104 網站 --- 爬蟲運作流程

目的

1. 取得「職缺、公司、工作內容、工作經歷、職缺連結、薪水、擅長工具」的表格(.csv)
   (爬下來的資料都以 DataFrame 形式去處理)
2. 取得擅長工具的統計圖表
3. 取得工作經歷的統計圖表

Github 連結

========== 流程說明 ==========

- 輸入想搜尋的關鍵字，format(關鍵字+頁數)到 URL

```
keyword = input('Search job from 104: ').replace(' ', '%20')
page = 1
url="https://www.104.com.tw/jobs/search/?ro=0&kwop=7&keyword={}&expansionType=area%2Cspec%2Ccom%2Cjob%2Cwf%2Cwktm&order=15&asc=0&page={}&mode=s&jobsource=2018indexpoc".format(keyword, page)
```

- 取得以下資料:

1. 職缺名稱 **Job**
2. 公司名稱 **Company**
3. 工作內容 **Content**
4. 工作經歷 **WorkExp**
5. 該職缺的頁面連結 **Url**
6. 薪水 **Salery**
7. **擅長工具** (本次結果以 Data Engineer 為範例)
   因為搜尋關鍵字是 input 內容，擅長工具會因職缺而不同
   所以只取出最先出現的 20 個

```
8.  # create final_skill DataFrame
    final_skill_columns = list()
    for skill_list in final_data['Skills']:
        for skill in skill_list:
            # get first 20 skill only
            if len(final_skill_columns) == 20:
                break
            elif skill in final_skill_columns:
                pass
            else:
                final_skill_columns.append(skill)
    final_skill = pd.DataFrame(columns=final_skill_columns)
```

- 為了跑圖表，將 worKExp 轉成 int，若空值則填入 0

```
# if workExp is empty
try:
    workExp = int(re.findall('\d', jobData['data']['condition']['workExp'])[0])
except IndexError:
    workExp = 0
```

- 將所有資料依據填入 df

```
tmpData.append(pd.Series([job, company, content, skill, salary, workExp, jobUrl],
index=final_data.columns))
```

- 每爬 1 頁 print 完成訊息；每爬 10 頁休息 20 秒

```
print("finish page {}".format(page))
# sleep 10 seconds for every 10 Pages
if (i % 10 == 0):
    time.sleep(20)
```

- 為了做成圖表把 skill list 做成[1, 1, 0, 0, 1, 0, 0.....]的格式

```
# one-hot-encoding for skill & plot
for skill_list in final_data['Skills']:
    skill_ohe = [1 if skill in skill_list else 0 for skill in final_skill_columns]
    tmpskill.append(pd.Series(skill_ohe, index=final_skill_columns))

final_skill = final_skill.append(tmpskill, ignore_index=True)
```

- 將結果以 input 的關鍵字作為檔名儲存

```
plt.savefig('./work104/skill_for_{}.jpg'.format(keyword).replace('%20', '_'))
plt.savefig('./work104/workExp_for_{}.jpg'.format(keyword).replace('%20', '_'))
final_data.to_csv('./work104/work104_for_{}.csv'.format(keyword).replace('%20', '_'),
encoding='utf-8-sig')
```

(成果截圖在下一頁)

	Job	Company	Content	Salary	WorkExp	Url	C#	Java	Python	MySQL	HTML	Verilog
0	KLA美商科磊 2021校園新鮮人招募專區	美商科磊股份	【KLA 校園新鮮人招募專區】	待遇面議	0	https://www.104.com.tw/jot	0	0	0	0	0	0
1	"歡迎軟體行銷優秀人才" Product Marketing Role	ThunderCore_閃	Roles & Responsibilities • Maintain, implement and	待遇面議	3	https://www.104.com.tw/jot	0	0	0	0	0	0
2	資訊-大數據工程師(海外地區)	臻鼎科技股份	1. 各廠區EDA(工程數據分析)專案導入：包含需求訪	月薪36,000元以上	0	https://www.104.com.tw/jot	1	1	1	1	1	0
3	CPU Design Verification Engineer	金麗科技股份	Job description • Work with IC architects and	待遇面議	3	https://www.104.com.tw/jot	0	0	0	0	0	1
4	研發類 - SSD韌體/軟體研發工程師/經理(MP400)	旺宏電子股份	1.NVMe SSD 開發經驗，設計專案已量產尤佳	待遇面議	3	https://www.104.com.tw/jot	0	0	0	0	0	0
5	Technical Support Engineer / 技術支援工程師	逸空間有限公	排除第一線技術問題以維護產品的正常運作，並且擔任	月薪32,000~45,000元	0	https://www.104.com.tw/jot	0	0	0	0	0	0
6	【產品研發部】-技術經理	網際威信股份	工作內容： 1. 參與產品的功能規劃, 需	待遇面議	5	https://www.104.com.tw/jot	0	1	0	1	1	0
7	.NET C# 開發人員	拓樸龍科技股	【工作內容】 1. 平台會員功能設計開發	待遇面議	2	https://www.104.com.tw/jot	1	0	0	0	0	0
8	設備裝機工程師(台南) Product Installation Enginee	美商科磊股份	1. Perform customer-site installation and acceptance	待遇面議	0	https://www.104.com.tw/jot	0	0	0	0	0	0
9	設備工程師 - 南科 (濕蝕刻 Clean、蝕刻 Etch、薄	美商_科林研發	Technical 1. Responsible for providing	待遇面議	0	https://www.104.com.tw/jot	0	0	0	0	0	0
10	大數據 工程師	動力安全資訊	透過Big Data相關的技術，結合雲端服務，分析網路、	月薪30,000~40,000元	0	https://www.104.com.tw/jot	0	0	0	0	0	0
11	[TCXY08]大數據工程師_Big Data Engineer	優聘資訊科技	【代企業徵才，錄取後為該企業正職員工】	月薪38,000~80,000元	1	https://www.104.com.tw/jot	0	1	1	0	0	0