文字分析於校務研究的應用

淡江大學資訊處 曹乃龍

為何要作文字分析？

不要只依靠冰冷的數字

校園之外的資訊也很重要

LEVELS OF ANALYTICS

Analytics

文字雲

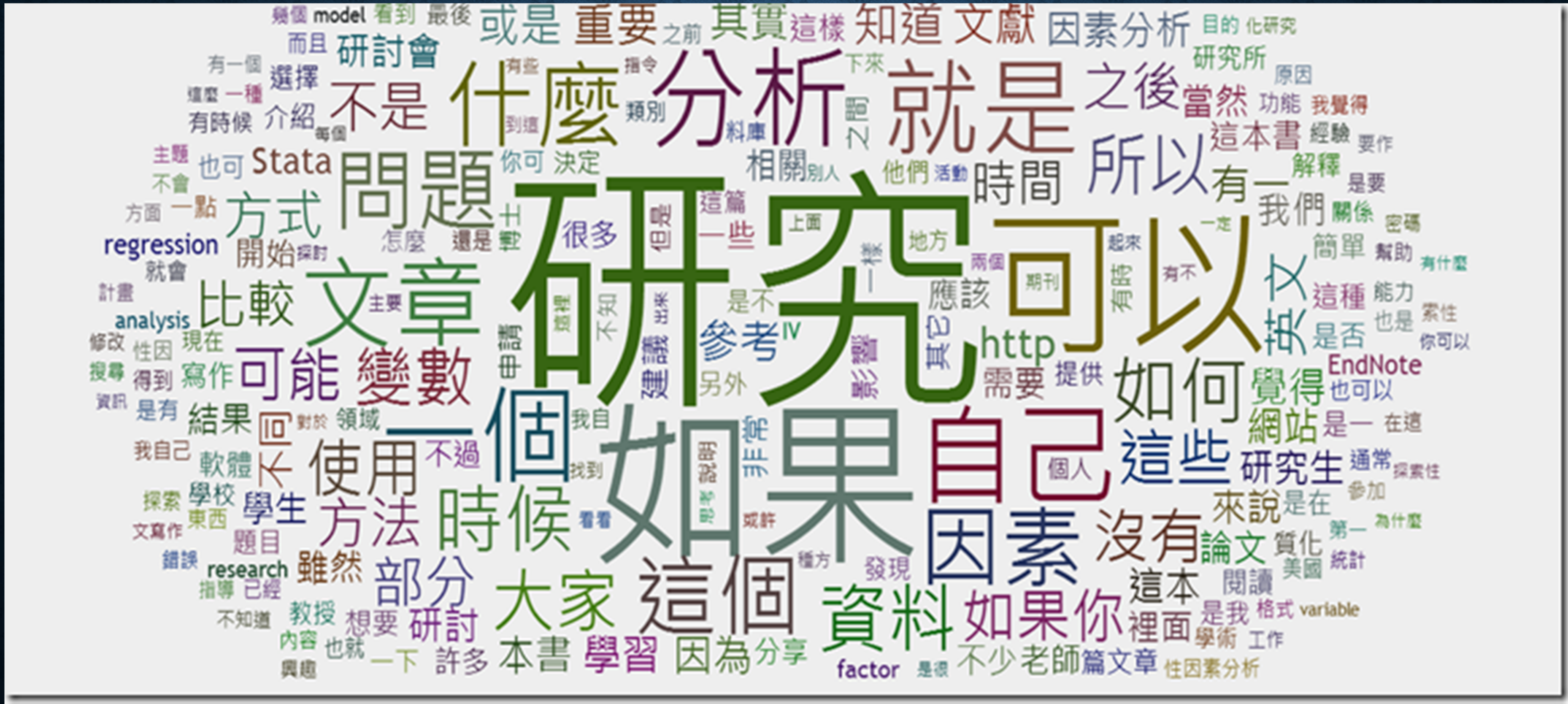
Handy Analytics

討論區評分

Actionable Analytics

課程設計輔助

ANALYTICS


研究
可以
如果
自己
分析
就是
什麼
問題
文章
一個
因素
這個
資料
如果你
時候
大家
使用
方法
部分
變數
可能
比較
方式
所以
如何
這些
沒有
之後
不是
重要
或是
其實
知道
文獻
因素分析
研討會
選擇
介紹
Stata
regression
analysis
research
factor
EndNote
http
參考
相關
時間
有一
我們
英文
網站
研究生
覺得
來說
論文
這本
閱讀
裡面
不少老師
篇文章
學習
因為
本書
許多
研討
想要
教授
雖然
學生
結果
寫作
軟體
不同
開始
很多
需要
提供
應該
另外
建議
非常
一些
個人
時間
當然
這本書
研究所
model
variable
性因素分析

