

R 語言與資料科學導論

Final Project

Group 10

知乎

B02611013 生機四 黃薇甄 / B02704033 國企四 張景淵 / B02705002 資管四 周詠瑤
B02705006 資管四 高偉立 / B02705027 資管四 陳信豪 / B02705028 資管四 冷俊瑩

一、介紹 (Introduction)

知乎是大陸的一個社會化問答網站，產品形態模仿了美國類似網站 Quora，也有點類似於台灣的 Yahoo 奇摩知識+，目前的註冊使用者已達 5000 萬，並擁有相當豐富的內容可以探勘。知乎網站架構下，有分為許多種主題，讓人們去發表文章、提問以及回答問題，而使用者們可以給這些內容物予以贊同、收藏和追蹤，全站目前累計產生了 1000 萬個問題，3400 萬個回答及 3500 萬贊同。

我們此次 Final Project 想要建立機器學習模型幫知乎回答給分數。我們將選擇一些特定主題，搜集其下的問題文本、回答文本、評論數、收藏數 等，並以贊同數當作我們的評分標籤。我們希望能在不依靠贊同數的情況下，透過這個模型為知乎的回答提供一個客觀的參考依據，讓發問者可以迅速知道哪些是好的回答以及讓回答者知道自己的回答水平如何。而我們的研究結果呈現在如下的網址，本報告將以解釋該頁面之功能為主。

知乎網站: <https://www.zhihu.com/>

Final Project 網站: https://omegappandda.shinyapps.io/shiny_dsr_zhihu/

二、資料搜集 (Data Collecting)

我們使用網路上的非官方 Python API，自己撰寫爬蟲程式來取得知乎平台上的資料，此為該 API 的來源：<https://github.com/7sDream/zhihu-oauth>。爬取的資料欄位如下表。

資料	說明
question_title	問題標題
question_detail	問題內容
question_time	發表問題時間
question_follower_num	問題追蹤人數
ans	回答內容
ans_time	發表回答時間
ans_upvote_num	回答贊同數
ans_collect_num	回答蒐藏數
ans_comment_num	回答評論數
author_follower_num	回答者被追蹤的人數
author_followee_num	回答者追蹤的人數
author_upvote_num	回答者獲得總讚數
author_thank_num	回答者獲得總感謝數
author_answer_num	回答者回答數
author_question_num	回答者發問數
author_post_num	回答者文章發表數
author_name	回答者名稱

我們最後蒐集到了 21 份資料集 (即 21 種不同話題的資料)，總計 1,300,000 筆資料。我們將運用以上的資料來進行我們後續的資料分析。

三、數據資料檢視 (Numeric Data Observing)

在左方的側欄中選擇想要檢視的資料集來檢視，若想要同時檢視多筆資料進行比較可以直接選取多筆資料，點選更新頁面後即可檢視以下初步分析資料：

1. 各資料集的資料筆數
2. 回答贊同數與時間的交互關係
3. 回答贊同數與回答者總讚數的交互關係
4. 回答贊同數與回答者被追蹤的人數的交互關係
5. 回答贊同數與回答者追蹤的人數的交互關係
6. 回答贊同數與回答者獲得總感謝數的交互關係
7. 回答贊同數與回答者發問數的交互關係
8. 回答贊同數與回答者回答數的交互關係

9. 回答贊同數與回答者文章發表數的交互關係
10. 回答贊同數與回答蒐藏數的交互關係
11. 回答贊同數與回答評論數的交互關係

四、文本前處理 (Text Preprocessing)

這個部分是針對剛利用爬蟲程式抓下來的資料進行初步整理及過濾，主要包括以下三個步驟：

1. 過濾網頁元素標籤和過多的冗贅無意義字元。
2. 斷詞：使用 JiebaR 來進行斷詞。
3. 過濾 stop words：取出所有文本中 df 值排名在 2.5% 以上的字詞進行過濾，因為這些字詞通常不具重要意義，例如：的、了、是、我、都.....。另外我們也使用了 tmcn 套件裡的 stopwordsCN() 作為我們停止詞的補充。

為了讓大家能更明白我們前處理的效果，在我們的網頁中有展示前處理的區塊，在 `text area` 中輸入利用爬蟲程式抓下來的內容，點擊送出文本後將立即開始進行上述的處理，並且可以看到我們文本清理和斷詞的結果，以下是一個簡單的範例。

