中天直播留言分析

組名 :

SedNeoCat

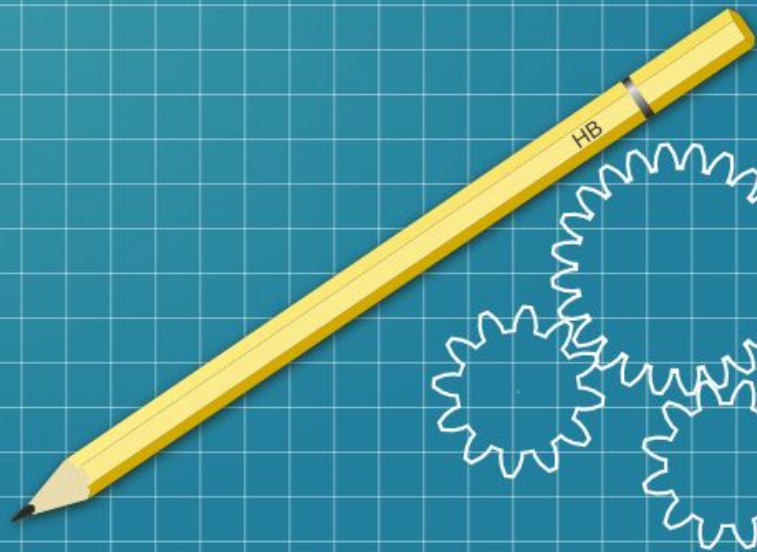
組員 :

蘇子權/40947023S

余原齊/40947027S

吳文元/40947030S

洪盛益/40947047S



What is the problem?

- 這個題目是什麼，它能解決什麼問題？
  - 協助社會學者藉由資料分析大眾對傳媒轉新媒的觀感。
  - 觀察中天新聞受眾的轉變。

Why this is important?

- 這個題目的緣起動機是什麼？你注意到什麼需求、現象、問題？
  - 由於近期中天新聞被NCC關台，轉戰YouTube頻道，因此我們想了解這轉變帶來的影響。
- 為什麼這個問題值得被解決？
  - 因為中天新聞是台灣第一個傳統媒體完全轉型為新媒體的案例。

What to do and how you do it?

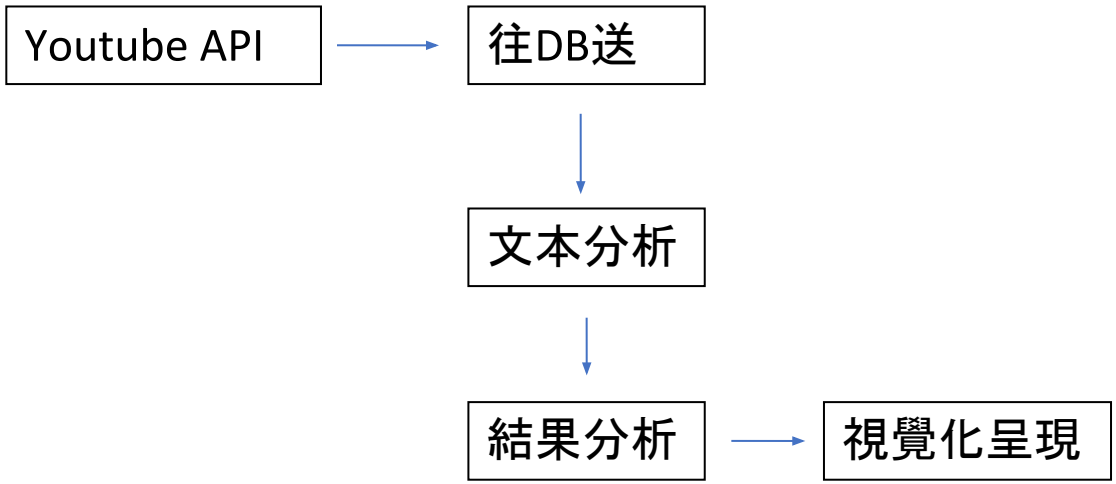
- 你需要做什麼來解決這個問題？你要做的這些事情裡的先後順序是什麼？
  - 蒐集YouTube聊天室內容
  - 分析文本、結果分析
  - 呈現圖形化結果
- 你會需要哪些工具？
  - python bot
  - ~~Jieba~~ -> Articut API
  - Matplotlib

Where do the data come from?