HANDY ANALYTICS

- 定義：進行後可即時實際使用的分析結果
- 範例：討論區評分
  - 合作夥伴：滁州学院计算机与信息工程学院 程曦資訊
  - 資料來源：滁州学院計算機概論MOOC平台討論內容，共有4,868主題及42,135回應。
  - 外部資料：Google Search

使用討論區評分（Before）

- 目的：利用課程討論區了解學生參與程度並給予評價
- 方法：利用發表主題及回應的次數及字數作為評分標準
- 問題：
  - 主題：哪个杀毒软件好？
  - 回應一： 360卫士
  - 回應二: 我不知道！

使用討論區評分（After）

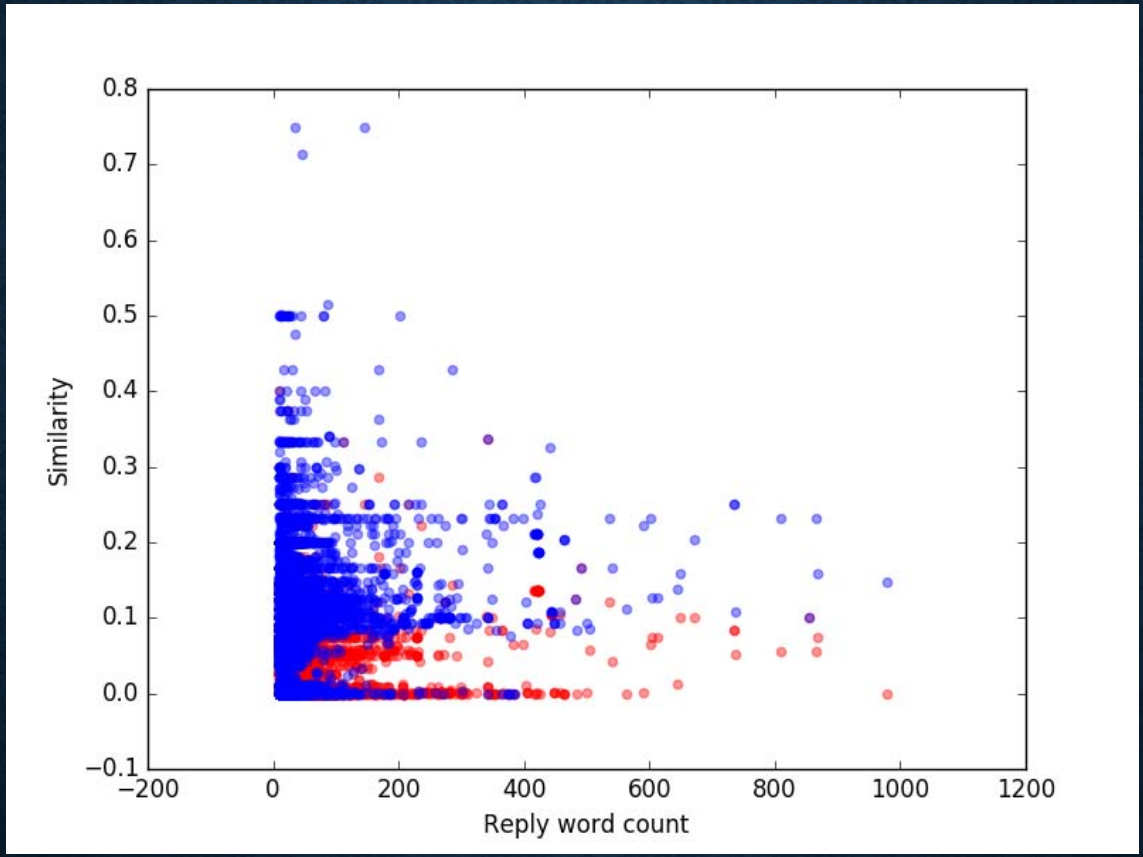
- 目的：分析回應內容判斷是否符合主題，以利給與正確評價
- 方法：利用Google Search 對回應進行內容擴充
- 問題：
  - 主題：哪个杀毒软件好？
  - 回應一： 360卫士，擴充後包含杀毒软件
  - 回應二: 我不知道！，擴充後不知所云

