KYPER
DATA TECHNOLOGIES

巨量資料分析

文字探勘基礎介紹

2017/05/19

Kyper Data Technologies

大綱

- 文字探勘簡介
- 工具介紹(英文&中文)
- 案例與實作分享
- 參考

文字探勘簡介
從文本產生有價值的訊息

- 透過模式識別等工具處理資料
  - 從非結構化到結構化
  - 分析結構化資料並得到Insight
- 相關領域
  - 自然語言處理(Natural Language Processing)
  - 資訊檢索(Information Retrieval)

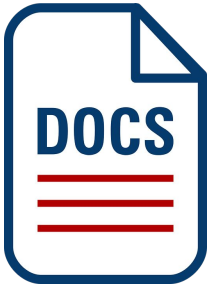
文字探勘應用
垃圾郵件偵測

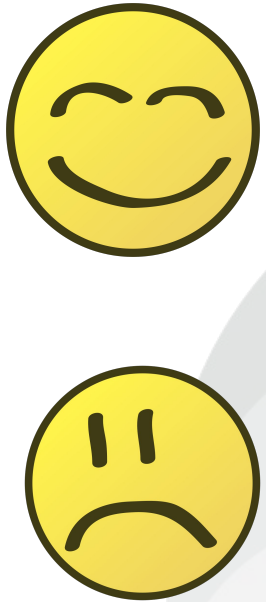
文字探勘應用
命名實體辨識

受到美股再創歷史新高的激勵下，加權指數在今天（5月9日）也一度重回萬點，然而隨後下殺收黑的表現卻讓投資人在欣喜之餘出現了一絲擔憂的心情，在 全球主要股市一片歡騰的情況下，投資人現在對於投資市場保持戒慎恐懼究竟是合理還是多慮呢？或許你越了解 巴菲特就越能知道該怎麼做。

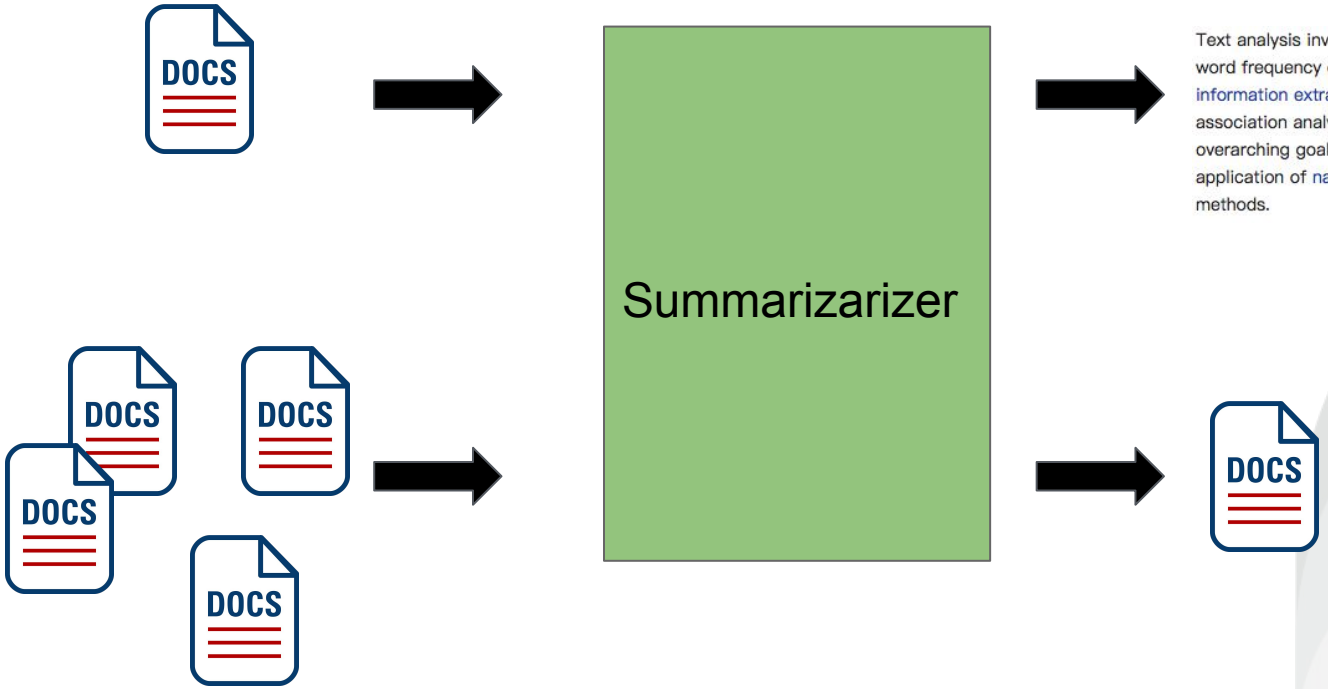
我們先從台灣的觀點來看，如果要挑一檔對加權指數影響最大的 美國個股，我想大家應該都會直覺的想到 蘋果，光是媒體上常提到的一顆 蘋果救台灣甚至是 蘋果概念股都相當受到重視，而無巧不巧的是，為何要買進蘋果股票也是這次 波克夏股東會的重要焦點，為何 巴菲特會想要投資過去不了解的蘋果公司？

文字探勘應用
情感分析





文字探勘應用
文件摘要


DOCS
DOCS
DOCS
DOCS
DOCS
Summarizarizer
DOCS
Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

N元語法(N-gram)

N-1階馬可夫鏈模型 - 用前N-1個估計第N個的機率

- 語句的表示方法(Representation)