- 你的資料從哪裡來？
  - 中天新聞YouTube頻道聊天室
- 怎麼收集？
  - Python Bot
- 怎麼知道量夠不夠？
  - 實做時在資料中可以發現明顯差異時。

Work flow

- 程式流程方塊圖



成果展示

Bot

- fetch data from Youtube API
- **füçk y0u Quota**
  - Data cache reduce quota waste
  - update data when request failed due to outdated data
- YT API return duplicated message
  - DB composite primary key!
  - Do nothing when dulicated
- Crontab trigger data fetch

前處理

- 從Database取資料
- 去除emoji，並將有意義的內容轉回文字
- 全形轉半形
- 去除 +1類型文字

Process

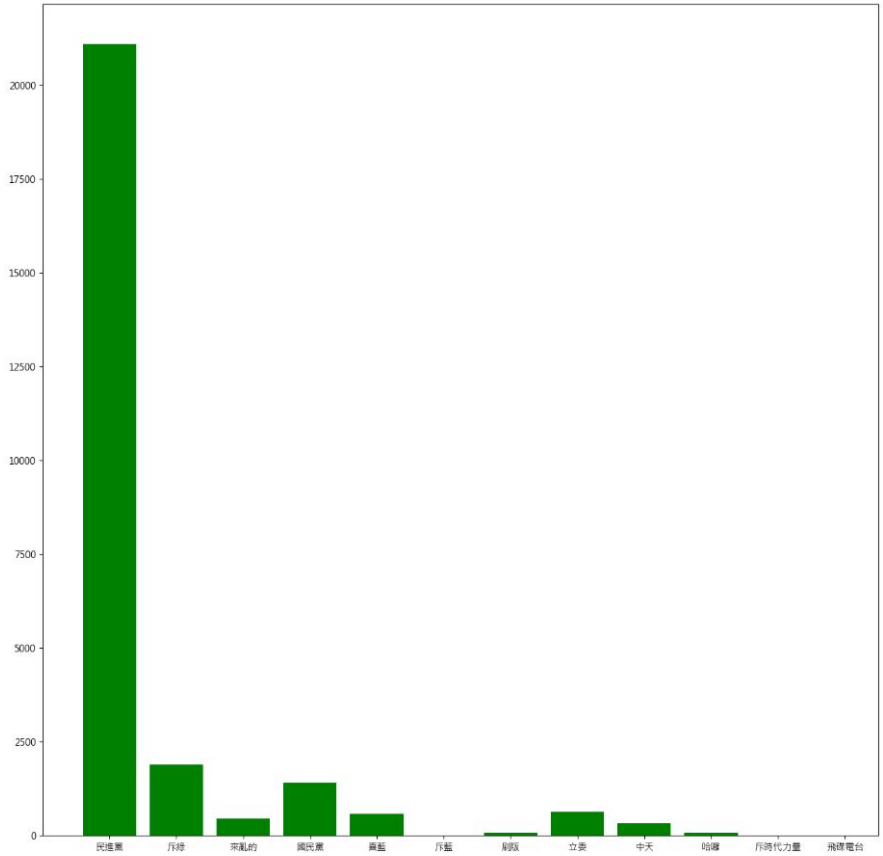
- ctp: 用拼音解決繁、簡體
- nltkSentiment: (情感分析)
- nameJob: name funtion的多執行緒切UserDefinedDictFile的name.json
- name: 數名字出現的次數
- attribute: 找出自定義相關詞句並歸類
- donate: 算出觀眾捐了多少錢給中天，有分辨幣值。

工人智慧

- 當個工人(?
  - ~~NLP的最佳實踐~~
- 把關鍵詞用UserDefinedDictFile包成三個json
- attribute.json
  - 將怪怪的用語或是故意錯別字歸類
- emoji.json
  - 表情歸類
- name.json
  - 將綽號、代名詞歸類

