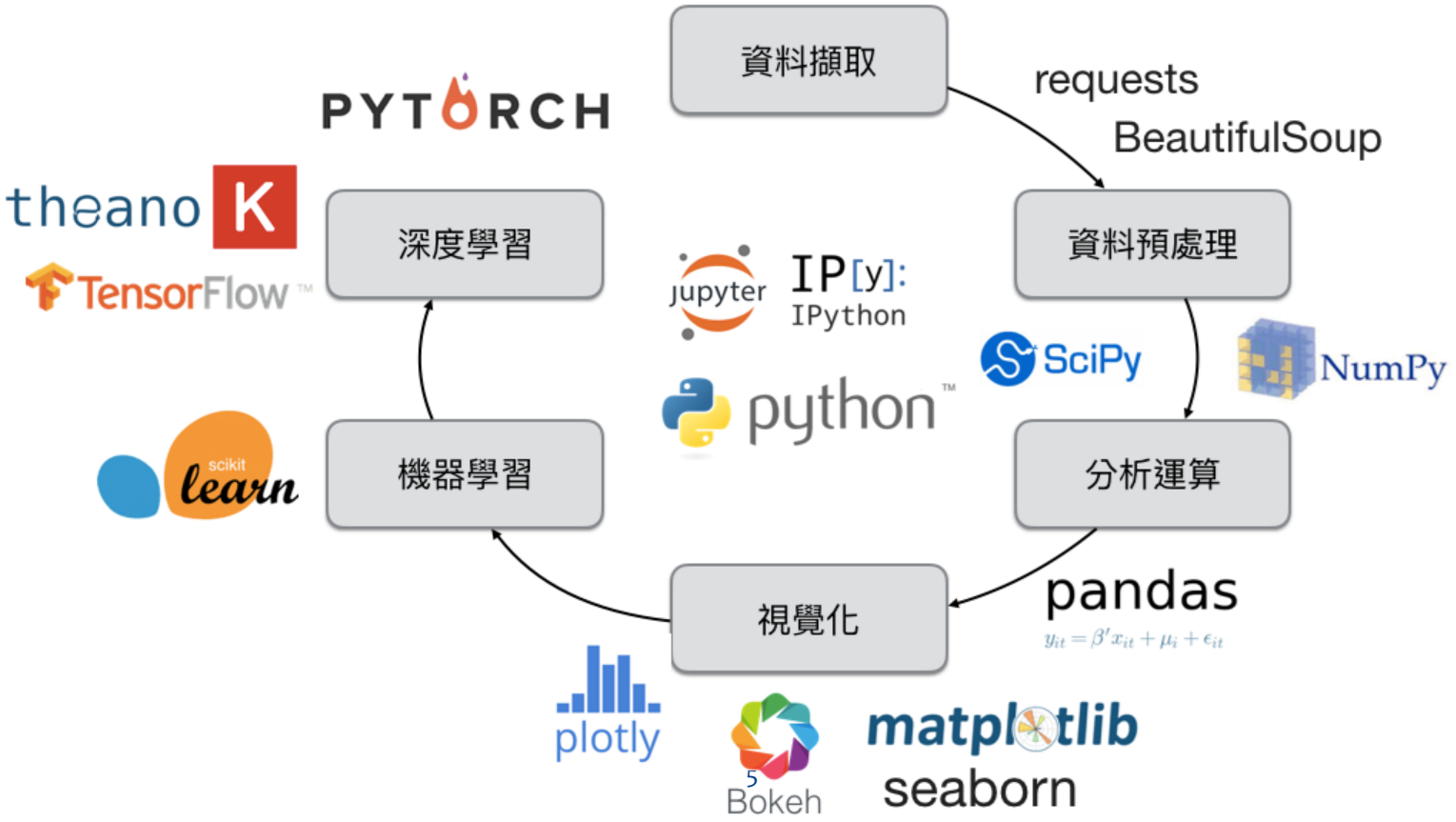
Python資料處理與儲存

Louis

Python資料處理與儲存

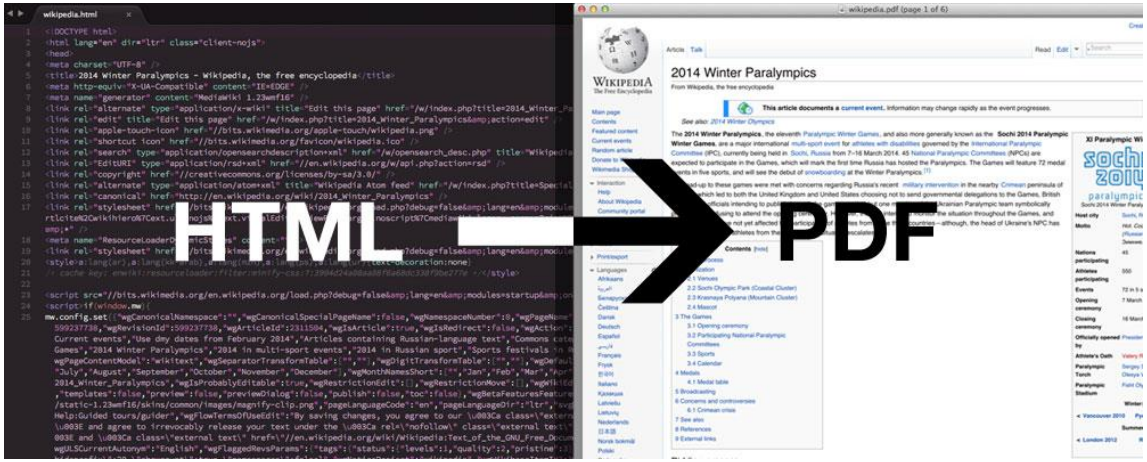
* Pdf資料處理
* CSV資料處理
* Json \ Yaml資料處理
* Xml資料處理
* sqlite3資料處理
* Mysql資料處理
* Google試算表處理

Python 應用範圍 - 資料分析



Pdf資料處理
安裝wkhtmltox

* 安裝wkhtmltox
  * 執行該執行檔，並記錄其安裝路徑



Anaconda install Python-pdfkit

* To install this package with conda run one of the following:
  pip install pdfkit

* 如果不行才繼續做以下動作
* conda install -c conda-forge python-pdfkit
  conda install -c conda-forge/label/gcc7 python-pdfkit
  conda install -c conda-forge/label/cf201901 python-pdfkit

Pdf資料處理-測試例子

* 抓取網頁、字串或是網頁檔→轉換成pdf
  * E1.py
* 讀寫pdf檔(pip install PyPDF2)
  * E2.py
* 讀寫pdf檔(從第3頁另存新檔)