School Inisght

Text Analytics Engine

Text Similarity

Google Search



Red: Before expansion, Blue: After expansion

SCHOOL INSIGHT INTEGRATION

- Data import
- Create dashboard menu item
- Pick attributes for visualization

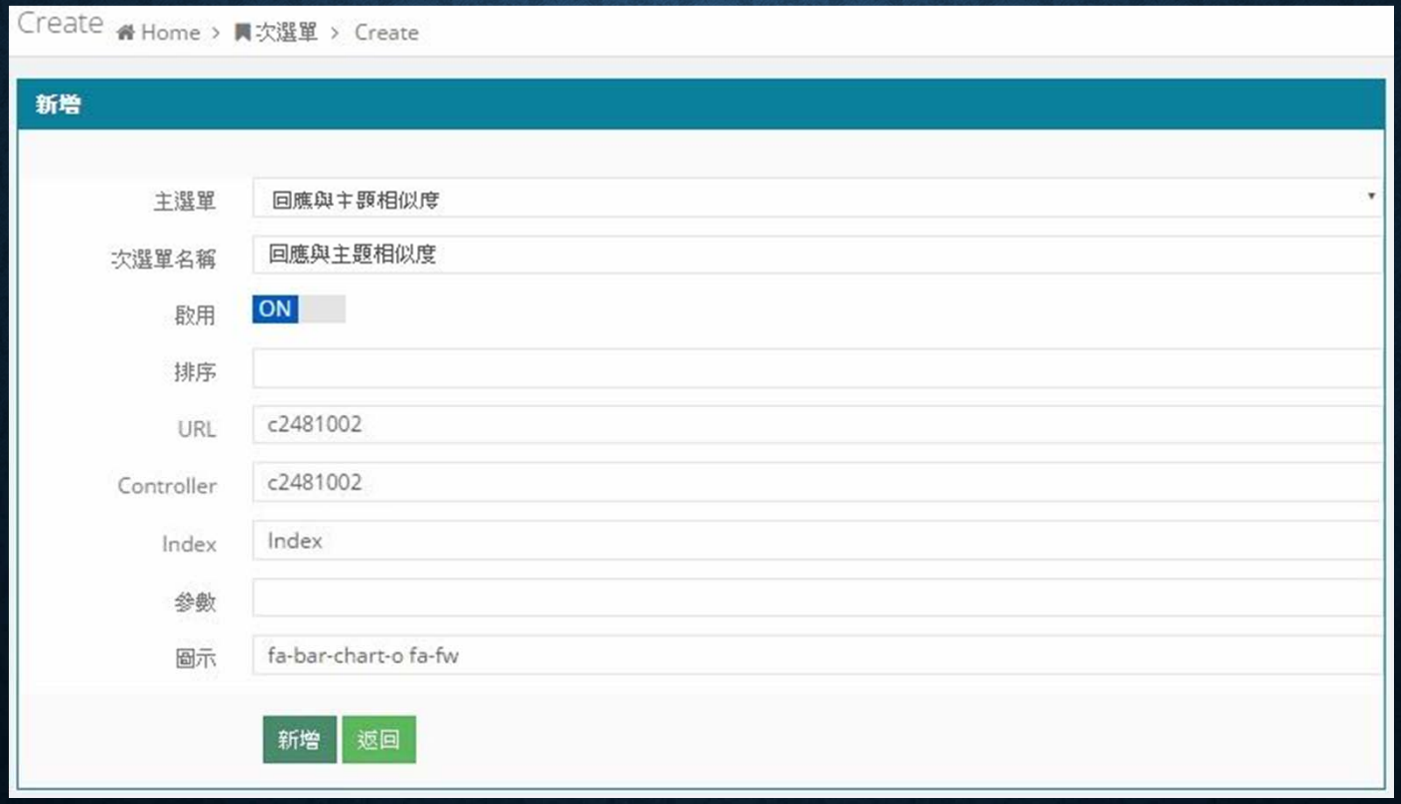
DATA IMPORT

cx_topic_reply_sim	cx_topic_reply_sim	14	1	674319	1001281623	357820
cx_topic_reply_sim	cx_topic_reply_sim	19	1	327710	1001280939	816933
cx_topic_reply_sim	cx_topic_reply_sim	22	1	660827	1001280589	157471
cx_topic_reply_sim	cx_topic_reply_sim	24	1	846383	1001280665	357598
cx_topic_reply_sim	cx_topic_reply_sim	26	1	327728	1001280978	816933
cx_topic_reply_sim	cx_topic_reply_sim	29	1	753723	1001280667	275322
cx_topic_reply_sim	cx_topic_reply_sim	32	1	491597	1001280611	816855
cx_topic_reply_sim	cx_topic_reply_sim	41	1	262222	1001509469	816933
cx_topic_reply_sim	cx_topic_reply_sim	44	1	987501	1001281036	443025
cx_topic_reply_sim	cx_topic_reply_sim	48	1	131162	1001239606	184964
cx_topic_reply_sim	cx_topic_reply_sim	52	1	196703	1001282160	792160
cx_topic_reply_sim	cx_topic_reply_sim	60	1	327786	1001280978	816834
