

PTT Crawling and Text Clustering

陳俊廷

R3323057@ntu.edu.tw

0. Abstract

- Crawling and extracting contents
- Feature engineering: sentence segmentation, tf-idf
- Word cloud
- Text clustering

1. Crawling and Extract content

- Crawl 3946 postings on board *EAseries* (2016/8/5-2017/8/9).
- Retrieve the title, main content, comments of each posting and combine them.

2. Feature Engineering

- sentence segmentation: jieba
- tf-idf: top-10000 features

Note: The stop word list would be adjusted according to the result.

3. Word Cloud

- Drop all the segmented words and their tf-idf to Tableau.

4. Text Clustering

- Cluster all postings by k-means.
- Sum all the tf-idf of each category classified and get a tf-idf vector representing each category.
- Choose words corresponding to the top 10 tf-idf in each category as keywords.
- If any word exist in all category, add it into the stop word list and re-train.

關於 爆炸 離開 魔法 and 歡樂 整理 模式 廣告 趕快 整部 認真 態度 難道 權力 劇中 call ray 翻譯 戲份 聲音 還會 難題 關鍵 講話 歡迎 變得 續約 ww kim big s1 12 ncis 路上 感受 倦眼 等等 黑人 評價 經典 意思 腳色 暗示 當年 經過 電腦 感動 慢慢 21 溫哥 新劇 愛情 媽媽 遊戲 疑團 網路 機台 換笑 標準 厲害 精神 適合 壓力 轉折 還差 警察 犧牲 議題 luke 類型 醫療 這劇 終於 第五 情報 接著 殺手 理解 清楚 提供 晚上 第二 集幾 搞寫 麻煩 博士 this 稍微 第六 超愛 超過 媒體 bug 剩下 無聊 last 幾個 場景 幾次 等到 順便 想想 想起 概念 偷喝 結束 幫忙 有似 剛好 首先 剛刷 恐怖 限制 相當 有些 相關 拯救 看法 很大 神秘 真像 律師 演出 英國 首播 直中 粉絲 深刻 控制 理論 組織 眼睛 推推 部份 偽造 這點 這件 設計 tv 絕命 進去 連結 信任 放在 姊妹 彼此 表情 能夠 真心 要求 有起 相關 拯救 看法 很大 神秘 真像 律師 演出 英國 首播 直中 粉絲 深刻 控制 理論 組織 眼睛 推推 部份 偽造 這點 這件 設計 tv 絕命 進去 連結 別人 作為 沒人 身上 完美 爸爸 弟第 13 兩位 依舊 武器 兒子 身體 明顯 或者 事件 很少 幽默 思考 刻意 抱歉 其中 況且 屍體 則是 空間 16 多年 使用 系列 每集 忘記 更是 28 老實 改變 即將 有夠 字多 多久 同一 再來 好好 名字 兇手 完成 有所 宇宙 26 改編 君臨 行為 攻擊 而首 利用 老師 完整 作品 收復 回去 男女 23 老公 安排 奇怪 性格 作者 多次 介紹 15 夫妻 出場 今年 出戲 反串 片頭 文化 fox 喜劇 重要 人員 朋友 5 接受 結束 52 兩人 01 frank barry 播出 甚麼 並沒有 be 是否 演出 結局 abc 畫面 記得 謝謝 內容 ford 機器 suits 節目 感謝 主要 