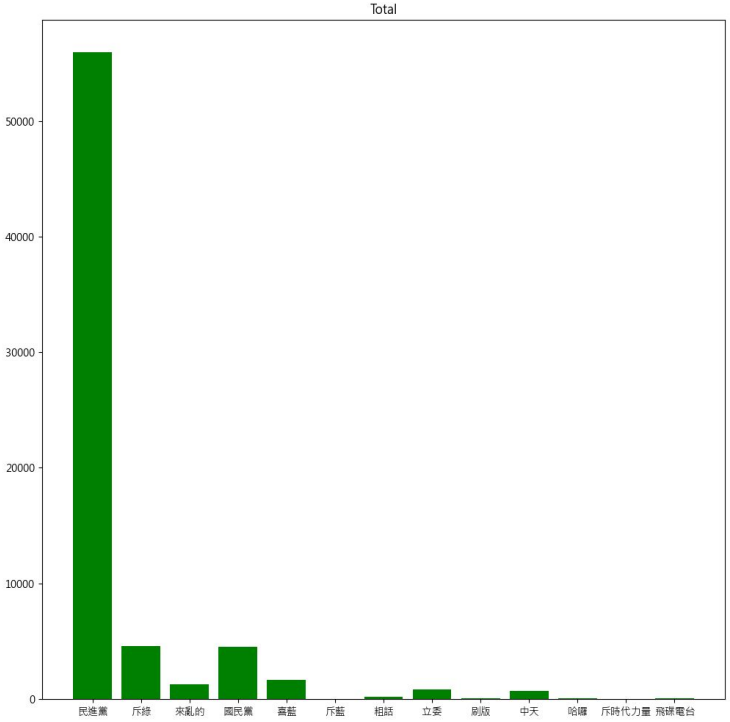
目的: 能切出自定義的詞

視覺化圖表

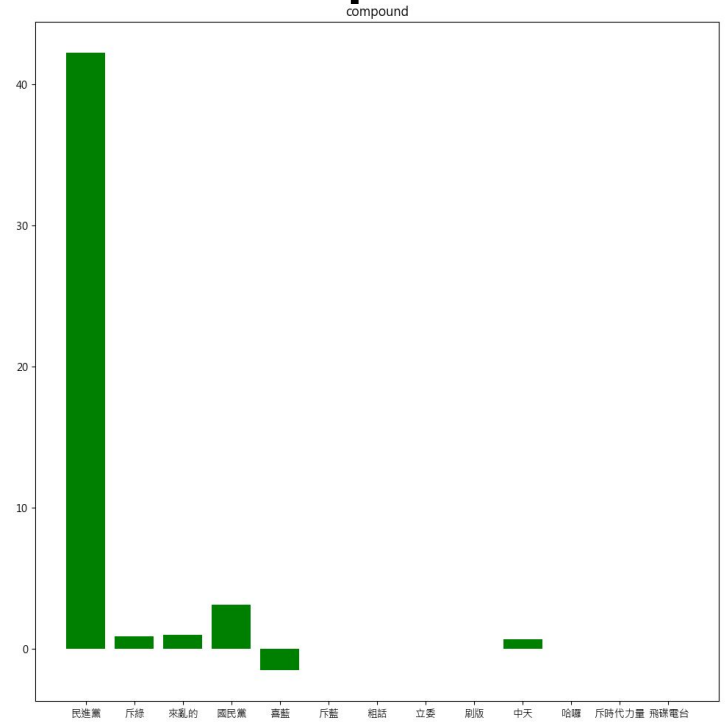

20000
17500
15000
12500
10000
7500
5000
2500
0
民進黨
庶時
來亂的
國民黨
真藍
庶藍
刷版
立委
中天
啃嚼
庶時代力量
飛碟電台