cx_topic_reply_sim	cx_topic_reply_sim	63	1	753686	1001280667	364590
cx_topic_reply_sim	cx_topic_reply_sim	66	1	125634	1001280695	392228
cx_topic_reply_sim	cx_topic_reply_sim	69	1	327831	1001280970	151123
cx_topic_reply_sim	cx_topic_reply_sim	72	1	327838	1001280978	816876
cx_topic_reply_sim	cx_topic_reply_sim	75	1	1037670	1001283636	157516
cx_topic_reply_sim	cx_topic_reply_sim	78	1	196783	1001240271	357667
cx_topic_reply_sim	cx_topic_reply_sim	99	1	557232	1001281692	816855
cx_topic_reply_sim	cx_topic_reply_sim	105	1	688312	1001282871	357598
cx_topic_reply_sim	cx_topic_reply_sim	108	1	327867	1001280939	143206
cx_topic_reply_sim	cx_topic_reply_sim	110	1	649930	1001281902	185162
cx_topic_reply_sim	cx_topic_reply_sim	116	1	327875	1001280970	148901
cx_topic_reply_sim	cx_topic_reply_sim	119	1	131273	1001281091	357751
cx_topic_reply_sim	cx_topic_reply_sim	120	1	933303	1001281288	157479
cx_topic_reply_sim	cx_topic_reply_sim	123	1	327885	1001280978	357598
cx_topic_reply_sim	cx_topic_reply_sim	126	1	884951	1001280390	377781
cx_topic_reply_sim	cx_topic_reply_sim	129	1	384085	1001239915	816906