  * E3.py
* 合併pdf檔
  * E4.py

CSV資料處理

* Csv檔的讀寫與置換內容(csv例子)
  * CSV_E1.py
  * CSV_E2.py
  * CSV_E3.py

Json \ Yaml資料處理

* Json檔的資料處理 (Json例子)
  * Json_E1.py
  * Json_E2.py
* Yaml 檔的資料處理 (Json例子)
  * Yaml_E1.py (E1-3-3-1.py)
  * Yaml_E2.py (E1-3-3-2.py)

Xml資料處理

* Xml檔的讀寫與置換內容(Xml 例子)
  * Xml_E1.py
  * Xml_E2.py
  * Xml_E3.py
  * Xml_E4.py

sqlite3資料處理

* sqlite檔的讀寫與置換內容(sqlite csv例子)
  * sqlite_E1.py
  * sqlite_E2.py
  * sqlite_E3.py
  * sqlite_E4.py
  * sqlite_E5.py
  * sqlite_E6.py

Mysql資料處理

* 開放源始碼的關聯式資料庫管理系統

* MySQL在過去由於效能高、成本低、可靠性好，已經成為最流行的開源資料庫，因此被廣泛地應用在Internet上的中小型網站中。

* 隨著MySQL的不斷成熟，它也逐漸用於更多大規模網站和應用，比如維基百科、Google和Facebook等網站。非常流行的開源軟體組合LAMP中的「M」指的就是MySQL