arrow 表現 成功 in 能力 小號 cw 總人數 還沒 當不 懂 無法 gatham 07 黑人 女 超能力 03 收視率 兩個 搞不好 台灣 一口 找到 印象 各種 不得不 survivor oliver 差不多 chicao go 久 michael girl 第四季 什麼 出現 第一次 arnold xddd 25 這季 無法 waym tbbt sherlock 表示 breaking amazon 女兒 xdd 其他人 外星 人 john 男主角 一開始 三個 新聞 一段 看下去 sense8 女主角 j netflix 第一集 不同 20 編劇 一季 期待 電影 以前 主角 xddd cbs westworld 美國 一切 中文 字幕 世界 0 推劇 戲 吸 血 鬼 男 主 角 一 開 始 三 個 新 聞 一 段 看 下 去 sense8 女 主 角 j netflix 第二集 大部分 narcosis 火之歌 better bernard lee things dolores 看不出 if american 莫名其妙 family 世界 0 推劇 戲 吸 血 鬼 男 主 角 一 開 始 三 個 新 聞 一 段 看 下 去 sense8 女 主 角 j netflix 女奴 生活 工作 jimmy james 奇物語 10 house 哈哈 xddd 解釋 是不是 設定 還不錯 william 東西 本劇 這個 0 5 好 笑 18 繼續 第二集 life 演員 第一 season 01 有趣 最近 觀眾 got hbo push flash 意外 人物 回到 算是 十分 00 每個 可惜 dr david 影響 啊 啊 啊 story 超級 畢竟 s3 特別 那 精彩 你 哈哈 1 night game fargo with 一邊 一位 力量 也許 不 不夠 一路 女 主 女 孩 27 女性 teddy jane 公布 女 王 小 鎮 不 知 中 國 不 行 代 表 公 司 公 主 左 右 只 好 平 台 心 理 方 法 世 紀 本 身 回 歸 02 如此 好友 再度 取 回 安 全 呈 現 快 轉 更 好 有 甯 老 婆 成 績 自 我 努 力 fb 可 憐 共 同 可 怕 本 人 目標 手段 再次 充 滿 名 稱 有 關 在 意 考 慮 任 務 成 員 存 在 4 死 亡 09 回 家 正 在 角 度 那 裡 受 到 社 會 政 治 為 何 相 信 19 肯 定 便 當 有 得 重 捕 科 幻 背 叛 推 文 討 論 記 者 效 果 告 訴 那 位 季 末 妻 子 法 律 刺 激 劇 份 快 客 車 的 關 宣 布 01 的 劇 易 昨 天 毒 師 值 得 原 創 負 責 故 意 面 對 下 情 緒 家 人 從 頭 到 尾 推 出 這 對 通 常 猜 測 最 高 最 細 節 最 佳 程 度 最 終 最 終 車 陣 痛 苦 登 場 傷 害 感 到 超 女 當 作 開 槍 開 捕 報 導 最 好 這 也 最 大 現 場 第 7 普 通 速 度 開 心 描 述 11 結 婚 傳 奇 無 論 感 情 資 料 集 到 abc 節 奏 開 頭 準 備 照 片 監 聽 實 際 想 像 認 同 想 法 線 索 過 程 電 視 對 象 錄 音 路 線 瘋 狂 說 過 戲 嗎 總 統 節 演 網 站 數 字 對 方 網 球 實 況 過 去 mib 經 理 的 電 話 緊 張 說 明 確 定 戰 爭 錯 誤 24 街 突 預 定 選 擇 隨 便 整 手 操 攔 與 趣 整 體 概 概 疑 重 調查 擁有 擔心 獲得 醫生 覺醒 邏輯 new jack to 歷史 游 樂 船 女 橋 段 隨 著 還 滿 爆 點 魔 王 主 導 演 類 似 tom s01 幫助 確 定 戰 爭 錯 誤 24 街 突 預 定 選 擇 隨 便 整 手 操 攔 與 趣 整 體 概 概 疑 重 調查 擁有 擔心 獲得 醫生 覺醒 邏輯 new jack to 魔 王 主 導 演 類 似 tom s01 幫助 確 定 戰 爭 錯 誤 24 街 突 預 定 選 擇 隨 便 整 手 操 攔 與 趣 整 體 概 概 疑 重 調查 擁有 擔心 獲得 醫生 覺醒 邏輯 new jack to