文前本處理

請輸入要清理的文本

```
<div><div></div></div> ②參加附近行業的沙龍和論壇培訓等。 跟他人接觸舉別人
來提升自己。 <div><div></div></div> ③如果你有團隊那就團隊之相互討論開
會什麼的（我就有個舍伙人，是我高中好基友，連袂帶囑好夕給拉過來了）
相互扶持相互努力，這樣快樂X2 痛苦+2 真的非常非常感謝有朋友的一路相
伴！ <div><div></div></div> <br><br><br><br>還貨說完了 該說重頭戲了： <br><br><br><br>乾貨
</div></div><br><br>技巧和方 思路（這玩意我自己都沒懂） <br><br>這是一個
系統的科學，從頭開始。 * * <br> 打字太多了 好累啊 休息休息 看看有沒有
人愿意看吧 ~~~~~ <br><br>
+++++
+++++ <br>貌似被群嘲了 ~~~~ 來吧 讓潔白的兩點來的更猛烈點
<br><br>
```

送出文本

clean_text()

過濾

網頁標籤、url、

過多的連續標點符號與空白

（個人調節）②參加附近行業的沙龍和論壇培訓等。 跟他人接觸讓別人來提升自己。 （圈子調節）③如果你有團隊那就團隊之相互討論開會什麼的（我就有個舍伙人，是我高中好基友，連袂帶囑好夕給拉過來了） 相互扶持相互努力，這樣快樂X2 痛苦+2 真的非常非常感謝有朋友的一路相伴！ （團隊調節）還貨說完了 該說重頭戲了：乾貨技巧和方 思路（這玩意我自己都沒懂） 這個是一個系統的科學，從頭開始 打字太多了 好累啊 休息休息 看看有沒有人願意看吧 貌似被群嘲了 來吧 讓潔白的兩點來的更猛烈點

Jieba 斷詞後濾掉 stop word 的結果

調節 / 參加 / 附近 / 行業 / 沙龍 / 論壇 / 培訓 / 接觸 / 提升 / 圈子 / 調節 / 團隊 / 團隊 / 互相 / 討論 / 開會 / 舍伙人 / 基友 / 連袂帶 / 順 / 好夕 / 拉過來 / 相互 / 扶持 / 相互 / 快樂 / X2 / 痛苦 / 非常感謝 / 一路 / 相伴 / 團隊 / 調節 / 漏 / 貨 / 說完 / 重頭戲 / 乾貨 / 技巧 / 思路 / 玩意 / 沒懂 / 系統 / 科學 / 從頭開始 / 打字 / 太多 / 好累 / 休息 / 休息 / 看吧 / 貌似 / 群嘲 / 潔白 / 兩點 / 猛烈