Google試算表處理

* Google 試算表是一款線上版的試算表應用程式，可讓您建立試算表並設定格式，以及和他人協同合作。

python™

參考資料

* 維基百科，自由的百科全書
* TQC+ Python3.x 網頁資料擷取與分析 特訓教材

Python網頁資料擷取與轉換

Louis

Python網頁資料擷取與轉換

* Requests
* BeautifulSoup
* Pandas
* Requests & BeautifulSoup & Pandas Ex
* Urllib
* Urllib Ex
* Json & sqlalchemy
* Requests & Json & sqlalchemy Ex
* Selenium & Selenium Ex
* 資料視覺化Matplotlib
* 綜合應用

Requests

* 導入 Requests 模組：
  * >>> import requests
* 獲取某個網頁。
  * Github 的公共時間線：
  * >>> gevent = requests.get('https://api.github.com/events')
* gevent 的 Response 對象。可以從這個物件中獲取網頁中想要的資訊。

Requests_httpbin

* A simple HTTP Request & Response Service(httpbin.org)
  * 可向httpbin.org發送請求，會照指定的規則返回該請求的回覆。就像是echo伺服器
  * 支持HTTP/HTTPS，支持所有的HTTP動詞，亦可以返回一個HTML檔或一個XML檔或一個圖片檔！

Requests

* Requests –API
* >>> gevent = requests.post('http://httpbin.org/post', data = {'key':'value'})
* >>> gevent = requests.put('http://httpbin.org/put', data = {'key':'value'})
* >>> gevent = requests.delete('http://httpbin.org/delete')
* >>> gevent = requests.head('http://httpbin.org/get')
* >>> gevent = requests.options('http://httpbin.org/get')

Requests

* 傳遞 URL 參數
  * >>> payload = {'key1': 'value1', 'key2': 'value2'}
  * >>> r = requests.get("http://httpbin.org/get", params=payload)

* 列印輸出該 URL，看到 URL 已被正確編碼：
  * >>> print(r.url)
    http://httpbin.org/get?key2=value2&key1=value1

BeautifulSoup

* Beautiful Soup 是一個可以從HTML或XML檔中提取資料的Python套件.可透過適當的轉換器來對文件做查找,修改文件的方式 。

* 簡單的說就是可以對抓取下來的網頁進行資料的擷取及處理。

* BeautifulSoup Ex...

Pandas

* Pandas 是 數據分析 函式庫，其資料格式(Data Frame)可以快速操作及分析資料，特色如下：

* 數據的讀取、轉換和處理上，都讓分析人員更容易處理，例如：從列欄試算表中找到想要的值。

* 其資料結構有以下二種
* Series、 DataFrame。
  * Series 處理時間序列相關的資料-感測器資料，主要為建立索引的一維陣列。

  * DataFrame 處理結構化的資料，有列索引與欄標籤的二維資料集-CSV檔，資料庫檔…等等。

Pandas

* 透過結構化物件所提供的方法，來快速地進行資料的前處理，如資料補值，空值去除或取代等...類似統計學裡的SPSS軟體的部份功能。

* 亦可從資料庫讀取資料進入 Dataframe，也可將處理完的資料存回資料庫。

* Pandas Ex...

Requests & BeautifulSoup & Pandas Ex

* Requests & BeautifulSoup & Pandas Ex
  * RBP-1.py
  * RBP-2.py
  * RBP-3.py

urllib

* Urllib是Python裡的標準函式庫，無需安裝，可直接可以用。提供了如下功能：
  * 網頁請求
  * 回應獲取
  * 代理和cookie設置
  * 異常處理
  * URL解析

urllib

* urllib.request for opening and reading URLs
* urllib.error containing the exceptions raised by urllib.request
* urllib.parse for parsing URLs
* urllib.robotparser for parsing robots.txt files

urllib

* urlopen 語法
  * # request:GET
  * import urllib.request
  * response = urllib.request.urlopen('http://www.baidu.com')
    print(response.read().decode('utf-8'))

* Request語法
* urlparse:拆分URL
* urlunparse:拼接URL，為urlparse的逆操作

* Urllib_EX…..