1: 一般心得	35.70%	4: Marvel	6.77%	7: Westworld	5.37%
【出現，兩人，'主角'，'決定'，'無法'，'jimmy'，'編劇'，'找到'，'生活'，'女兒'，' Barry'，'flash'，'閃電俠'，'綠箭'，'超女'，'英雄'，'dd'，'反派'，'arrow'，'ford'，'host'，'dolores'，'arnold'，'機器'，'bernard'，'westworld'，'maeve'，'mib'，'推薦'】動作影集THE SHOOTER狙擊生死線 S1微雷		【心得】 DC英雄四劇運動		【心得】 Westworld S01E07 (雷)	
【心得】 SKAM 第四季完結的小心得(雷)		【心得】 Supergirl S02E14 Homecoming (雷)		[LIVE] Westworld S01E09	
【心得】 Gotham S3E10		【心得】 The Flash S03E11 (雷)		Re: [心得] Westworld S1E05 (雷)	
【心得】 Shameless S07E07		【心得】 Iron Fist S01 (雷)		Re: [心得] Westworld S01E09 (雷)	
【心得】 Broadchurch S03E05 (雷)		【情報】 NETFLIX 漫威盧克凱奇 最新預告		【討論】 WestWorld S1 一點疑惑 (雷)	
【心得】 Big Little Lies S01E02 (雷)		【閒聊】 夜渡俠為何評價比閃電俠高?		【心得】 WestWorld S01E10 - MIB (雷)	
Re: [心得] Homeland S06E12 本季終 (雷)		【心得】 Legends of Tomorrow S02E12 (雷)		【請益】 幾個westworld問題	
【心得】 OITNB S1-S5 (雷)		Fw: [閒聊] 影集-Iron Fist應該要如何籌備		【閒聊】 Westworld & 攻殼	
【閒聊】 This is us S01E11		【心得】 The Flash S03E15 (雷)		【心得】 Westworld S01E08 (雷)	
【心得】 13 Reasons Why S1 (雷)		【心得】 Supergirl S02E16 (雷) 繁星		【請益】 Westworld S1E5問題(雷)	
2: 問題、閒聊	27.62%	5: GOT	5.98%	8: 收視率	2.03%
【Netflix，'期待'，'最近'，'主角'，'一季'，'電影'，'謝謝'，'好笑'，'推薦'，'台灣'，'小指'，'龍女'，'小惡'，'詹姆'，'北境'，'got'，'君臨'，'布蘭'，'龍媽'，'cbs'，'abc'，'nbc'，'fox'，'00'，'今晚'，'收視率'，'成績'，'節目'，'cw'，'Netflix'，'續訂'第四季		【請益】 權力遊戲第一季疑問		【新聞】 初步收視率 10/14(五)及續約等新聞	
【討論】Suits Season6 Finale		Re: [閒聊] GameOf Thrones S702 戰略地圖分析 (雷)		【新聞】 初步收視率 10/3(一)	
【討論】 Westworld 甚麼是 The Maze		Re: [請益] 權力遊戲-泰溫 (有關龍王新相)		【新聞】 收視率報告 3/1(三)及AMAZON預定影集新聞	
Re: [心得] Black sail S4E3		【閒聊】 GOT 疑問		【新聞】 初步收視率 11/28(一)	
【請益】 追劇服務選擇		【心得】 GOT S07E03 防雷 運輸問題		【新聞】 初步收視率 11/21(一)	
Fw: [情報] Constantine 新動畫系列		Re: [心得] 冰與火之歌 S03-09 我下巴掉了!!!!		【新聞】 初步收視率 10/2(日)	
【心得】 The Magicians S02E05		【閒聊】 今年出現GOT演員的電影~		【新聞】 初步收視率 11/22(二)	
【心得】 超虐的Happy Valley S01E03-4 (有雷)		【閒聊】 傳聞劇透冰與火之歌S07E04(可能巨雷)		【新聞】 5/9(二)初步收視率	
【討論】 SKAM S3中的聖經隱喻，神奇數字21:21		Re: [心得] Game of Thrones S07E02 (雷)		【新聞】4/13(四)初步收視率 - ABC夏季檔公布及取消	
【閒聊】 Suits S05E01-04		Re: [閒聊] GOT S7E02		【新聞】初步收視率3/6(一) CBS2017-18只預定4新劇	
3: 新聞	7.30%	6: 懸疑、驚悚: Better Call Saul, Breaking Bad	5.83%	9: 其他	2.03%
【'新聞'，'喜劇'，'of'，'預定'，'本劇'，'播出'，'hbo'，'netflix'，'宣布'，'電影'，'jimmy'，'saul'，'bb'，'老白'，'chuck'，'breaking'，'bad'，'better'，'call'，'00'，'21'，'23'，'11'，'22'，'20'，'10'，'06'，'16'，'17'，'那些'，'年'，'我'，'看過'，'的'，'笑劇'，'國家'，'地理'，'世紀'，'天才'：'愛因斯坦'，'07'，'那些'，'年'，'我'，'看過'，'的'，'笑劇'，'超級'，'英雄'，'SPN'，'開分店'，'Fw: [情報] 'Supergirl' (女超人) 第二季 on Warner TV。'討論' Bates Motel S05E03 (雷) (問卷) (已截止) 追劇行為研究 (填問卷抽Line代幣) 'Fw: [情報] 'The Flash' (閃電俠) on Warner TV。'討論' Bates Motel S05E06 (雷) '閒聊' 為什麼新版馬蓋先不找Jared演? [LIVE] AXN 07-08(六) 1900 新世紀潘羅斯第_4 Fw: [情報] 'The Flash' (閃電俠) on Warner TV。 (情報) 'DC's Legends of Tomorrow' (明日傳奇) on Warner TV。】		【心得】 Better Call Saul S03E05		【新聞】 國家地理 世紀天才：愛因斯坦 07	
Re: [心得] Powerless S01E01 (雷)		【請益】 絕命律師裡有絕命律師的人物嗎?		Fw: [情報] 'Supergirl' (女超人) 第二季 on Warner TV。	
【新聞】 Luke Evans及Daniel Bruhl主演TNT新劇		【討論】 Breaking Bad 老白的媽 (有雷)		【討論】 Bates Motel S05E03 (雷)	
【新聞】 HBO歐連軍超英影集? SPN將開分店?		【閒聊】 看Breaking Bad後藍壓印感		【問卷】 (已截止) 追劇行為研究 (填問卷抽Line代幣)	
【新聞】 奪最佳影集 '冰與火' 艾美獎大贏家		Re: [心得] Better Call Saul 第三季 S03E05		Fw: [情報] 'The Flash' (閃電俠) on Warner TV。	
【請益】 金融相關影集推薦		Re: [心得] Better Call Saul 第三季 S03E05		【討論】 Bates Motel S05E06 (雷)	
【新聞】 不只陰屍路! 2016年歐美口碑新劇大盤點		【請益】 絕命毒師某橋段		【閒聊】 為什麼新版馬蓋先不找Jared演?	
【請益】 新手求劇		【請益】 better call saul 哪集數有退休生活?		[LIVE] AXN 07-08(六) 1900 新世紀潘羅斯第_4	
【新聞】 CBS季中檔第一刀確定，CW新劇主角公布		【請益】 請幫忙推薦整季連看的影集		Fw: [情報] 'The Flash' (閃電俠) on Warner TV。	
【請益】 能很快吸引人的影集		【閒聊】 大家最想吃的'劇中美食'有哪些?		(情報) 'DC's Legends of Tomorrow' (明日傳奇) on Warner TV。	