探勘文本主題有幾種方法，在這裡我們主要是運用了文字雲、tf-idf 關鍵詞 和 LDA 這三種方法來進行觀察，以下我們以 pokemon 資料集為例。

根據詞頻產生文字雲。

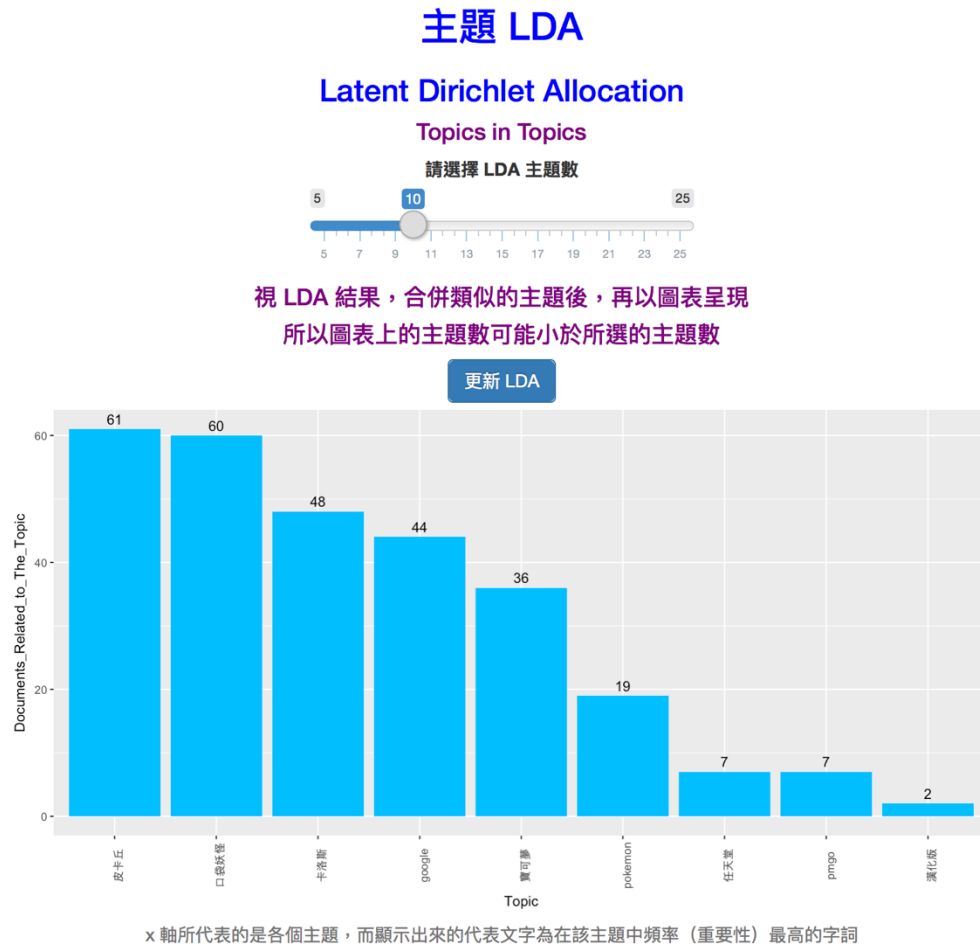


我們將主題中 tf-idf 值進行排序，前幾名高的詞彙對該主題的代表性就越強。



3. 主題 LDA (Latent Dirichlet Allocation) 分析

LDA 是一種非監督式的文本主題探勘模型，我們將資料集中的所有文章透過 LDA 分析後就可以發現其中隱藏的子主題。(Topics in Topics)



4. 除了上述三種方法外，我們的網站也有將熱門的問題和答案做成 table，大家可以透過操作看看有沒有什麼有趣的問題和答案。

請選擇一個問題

寶可夢日月要如何使用金手指？ ▼

看問題內容與熱門回答

問題內容

好久不登賬號，不知道是誰改的問題，現在改回來了。這個問題一開始只是發一下牢騷，沒想到會火 原問題描述 其實我畫畫也不怎麼好，只會臨摹，小時候由於無聊，上課又不愛聽課，平時都被鎖家裡不讓出門，實在沒事幹所以天天畫畫，不知不覺畫畫水平有點提高，也就是臨摹什麼都有個七八分像的感覺。現在上大學後，開始討厭畫畫了，知道自己不是這方面天才，只是無聊畫的多才稍微畫的比一般同學好一點點而已。但我爸以為我非常有天賦，讓我好好學畫畫將來在路邊給人畫像謀生 我就很鬱悶，我雖然學業不行，但也不至於落的個在馬路牙子上給人畫臉的地步吧 A 類似問題：畫畫有什麼用？ - 設計

