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.

5. Results

Figure 1: word cloud by tf-idf

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	[請益] DC英雄四劇連動	[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] 夜魔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', '我	['cbs', 'abc', 'nbc', 'fox', '00', '今晚', '收視率', '成績', '節目', 'cw']
[新聞] iZombie 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] GameofThrones 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（有關龍王獅相）	[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 Soul, 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']
[心得] 那些年我看過的英劇	[心得] Better Call Saul S03E05	[LIVE] 國家地理 世紀天才：愛因斯坦 07
Re: [心得] Powerless S01E01 (雷)	[請益] 絕命律師裡有絕命毒師的人物嗎？	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欲進軍超英影集? SPN將開分店?	[閒聊] 看Breaking Bad後壓抑感	[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) 19:00 新世紀福爾摩斯_第4
[新聞] CBS季中檔第一刀確定，CW新劇主角公佈	[請益] 請幫忙推薦整季連貫的影集	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。

Figure 2 : clustering results