以fomc發言預測利率走勢

|  |  |
| --- | --- |
| 網路爬蟲 | 邱子軒 |
| 資料庫建立 | 鄭元甦 |
| 文字轉向量word2vec | 余柏瑞 |
| 估算FOMC升降息機率 | 黃煒翔 |

一、網路爬蟲

## 網路爬蟲的代碼

from urllib.request import urlopen  
import re  
import requests  
from bs4 import BeautifulSoup  
import json  
html = urlopen(  
 "https://www.federalreserve.gov/newsevents/speech/2019-speeches.htm"  
).read().decode()  
res = re.findall(r'href="(.\*?)"', html)  
matchers =['newsevents/speech/']  
matching = [s for s in res if any(xs in s for xs in matchers)]  
K='https://www.federalreserve.gov'  
matching=[K+x for x in matching]  
aa=len(matching)  
for i in range(0,aa):  
 url = matching[i]  
 page = requests.get(url).text  
 soup = BeautifulSoup(page)  
 p\_tags = soup.find\_all('p')  
 p\_tags\_text = [tag.get\_text().strip() for tag in p\_tags]  
 sentence\_list = [sentence for sentence in p\_tags\_text if not '\n' in sentence]  
 sentence\_list = [sentence for sentence in sentence\_list if '.' in sentence]  
 article = 'NEW ARTICLE '+url+' '.join(sentence\_list)+' END ARTICLE '  
 my\_json\_string = json.dumps(article)  
 with open('2019', 'a+', encoding='utf-8') as f:  
 json.dump(article , f, ensure\_ascii=False, indent=4)

說明

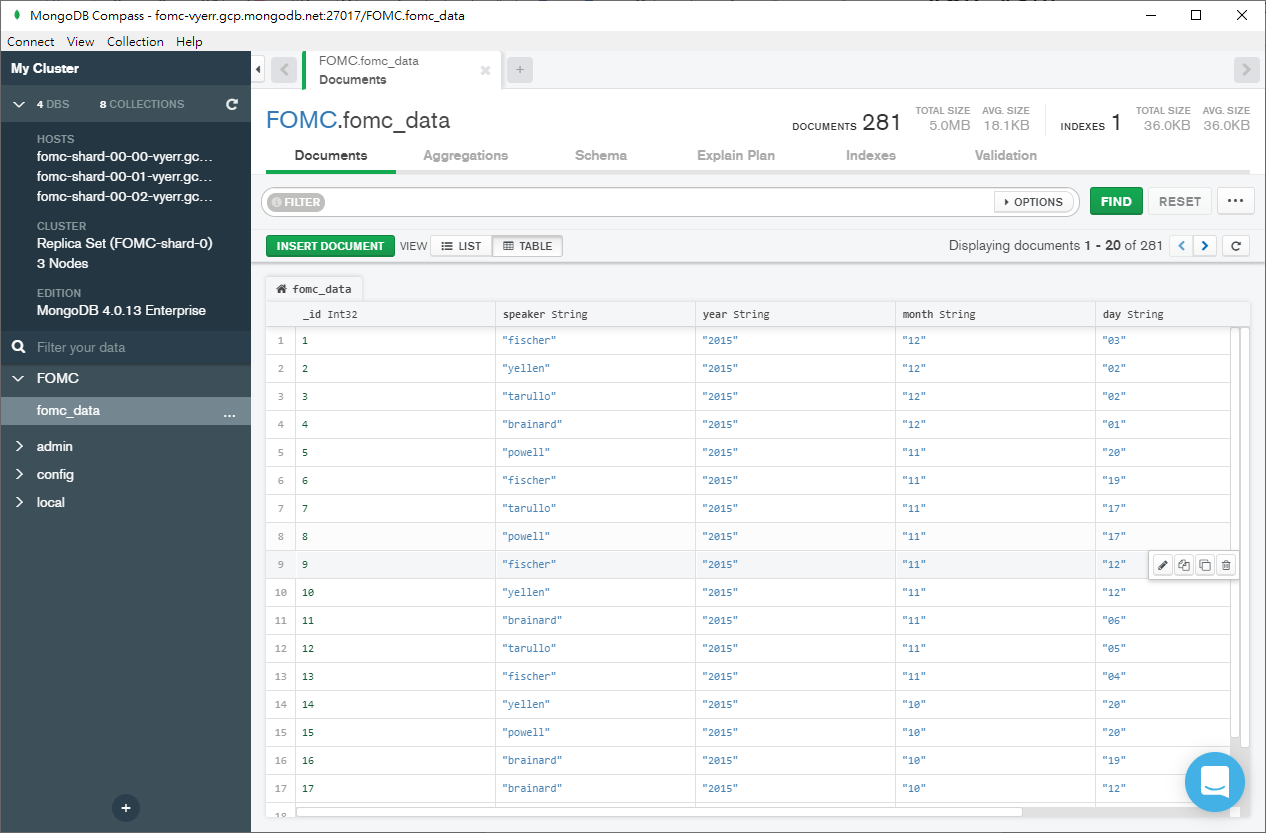
我們使用python BeautifulSoup套件，去進行資料整理。首先必須克服的問題是如何有效大量且短時間一次下載所有的文檔。這必須要去研究聯邦公開市場委員會網站的ＨＴＭＬ代碼。之後再把下載的檔案轉成json檔方便建立資料庫。