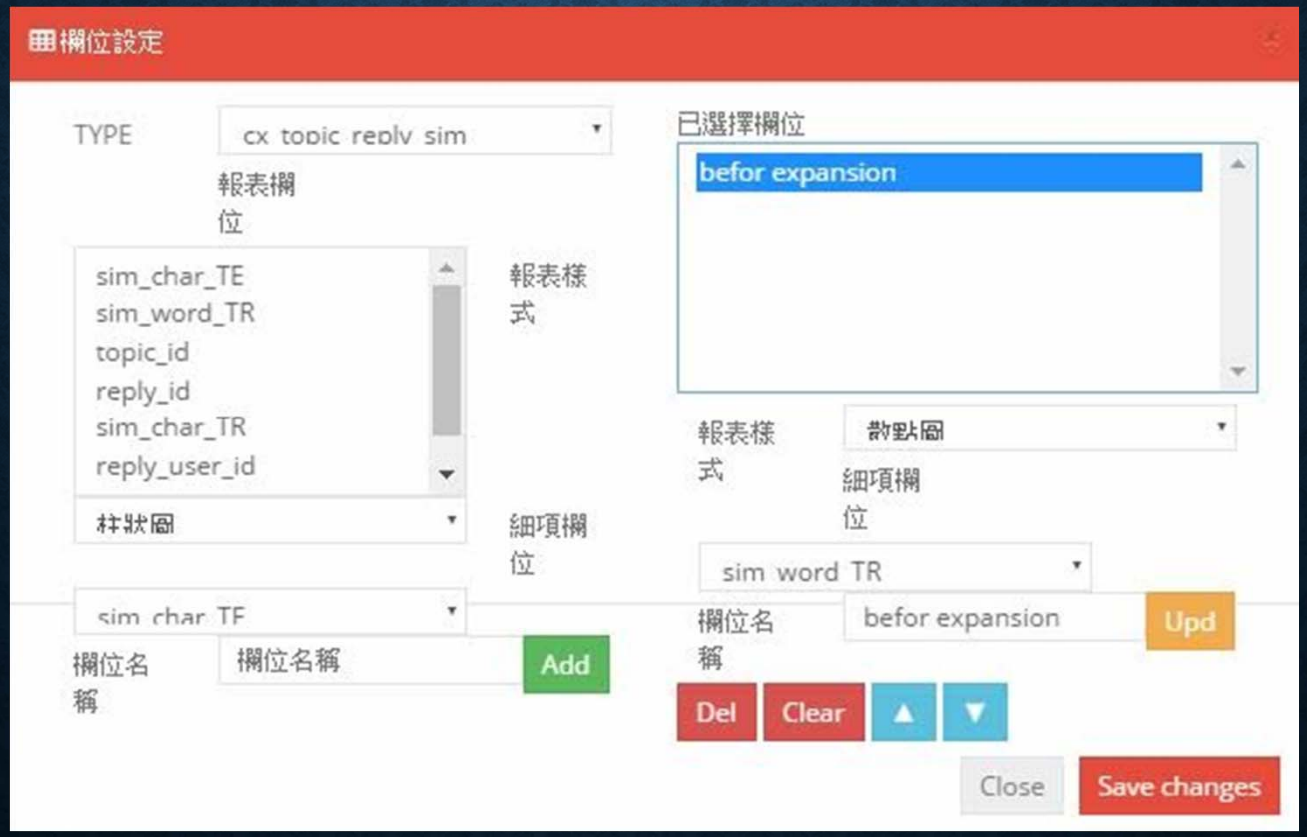
原始数据


```
{
    "_index": "cx_topic_reply_sim",
    "_type": "cx_topic_reply_sim",
    "_id": "14",
    "_version": 1,
    "_score": 1,
    "_source": {
        "nst_sim": [
            {
                "reply_id": 674319,
                "reply_user_id": 1001281623,
                "topic_id": 357820,
                "sim_char_TR": 0.000018,
                "sim_char_TE": 0.071516,
                "sim_word_TR": 0.000018,
                "sim_word_TE": 0.125031,
                "reply_word_count": 12
            }
        ]
    }
}
```