熱門回答

Ans

Upvote

2266

題主改了題目，原問題是 畫畫有什麼用，我的回答是會畫畫簡直讓人生活充滿樂趣！看來題主是迫切需要一些工作上的指引，正好來更新一下，寫下這個答案到現在過了半年了，我學會了畫除了自己生活以外的故事，在一些平臺發佈自己的作品，也有些人賞識，學會了與人溝通，已經和出版社合作了幾本書的插圖工作，預計下半年會出版。當初沒想到這些小畫真的會帶給我工作收入，只是很喜歡，所以一直畫著，畫畫教會我的就是不會著急，不會想著做一件事就一定馬上有什麼回報，只要你願意堅持去思考，去努力，回報自然會有的。以下是原答案居然破千了超感謝你們的贊！！萬聖節還要上班的我抓緊一點時間塗了張送給你們～沒穿衣服的我（此處應有表情）～怎麼礙為了閱讀起來更順暢，把所有更新都放進答案正文里以下是正文 會畫畫簡直讓人生活充滿樂趣！比如說前一陣子剛擺完酒，這是我自己畫出來製作的婚禮請柬這裡就不展開說了，對請柬感興趣的同學可以詳見這裡有哪些好看的婚禮請柬？- 鐘小朋友的回答整場婚禮沒有請婚慶公司，全是自己的佈置，把平時的小畫打印出來每張桌子擺幾張，走廊上擺著一排畫架放著自己的畫...不是什麼豪華的婚禮，但是是屬於自己的，小小的，很溫馨。當天太忙了都沒有把這些佈置拍下來，找到兩張跟拍的攝影師拍的.....從小一直有寫日記的習慣，直到有一天突發奇想用畫畫來記錄，然而一開始的畫是這樣的...更醜的我就不好意思拿出來了 但我相信只要一直堅持畫總會有進步的！於是過了幾個月大概是這樣的...嘿好像有了那麼一丟丟進步...再繼續畫...感覺有點得心應手了啊對不對！後來我又畫起了故事畫了好多篇小故事...然後就不滿足於黑白了開始畫彩色！剛開始畫的也是慘不忍睹 雖然是科班出身但從來沒畫過水彩...所以一開始顏色完全摸不著頭腦，顏色糊糊的 自己被打擊到了...有一段時間沒有畫直到有一天看到一個外國插畫家的圖，覺得好贊，於是借用它的背景半

六、情感分析 (Sentiment)

情感分析的部分，我們在網頁上製作了一個情感分析機，使用者可以輸入一段文字並得到一個情感指標。

1. 請輸入一段文字

你正在想什麼呢？

藍瘦，香菇，本來今顛高高興興，泥為什麼要說這種話？藍瘦，香菇在這裡。第一翅為一個女孩使這麼香菇，藍瘦。泥為什麼摸要說射種話，丟我一個人晒這裡，香菇，藍瘦在這裡，香菇...

2. 依據文字內容作情感分析並顯示心情指數

● 正負面詞彙文字雲

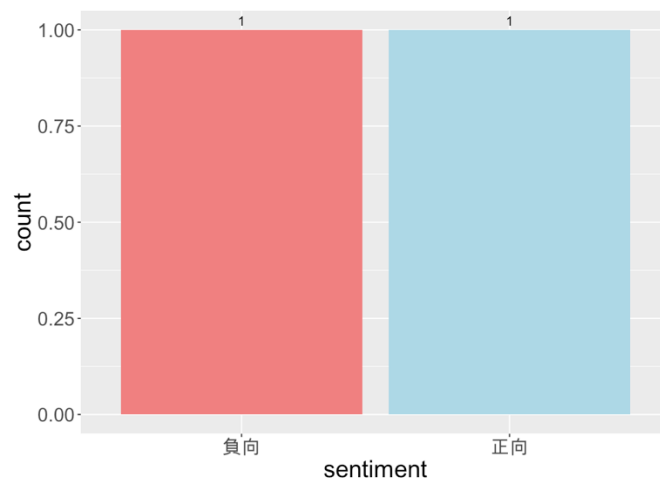
Positive WordCloud

興
興
嘔
嘔

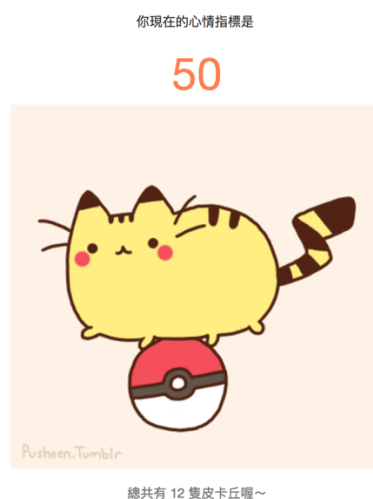
Negative WordCloud

瘦

● 詞頻統計



● 心情指數和一隻代表你心情的皮卡丘



3. 如果覺得不準，也可以自行新增正、負面詞彙

不準啦！
我要...