二、資料庫建立

本組使用非關聯式資料庫來儲存爬蟲結果。使用非關聯式資料庫的原因在於，非關聯式資料庫較關聯式資料庫具儲存彈性；且因爬蟲結果檔案較大，使用非關聯式資料庫在讀取上較具效率。

而因本組儲存內容為大量文字，故本組使用MongoDB作為資料庫。在雲端上建置資料庫後，便將爬蟲結果放置至雲端空間。建立資料庫的過程除了在MongoDB建立帳號並且設定存取權、安裝之外，在本地使用MongoDB Compass以便確認資料庫是否能夠正常連線及確保爬蟲的結果能夠順利地儲存在資料庫內。

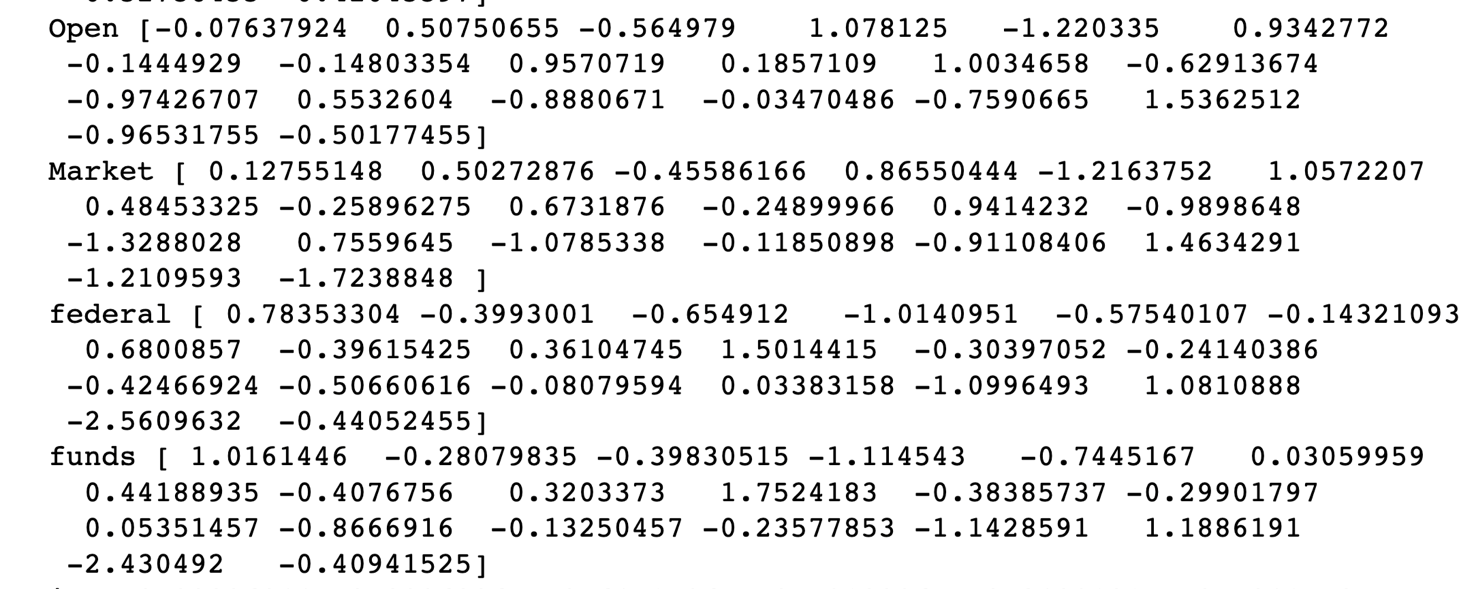
當資料庫建立後，即可開始以程式語言快速的將資料放入資料庫中。我們選擇Python作為讀取MongoDB之程式語言，並且使用其套件pymongo作為讀寫資料庫之手段。在爬蟲檔案中，依據一定的規則，找出了發言人、發言日期、以及發言內容，作為讀取之鍵值。