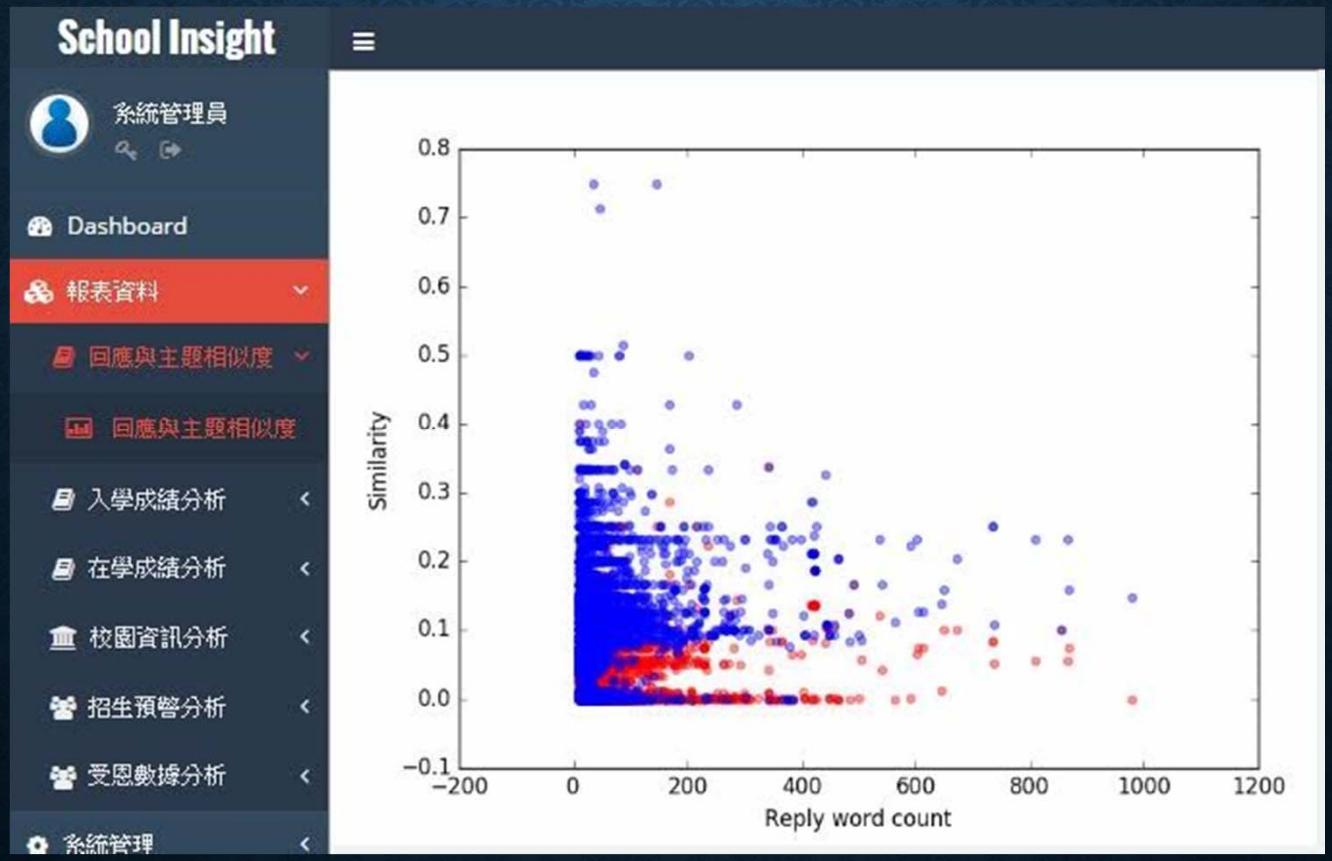
CREATE MENU ITEM



PICK THE ATTRIBUTES FOR VISUALIZATION



VOILÀ


School Insight
系統管理員
Dashboard
報表資料
回應與主題相似度
回應與主題相似度
入學成績分析
在學成績分析
校園資訊分析
招生預警分析
受恩數據分析
系統管理
Similarity
Reply word count
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−200
0
200
400
600
800
1000
1200

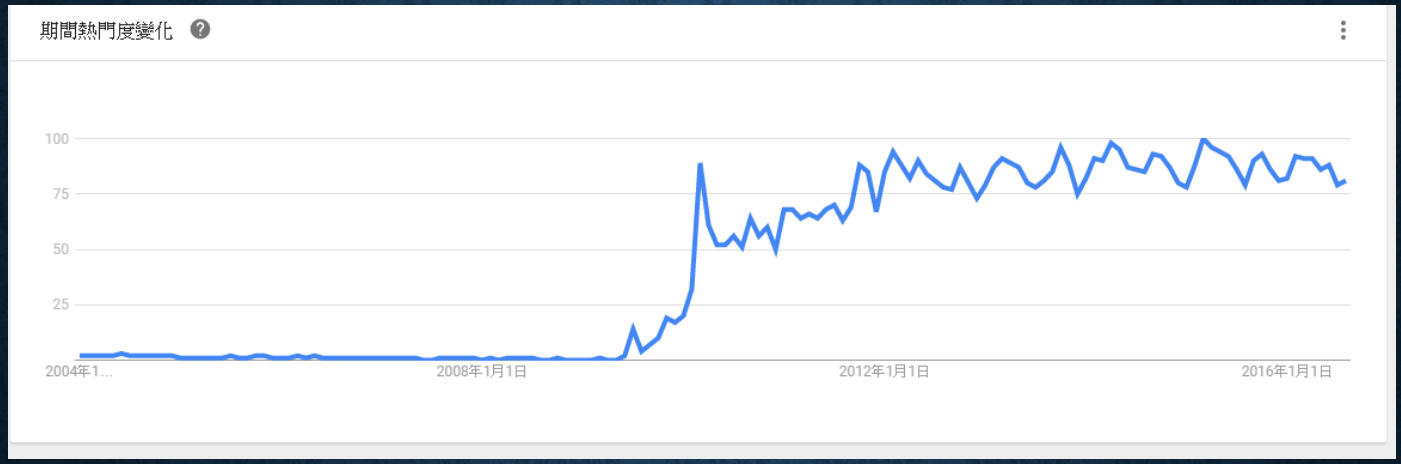
WHY SCHOOL INSIGHT?