例句	… to be or not to be ...
一元語法(unigram)	…, to, be, or, not, to, be, …
二元語法(bigram)	…, to be, be or, or not, not to, to be, …
三元語法(trigram)	…, to be or, be or not, or not to, not to be, …

- 單位可以是字或詞
- 可估計語句的機率，用在不同應用上

N元語法(N-gram)
英文套件

R **ngram** packages

```
> x <- "a b a b a"
> library(ngram)
> ng <- ngram(x, n=3)
> get.phrasetable(ng)
  ngrams freq      prop
1 a b a     2 0.6666667
2 b a b     1 0.3333333
```

- 以空白隔開詞彙
- 不適用於中文

練習時間
操作N-gram

- 使用預先定義的N-gram的函數
- 畫出trigram文字雲
- 試著調文字雲格式(套件：wordcloud2)

處理細節

- 停用詞(Stop words)
  - 主要為功能詞
  - 英文如‘the’, ‘at’等，中文如‘啊’, ‘喔’等
- 標點符號
  - 用以斷句
  - 或直接拿掉
- 濾掉低頻詞

中文處理 - 斷詞

將有意義的基本單位(詞彙)切割出來

- 英文常以空白隔開詞彙

  “Donald Trump and Recep Tayyip Erdoğan deliver joint statements at the White House on Tuesday in Washington DC.”

  - 英文處理 - Tokenization
- 中文詞與詞會連在一起，電腦無法直接理解
  - “下雨天留客天留我不留“
- R的斷詞套件 - JiebaR
  - 支援斷詞以及標注詞性

中文處理 - 斷詞

JiebaR用法

- 快速斷詞模式

```
> library(jiebaR)
> qseg <= "統計學是資料科學的基礎"
```

- 一般模式

```
> words = "統計學是資料科學的基礎"
> tagger = worker()
> tagger <= words
[1] "統計學" "是"    "資料"   "科學"   "的"    "基礎"
```

- 標記詞性

```
> tagger = worker("tag")
> tagger <= words
     nt       v       n       n      uj       n
"統計學"    "是"  "資料"  "科學"    "的"  "基礎"
```

從詞到文件 - 如何表示一個文件

tf-idf表示法

Term frequency

詞在檔案中出現的頻率

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Inverse document frequency

總檔案數除以含有某詞檔案數

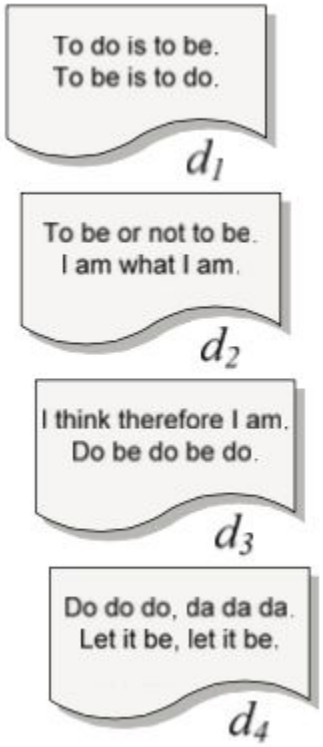
$$\mathrm{idf}_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

=> 某i詞在第j個文件中的比重

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

從詞到文件 - 如何表示一個文件

tf-idf矩陣



		d_1	d_2	d_3	d_4
1	to	3	2	-	-
2	do	0.830	-	1.073	1.073
3	is	4	-	-	-
4	be	-	-	-	-
5	or	-	2	-	-
6	not	-	2	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	2	-	-
10	think	-	-	2	-
11	therefore	-	-	2	-
12	da	-	-	-	5.170
13	let	-	-	-	4
14	it	-	-	-	4

搭配機器學習應用

- 監督式/非監督式學習
- 在許多應用上取得極佳成果
- 特徵工程
  - 描述內容的特徵，如寫作習慣
  - 搭配外部資源如字典
  - 看應用而定會有不同的metadata

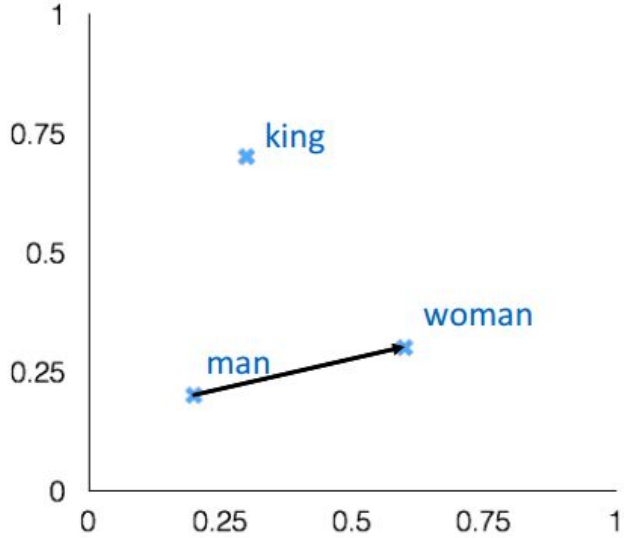
進階議題
深度學習熱潮

word2vec

- 將詞映射到向量空間中
- 有R套件**wordVectors**可使用

Recurrent Neural Networks

- Attention機制
- 在摘要、翻譯等應用取得極大進步



練習時間

- 比較Bigram跟斷詞形成的Tf-idf的差別
- 調整svm參數將準確度提高
- 觀察Clustering的結果
- Euclidean distance v.s. Cosine distance

參考

KYPER

Reference

- https://github.com/bmschmidt/wordVectors
- https://cran.r-project.org/web/packages/tm/tm.pdf
- https://www.r-project.org/nosvn/pandoc/jiebaR.html