請以空白相隔詞彙
輸入範例：藍瘦 香菇

增加正向詞彙

增加負向詞彙

4. 情感字典來源

- NTUSD:

<https://docs.google.com/forms/d/e/1FAIpQLSe20EyOE3bp9cKT0gF6R4DodTHOmriIGegkGYa03oHYejhi9g/viewform?c=0&w=1>

- der3318:

<https://github.com/der3318/SentimentAnalyzer/tree/master/docs>

- Tidytext bling:

<https://github.com/juliasilge/tidytext>

七、文本相似度 (Essay Similarity)

要計算兩個文本的相似性，我們利用了餘弦相似性，透過 cosine 我們可以算出 A 和 B 兩個 vector 的相似度。

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

在這邊我們令 A 代表問題文本詞頻的 vector、B 代表回答文本 tf-idf 值的 vector，並通過 cosine 計算出來兩者之間的相似度。

1. 首先選擇一主題

請選擇一個主題

current_event ▼

更換資料集

2. 選擇與此主題相關的問題以及其下的某一回答，就能顯示出回答與問題之間的相關程度。

以下是我們將 <問題詞頻> 與 <答案 tf_idf> 這兩串向量
通過 cosine 所計算出來的相似度

<p>請選擇一個問題</p> <p>不吹牛逼、八達嶺那事要是你、你連敢下去 和老虎玩命不？</p> <p>看問題內容</p>	<p>請選擇一個答案</p> <p>1</p> <p>看答案</p>	相似度
--	------------------------------------	-----

<p>新聞說作死的那個連沒死，死的是殺人的，別動不動就說人家缺心眼神經病什麼的， 為救家人敢和老虎生嗆也是種巨大的勇氣，不吹牛逼，遇到這樣情況，你敢麼</p>	<p>如果我的親人智商這麼下線，不帶腦子的話，我只能，嘔。</p>	40.82 %
---	-----------------------------------	------------

八、文本分群 (Clustering)

我們分群的方法是將 Term Document Matrix 用 PCA 降階後再餵給 kmeans。

1. 先選擇一個主題以及問題

請選擇一個主題

art

更換資料集

請選擇一個問題

有哪些演員顏值特別高的電影或電視劇？▼

選擇問題

2. 選擇答案分群數

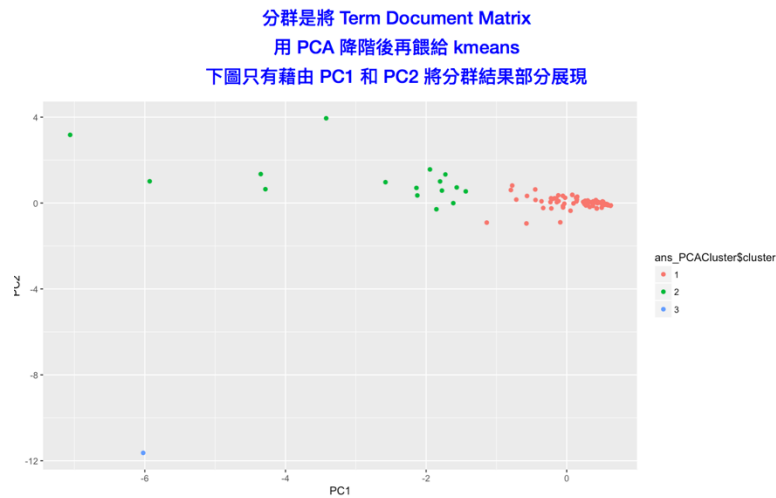
選擇答案分群數

2 3 8

2 3 4 5 6 7 8

開始分群

3. 展示分群效果 (以下只畫出 PC1 和 PC2 的關聯)

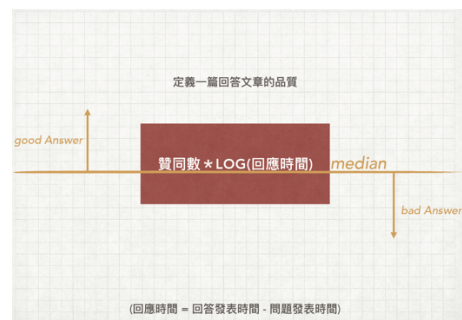


我們認為分群後可以將文本與各分群群簇的接近度作為文本 **features** 來看待，舉例來說假設今天我們的文本集是有關政治的文本集，那分群後的結果有可能各個群簇都各代表了某一政黨的言論。那以一篇文章來說，我們可以觀察其在分群後的與各群簇的接近度，並以此判斷他比較屬於哪一個或哪一些黨派的言論。