視覺化圖表

Total

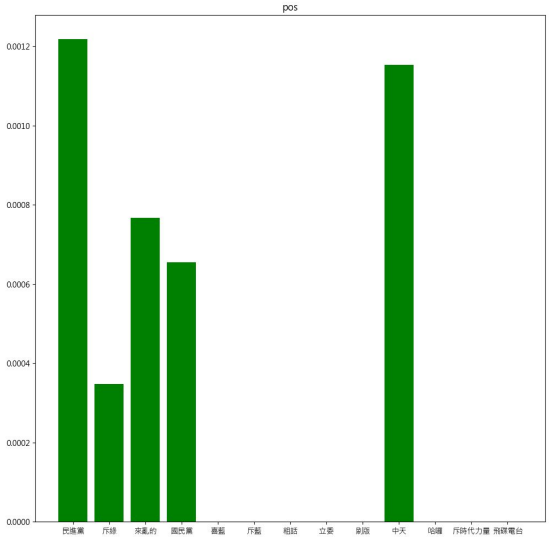


compound

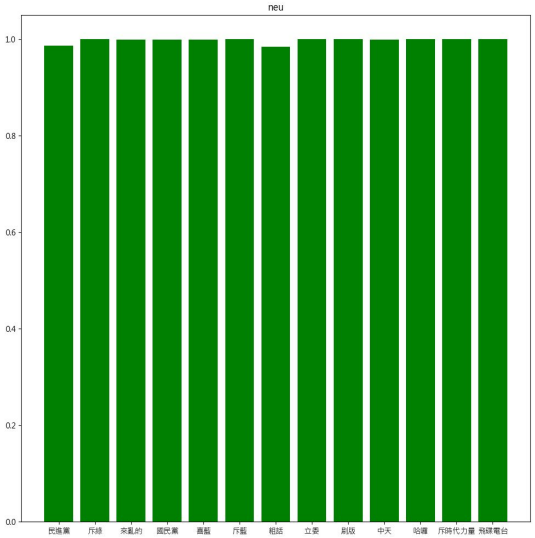


視覺化圖表

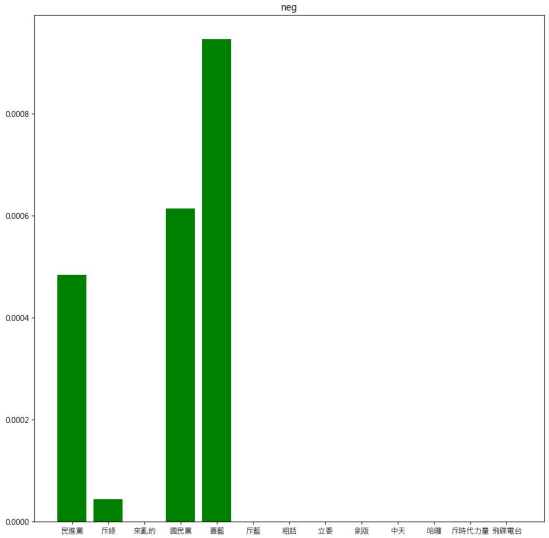
pos



neu



neg



一些娛樂

- entertainment.json
  - articut 不接受斷詞的留言。

過程中遇到的瓶頸

- 工人智慧會睛痛。
- Articut一分鐘80次的限制，需要耗費較多的時間等待。
- 用多執行緒直接塞Articut會炸掉->threading pool
- 畫圖時會中文字有字型缺陷，在linux-server上生成圖片會有問題
- database有可能被SQL Injection->跳脫字串
- YT API fetch 到底後，不會終止 -> 我就定時抓，定時停 www

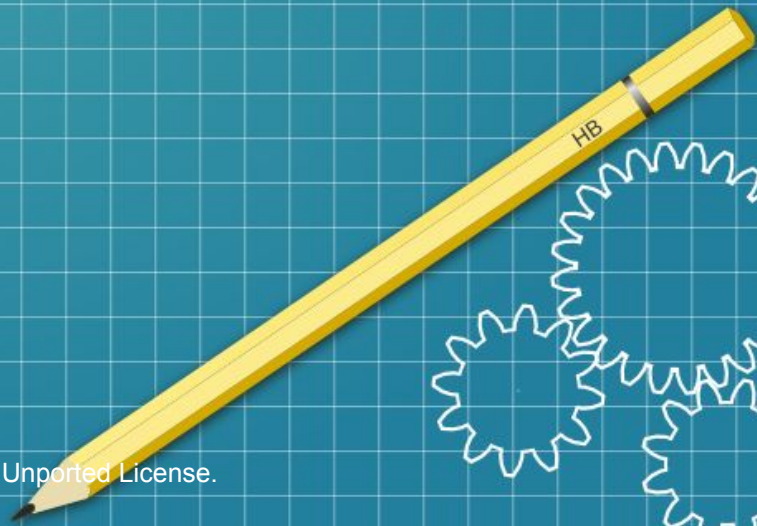
Who did what in your team

- 請介紹你的組員工作分配:
  - 蘇子權: 工人智慧
  - 洪盛益: 支那間諜 拼音大師 aka 情感分析大師
  - 余原齊: 超自然前處理與多執行續優化大師
  - 吳文元: YT API 與資料快取大師
- Github repo: https://github.com/jw910731/NTNU_TextProcessing_Final.git

From ArticutAPI Github Repo

```
{"雷姆":["小老婆"],
 "艾蜜莉亞":["大老婆"],
 "初音未來": ["初音", "只是個軟體"],
```

Q & A



Q & A

- 請簡單分析出來可能的結果為何?

  1. 政治傾向
  2. 支持度
  3. 來鬧的

  => 最後由圖形化結果呈現

Q & A

- 「協助社會學者」是指誰？或是哪一個領域？「從文本推測性別」，怎麼做？

  理論上研究新聞媒體轉型學者皆為團體，而且研究這類型結果還是會回到自己身上，因為我們不能控制哪些學者會採納我們的研究結果。因此我們的研究結果是公開在網路上讓研究學者、媒體「主動」看到我們的研究結果，也就是我們扮演自媒體

Q & A

- 政黨傾向怎麼分 畢竟現在政黨紛雜?

  我們只呈現支持與不支持，如果留言討厭綠，那只將其歸類在不支持綠，則不會歸類在支不支持綠以外的黨。

- 水軍如何處理?

  這點我們無法判斷，我們也無從分析誰是被買通的。我們也只能納入統計。

- 能否分辨機器人？

  通常機器人為固定時間留言，從留言時間判斷。

Q & A

- 想請問你新聞自由怎麼定義呢?

  言論自由，可報導想報導的事物，假新聞除外。

- 如果是反諷語氣模仿敵隊陣營用詞的怎麼處理

  對於段詞而言很難做到這點，而且反諷語氣有部分雖然是可以判斷出，但有些是「個人主觀的解釋以及語調」，因此字面上是無法判斷他的語氣。