1. Reusable modules
2. Easy to integrate web visualization libraries, such as Highchart, matplotlib, D3.js and Google Chart.
3. Easy to integrate analytics engine designed by common development tools, such as Python, Java and R.

ACTIONABLE ANALYTICS

- 定義：進行後可供實際決策的分析結果
- 範例：課程設計輔助系統
  - 最終目的：縮短學用落差
  - 方法：擷取企業所需且正在高成長的技術
  - 資料來源：104人力銀行、Wikipedia、Google Trend及教學計畫表

NOSQL



	數量	備註
104.com.tw	303職缺	2016.8.31資料
淡江課程	3門課	102-104, 同一教授

APPROACHES

1. Get job description
2. Extract key term candidates
3. Use Wikipedia to filter key terms
4. Extract Google Trend for these key terms
5. Count occurrence in teaching plans

GET JOB DESCRIPTION

- 方法：使用104 API
- 目標：職缺類別為軟體設計工程師
- 資料：共有4468個職缺
- 技術字彙來源：職稱、職缺描述及其他條件

EXTRACT KEY TERM CANDIDATES

- 使用斷詞器(Jieba)
- 以小單位為優先，例如「平行處理經驗」會被斷為「平行」 「處理」「經驗」
- 技術字彙擷取時除了考慮單字，也考慮bigram

Nosql	Asp.net MVC	Machine learning	Stored Procedure	Programming skills
○	○	○	○	○

USE WIKIPEDIA TO FILTER KEY TERMS

- Only keep those key terms which are also Wikipedia article title
- Future work: Chinese Wikipedia

Nosql	Asp.net MVC	Machine learning	Stored Procedure	Programming skills
○	○	○	○	×

EXTRACT GOOGLE TREND FOR THESE KEY TERMS

- Use the difference between trends of last year and first year

Nosql	Asp.net MVC	Machine learning	Stored Procedure	Programming skills
○	○	○	×	×

COUNT OCCURRENCE IN TEACHING PLANS

- 22,898 teaching plans from 102-1 to 104-1

Nosql	Asp.net MVC	Machine learning	Stored Procedure	Programming skills
○	○	×	×	×

Term	Freq	MI/RP	Trend	Plan
asp.net mvc	68	44.86	83	0
android studio	43	39.7	91	4
entity framework	37	38.7	89	0
deep learning	48	36.88	78	0
mvc	451	17.75	33	2
node.js	131	17.55	90	0
json	125	17.48	92	4
angularjs	114	17.35	95	0
nosql	104	17.21	85	3
mongodb	73	16.7	95	3
github	67	16.58	89	2
laravel	61	16.45	85	0
xcode	59	16.4	38	3
openstack	52	16.21	85	0
opencv	47	16.07	54	5
devops	46	16.04	83	0
redis	45	16.01	84	3
nginx	44	15.97	92	0

WHY IT'S ACTIONABLE?

- 反例：學生畢業高中分佈
  - 雖然能知道分佈情形，但由於只是數字多寡，而且逐年不同，管理者該如何利用這個分析並無明確做法。

- 為什麼這個分析可執行?
  - 管理階層可明確了解業界所需及高成長、高需求的技術內容
  - 系所及教師可根據分析結果加強課程內容或訂定研究方向

THANK YOU!