九、支援向量機 (SVM)

這部分是我們這份專題最關鍵的部分，希望根據我們之前分析的結果進行回答的評分。在這邊我們使用了 SVM 的機器學習模型，使用 **e1071** 的套件來完成這項任務。

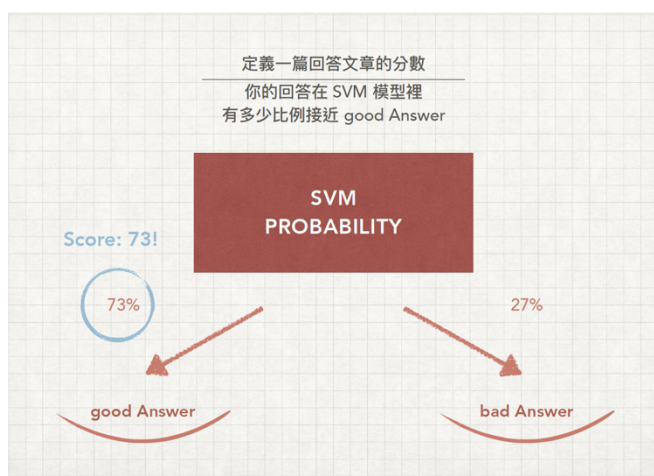
在幫回答評分之前我們必須先定義和區分 **training data** 中好的答案和壞的答案，在這邊我們認為答案之好壞可以從其贊同數來觀察，另外因為回應時間也是一個影響贊同數的因素，所以我們將 $\text{scale}(\text{贊同數} * \log(\text{回應時間}))$ 作為我們評判好壞的標準值，在標準值的中位數以下即代表壞答案，標準值的中位數以上即代表好答案。



在 SVM 這個監督式的機器學習模型裡，我們將答案的好壞作為我們的目標 label，然後將以下項目作為我們的 features 來進行 training。

1. 回答文本的情感分數
2. 問題文本與回答文本的相似度
3. 文本分群後抽出來的接近度 features
4. 回答文本的字數
5. 回答文本的詞數
6. 回答文本的 stop word 數
7. 回答文本的 stop word 比例
8. 回答者的被追蹤數
9. 回答者的追蹤數
10. 回答者獲得的總讚數
11. 回答者的總感謝數
12. 回答者的發問數
13. 回答者的回答數
14. 回答者的文章發表數

最後，SVM.probability 產出來的模型可以 predict 出答案有多少比例是接近好答案，有多少比例是接近壞答案，那我們就依其接近好答案的比例作為我們對該答案的評分。舉例來說假設今天有一個答案經過 SVM 的 prediction 後，得知他有 73% 會是好答案，27% 會是壞答案，那就將此答案的分數定為 73 分。



本頁有提供一個範例題目，使用者可以在欄位中任意輸入相關的答案，再填入個人的基本資料後點擊「看看我的回答水準」即可查看該回答之評分，評分越高者及代表你的回答品質越高。