下圖為本組資料庫之型態，雖然MongoDB為非關聯式資料庫，但透過MongoDB Compass，仍可以以關聯式資料庫之型態來檢視資料庫。本組蒐集了2015年至2019年底Fed的發言，共281筆資料。

三、文字轉向量 Word2vec

本組採用gensim模組內的word2vec套件，將文章內獲取之關鍵字轉成以向量表示之。在進行word2vec之前，得先將文章內的每個單字分開，稱為「分詞」，分詞之套件為NLTK，在分完詞後須將多餘之詞濾掉如「a」、「the」、「http」等，對文章意義無影響之詞，濾完且分完詞後便是成為我們將要訓練成為詞向量的資料。

資料備妥後便使用word2vec進行訓練，將會得出一套模型，模型所產生的資料便是之後建構趨勢預測模型的資料來源。我們可以利用這模型顯示出各詞之向量值，也可以找出最具相關性的詞，還會給出一相關係數，十分方便。也因此套件之方便性，可以任意調整模型之維度，本組調整採用維度為20，方便爲之後趨勢預測模型更簡單建置。

下圖為跑完模型後，部分詞之向量值。

四、估算FOMC升降息機率

本組發現在FOMC利率會議聲明中，最重要的結論莫過於升息、降息、維持，分別為”raise”、”lower”、”maintain”。而我們想知道哪些詞彙會導向升息和降息，藉以衡量升降之間的趨勢強弱，估算出升降息機率。

本組將訓練完成的word2vec模型，利用Most\_similar函數，排序出最相關的單詞及各自的相關係數。我們藉此列出兩種清單，分別是以raise和lower為主的前200名單詞清單(後文簡稱raise清單、lower清單)，為了便於回測與排版，於EXCEL進行後續分析。

首先在INPUT工作表放入今年所有FOMC利率會議聲明內容(共8次)，利用各單詞出現頻率對相關係數加權，運算出該次會議的raise指數、lower指數。最後再用線性內插法(兩端為100%升息、100%降息)，得出估計之升息或降息機率。實際值本組採用今年各次聲明發布時，美國公債期貨市場上的殖利率所隱含之升降息機率(數據來源：Bloomberg、財經M平方)。今年的預測方向大致與市場預期相同，估計機率值和實際機率值之差距都維持在33%之內。下方表格為該模型輸出結果之彙整。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 12/11 | 10/30 | 9/18 | 7/31 | 6/19 | 5/1 | 3/20 | 1/30 |
| Calculated Possibility (估算值，負數表示降息機率) | **-22.82%** | **-22.63%** | **-21.83%** | **-24.38%** | **-28.11%** | **-31.12%** | **-34.12%** | **-28.45%** |
| Market Implied Possibility  (實際值，負數表示降息機率) | **-8.90%** | **-22.90%** | **-43.80%** | **-56.50%** | **-20.80%** | **-10.00%** | **-2.00%** | **1.30%** |
| 估計誤差 | -13.92% | 0.27% | 21.97% | 32.12% | -7.31% | -21.12% | -32.12% | -29.75% |