Urllib Ex

* Urllib Ex
  * UrllibEx1.py
  * UrllibEx2.py
  * UrllibEx3.py

Json & sqlalchemy

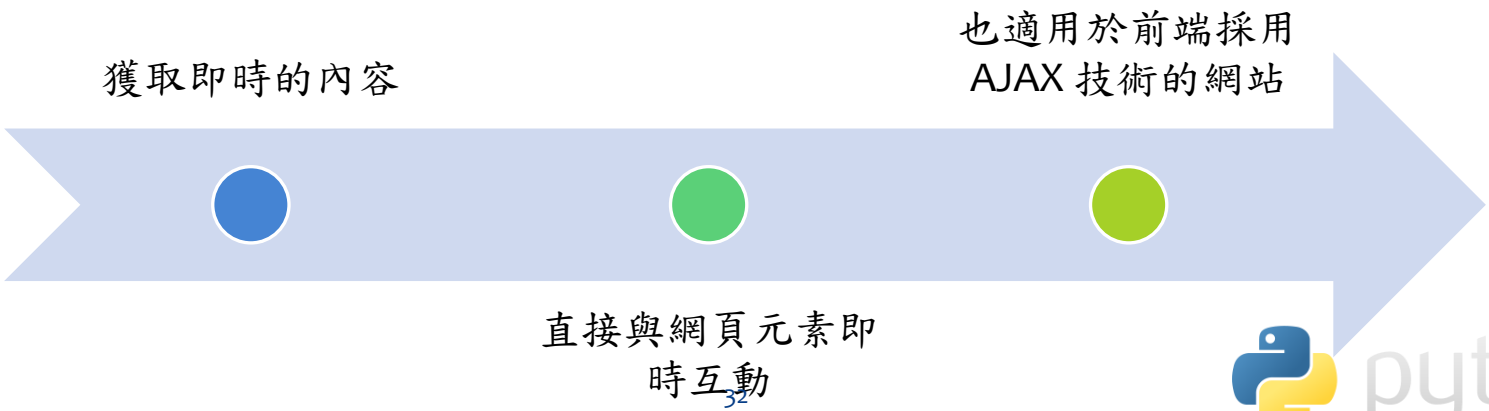
* **JSON**（**J**ava**S**cript **O**bject **N**otation，JavaScript物件表示法）是一種由道格拉斯·克羅克福特構想和設計、輕量級的資料交換語言，該語言以易於讓人閱讀的文字為基礎，用來傳輸由屬性值或者序列性的值組成的資料物件。
* **SQLAlchemy**提供了SQL工具包及物件關係對映（ORM）工具，使用MIT許可證發行。
* SQLAlchemy「採用簡單的Python語言，為高效和高效能的資料庫存取設計，實現了完整的企業級持久模型」。
* SQLAlchemy的理念是，SQL資料庫的量級和效能重要於物件集合；而物件集合的抽象又重要於表和行。
* 因此，SQLAlchmey採用了類似於Java里Hibernate的資料對映模型

Requests & Json & sqlalchemy Ex

* Requests & Json & sqlalchemy Ex
  * RJsqlEx-1.py
  * RJsqlEx-2.py

Selenium & Selenium Ex

* Selenium
  * 是為了可以讓瀏覽器自動化（Browser Automation）所設計的一套工具，讓程式可以直接驅動瀏覽器進行各種網站操作。
  * 運作在執行其「真實的瀏覽器」來進行網站操作的自動化，



Selenium & Selenium Ex

* Selenium Ex
  * SeleniumEx1.py
  * SeleniumEx2.py
  * SeleniumEx3.py

Matplotlib

* **matplotlib**是Python程式語言及其數值數學擴展包 NumPy的可視化操作界面。

* 它利用通用的圖形用戶界面工具包，如Tkinter, wxPython, Qt或GTK+，向應用程式嵌入式繪圖提供了應用程式接口（API）。此外，matplotlib還有一個基於圖像處理庫（如開放圖形庫OpenGL）的pylab接口，其設計與MATLAB非常類似。SciPy就是用matplotlib進行圖形繪製。

* Matplot現今來說歷史悠久也最多人使用，非常多的教學文章及範例，且畫圖功能多元。

* Matplotlib_Ex….

綜合應用

* 綜合應用
  * FinalApp1.py
  * FinalApp2.py
  * FinalApp3.py

參考資料

* 維基百科，自由的百科全書
* TQC+ Python3.x 網頁資料擷取與分析 特訓教材