1. 瀏覽問題

怎麼徹底的識別一個男人的本質？

我小姨的故事。先直接跳到結果，她因為不能生小孩被離婚，之後受打擊太大，得了癰疽，做了開釦手術，現在生活無法自理，思緒像小孩一樣，全靠外公外婆照顧。她年輕時很漂亮（很像舒暢），追求者眾多，但是現在連正常人都不知。而前夫娶了三之後，生了兒子，買房買車遷升了小學校長，日子過得很滋潤。我想強調的是：1 這個男人性格溫和和有耐心，溫文爾雅。這種十多年的長期相處，性格是裝不出來的。他偏感情很好，直到離婚。我小姨比較好強，自尊心強，要面子，強勢。2他是人貴上門的女婿，家裡比較窮，外公家比較富裕，所以他家在那年代受過恩惠和好處，說了困境，曾經有很長一段時間他的親戚們都住在外公家。3這個男人在離婚之後，我小姨開釦手術住院期間以及其後漫長的歲月里，沒揮霍過一次，沒有任何聯繫。即使那次手術小姨差點有生命危險不省人事，對一個被自己傷害過的女人。怎麼這麼無情呢我無法理解。一個看上去又顧家又有責任感的男人，卻也同時是弱雞的渣男。以此為例，我想知道，怎樣才能看透一個人的本性，從眾多假象和偽裝中看到本質。看評論發現我漏了一些很重要的信息1小姨比姨父大三歲，他倆離婚的導火索好像是其地，小姨父親在到另一個學校教書之後，不常回來（開車得2個多小時的路程），倆人見面次數少了。他認識了一個離異的女人（我看過照片，挺漂亮），家裡有錢，比小姨小10歲，跟前夫有孩子。後來他跟小姨就離婚了，之後又結婚生了一個兒子。2我小時候最喜歡的就是小姨父，超過對我爸爸的喜歡，會想他是我爸爸多好。這麼多年之後，離開小姨不說，我是懷念他的。他從來不發脾氣，帶我看影碟恐怖片，給我講腦筋急轉彎，比爸爸有耐心多了。只是現在是一個模糊的影子，長相記不清了。最後兒他的那年春節，家裡氣氛很詭異。我當時小不知道，但突然有了隔層的擔憂，他給我壓歲錢之後，我問他：XX叔叔，你明年還來玩嗎？媽媽瞪了我一眼，他笑著說：「會來的」。之後再沒講過話。有一次在學校遠遠的看到他，想跟他說話但沒去（我媽不讓）。他看到我了但沒跟我說話。另一次在音樂節上，遠遠看到他托舉著一個小嬰兒，笑得很開心。3爭議最大的是關於小姨的性格，她確實強勢，這是問題。但不是那麼強勢，小性子什麼的沒你們想的那麼誇張。而且她一直這樣，都知道。後來也有人追她，她沒答應。而且關鍵是那個小三也強勢啊管他很嚴。怪我沒說清楚。4他的工作能力？我不知道。他教中學的時候我讀小學。我讀中學的時候他調走了。之後就離婚了。5家人對他的看法怎麼從那以後沒有一個人提他。就像他從來不存在一樣。大家都不想揭別人傷疤。但我媽和外婆是討厭他的，外公不想談這個人。我爸爸只是說：「他胖了發福了，生了兒子，升了校長，小日子過得挺滋潤的（原話）」。也不想多談。沒人認為他比我們低一等。原來大家都喜歡他。6有人說我來找認同感，痛批此人是渣男。沒有，大家一起罵他是渣男對事實不會有任何改變。我只是無法把我心目中的那個溫和和有耐心、想讓我當爸爸的人、和後來這個人聯繫在一起。我花了很久才相信這是同一個人。7有人說我把小姨得離婚這事導致的完全怪罪到他頭上。那麼我只應這一個事實：離婚之後，小姨就像變了一個人。精神上消沉，具體過程我在讀書不清楚，但我媽清楚。之後的小姨跟之前的感覺完全不一樣了，她不再像以前一樣買很多好看的衣服，不再熱衷於打扮，而且她現在的白髮比我媽還多（我媽比她大6歲）。我想強調的是，離婚這事不全是離婚導致的，但必然與之脫不了關係。8有人說我帶有濃重的主觀色彩。確實，對於一個童年我曾經最喜歡的人，我是懷念他的。對於一個拋棄我至親的小姨，傷害她最重的人，我是恨他的。你要我拋棄感情去談這個事情，我做不到。明哈。看完了。等到有一天找到小姨父。親自跟他聊了之後再更人性複雜。人心變化。凡夫俗子的我們是看不透的。所以我們都只能被命運推來推去了麼。只能祈禱命運對我們溫柔一點。麼。你們願意認命嗎新消息。小姨要說一個三十歲左右的退伍軍人再婚了（年紀相差二十歲左右）感覺此人性格內向溫和。不善言辭。小姨的朋友在她結婚前夕找了前小姨父，故意告訴他小姨馬上就要跟一個各方面條件都不錯的人結婚了。想看他反應。他第一句話是。是不是看上她的錢了？後來讓轉告他的祝福，說如果小姨過不好這輩子他都不會安心。我後來知道，這些年他自己也動過心釦手術。而且大出血，在鬼門關走了一趟。原來他內心其實有愧疚的。他過得也沒那麼好。願問題：怎麼徹底的識別一個男人的本質？。戀愛

2. 輸入回答

請輸入你的答案

女人 豆腐心 海底針

男女以和樂為貴

還不夠如火如荼啊。

主要是我覺得目前還是誰有錢有權誰更容易風口浪尖，等有一天呂姓的家庭社會地位都高於男性了，分析這個問題會更有意義。

3. 輸入個人基本資訊

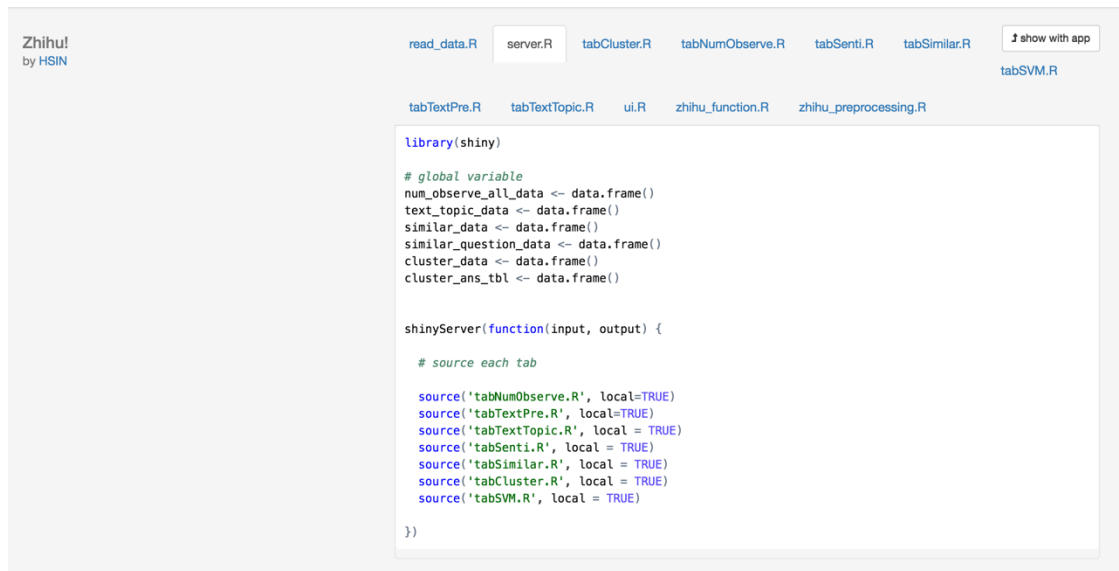
author_followee_num	9
author_upvote_num	27
author_thank_num	4
author_answer_num	7
author_question_num	4
author_post_num	11

4. 看分數



十、程式展現

本網頁由 R 語言撰寫，並經由 Shiny app 展現，在網頁的最下方有我們的 display code，可以完整的看到我們程式是如何寫的。



```
library(shiny)

# global variable
num_observe_all_data <- data.frame()
text_topic_data <- data.frame()
similar_data <- data.frame()
similar_question_data <- data.frame()
cluster_data <- data.frame()
cluster_ans_tbl <- data.frame()

shinyServer(function(input, output) {

  # source each tab

  source('tabNumObserve.R', local=TRUE)
  source('tabTextPre.R', local=TRUE)
  source('tabTextTopic.R', local = TRUE)
  source('tabSenti.R', local = TRUE)
  source('tabSimilar.R', local = TRUE)
  source('tabCluster.R', local = TRUE)
  source('tabSVM.R', local = TRUE)

})
```

十一、分工細項

1. 爬資料：陳信豪、張景淵、冷俊瑩
2. 清理資料（斷詞、濾 stop words、標記資料等）：陳信豪、周昺瑤、高偉立
3. 基本分析（數據統計部分）：周昺瑤
4. Model（機器學習部分）：陳信豪、高偉立
5. 資料呈現彙整（Shiny / Notebook html）：陳信豪
6. 書面：冷俊瑩、黃薇甄

十二、附錄

本專案之 Coding 和相關檔案：

https://github.com/OOmegaPPanDDa/shiny_dsr_zhihu