G03

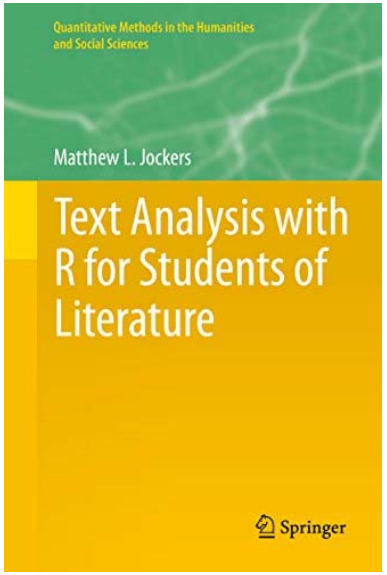
文字視覺化

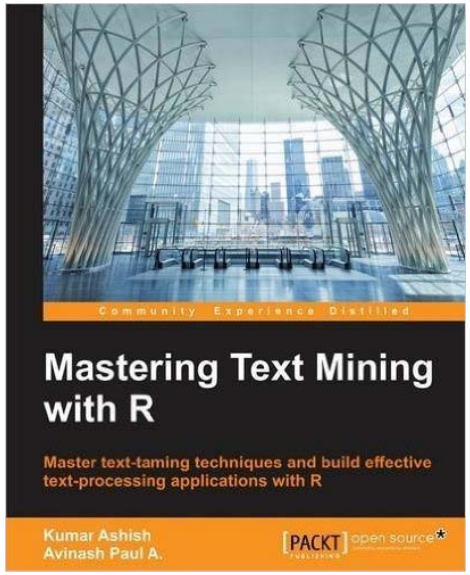
吳漢銘

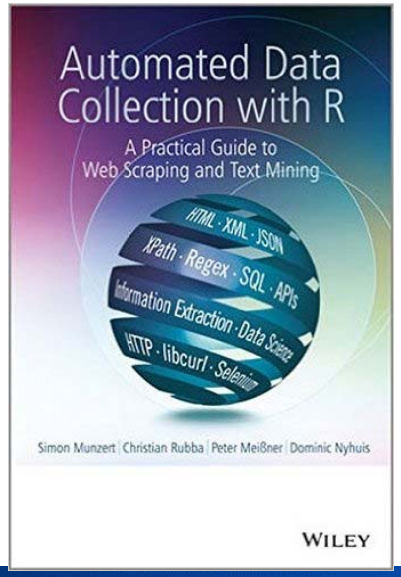
國立政治大學 統計學系

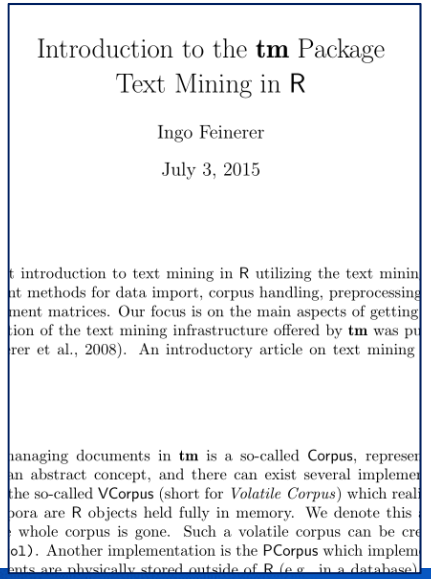
http://www.hmwu.idv.tw

References









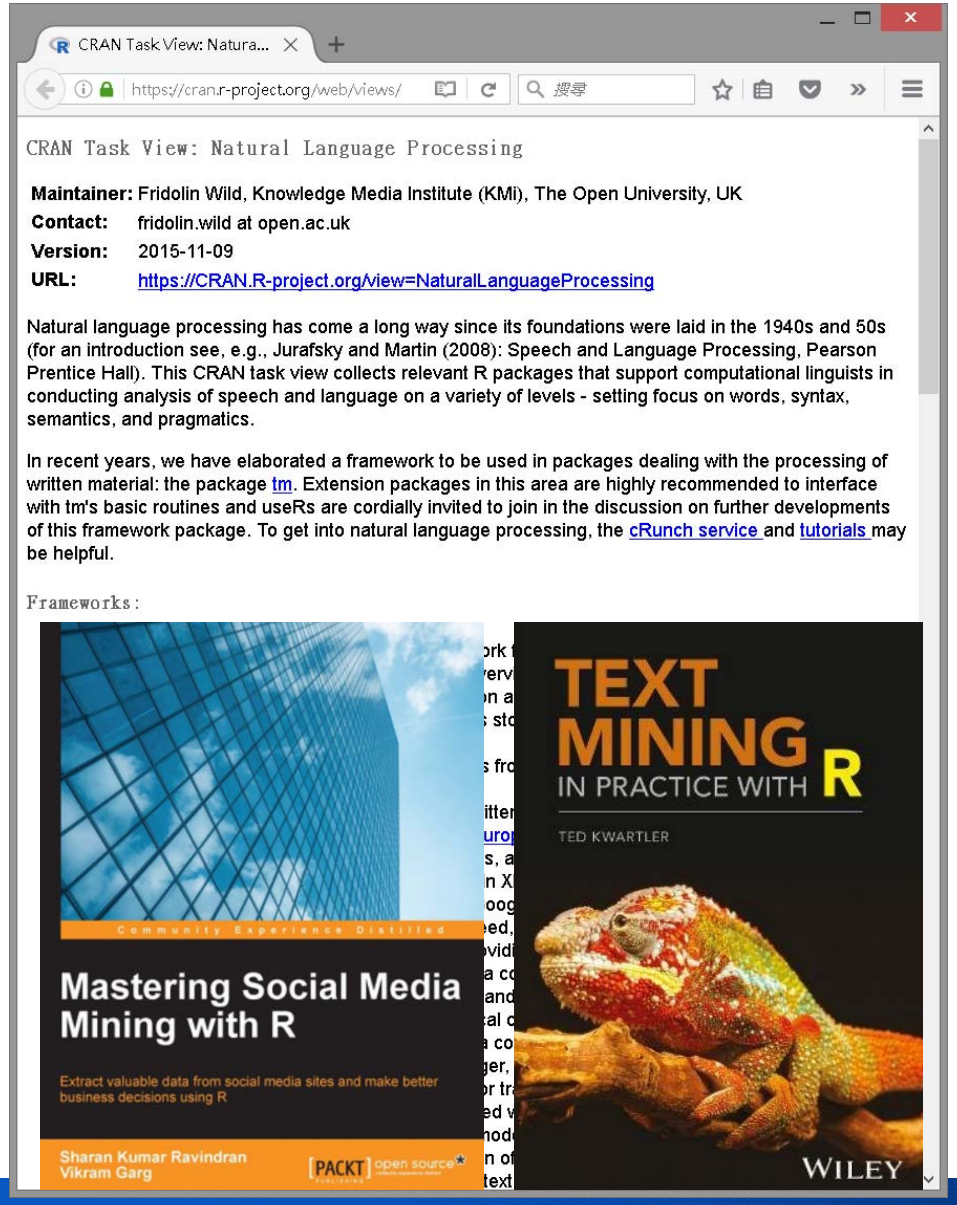
Introduction to the **tm** Package
Text Mining in R

Ingo Feinerer

July 3, 2015

t introduction to text mining in R utilizing the text minin
ut methods for data import, corpus handling, preprocessing
ment matrices. Our focus is on the main aspects of getting
tion of the text mining infrastructure offered by **tm** was pu
rer et al., 2008). An introductory article on text mining

nanaging documents in **tm** is a so-called Corpus, represen
an abstract concept, and there can exist several implemen
the so-called VCorpus (short for *Volatile Corpus*) which reali
pora are R objects held fully in memory. We denote this
whole corpus is gone. Such a volatile corpus can be cre
ol). Another implementation is the PCorpus which implem
nts are physically stored outside of R (e.g., in a database)



CRAN Task View: Natura...

https://cran.r-project.org/web/views/

CRAN Task View: Natural Language Processing

Maintainer: Fridolin Wild, Knowledge Media Institute (KMi), The Open University, UK
Contact: fridolin.wild at open.ac.uk
Version: 2015-11-09
URL: https://CRAN.R-project.org/view=NaturalLanguageProcessing

Natural language processing has come a long way since its foundations were laid in the 1940s and 50s (for an introduction see, e.g., Jurafsky and Martin (2008): Speech and Language Processing, Pearson Prentice Hall). This CRAN task view collects relevant R packages that support computational linguists in conducting analysis of speech and language on a variety of levels - setting focus on words, syntax, semantics, and pragmatics.

In recent years, we have elaborated a framework to be used in packages dealing with the processing of written material: the package tm. Extension packages in this area are highly recommended to interface with tm's basic routines and useRs are cordially invited to join in the discussion on further developments of this framework package. To get into natural language processing, the cRunch service and tutorials may be helpful.

Frameworks:

Community Experience Distilled
Mastering Social Media Mining with R
Extract valuable data from social media sites and make better business decisions using R
Sharan Kumar Ravindran
Vikram Garg
PACKT open source

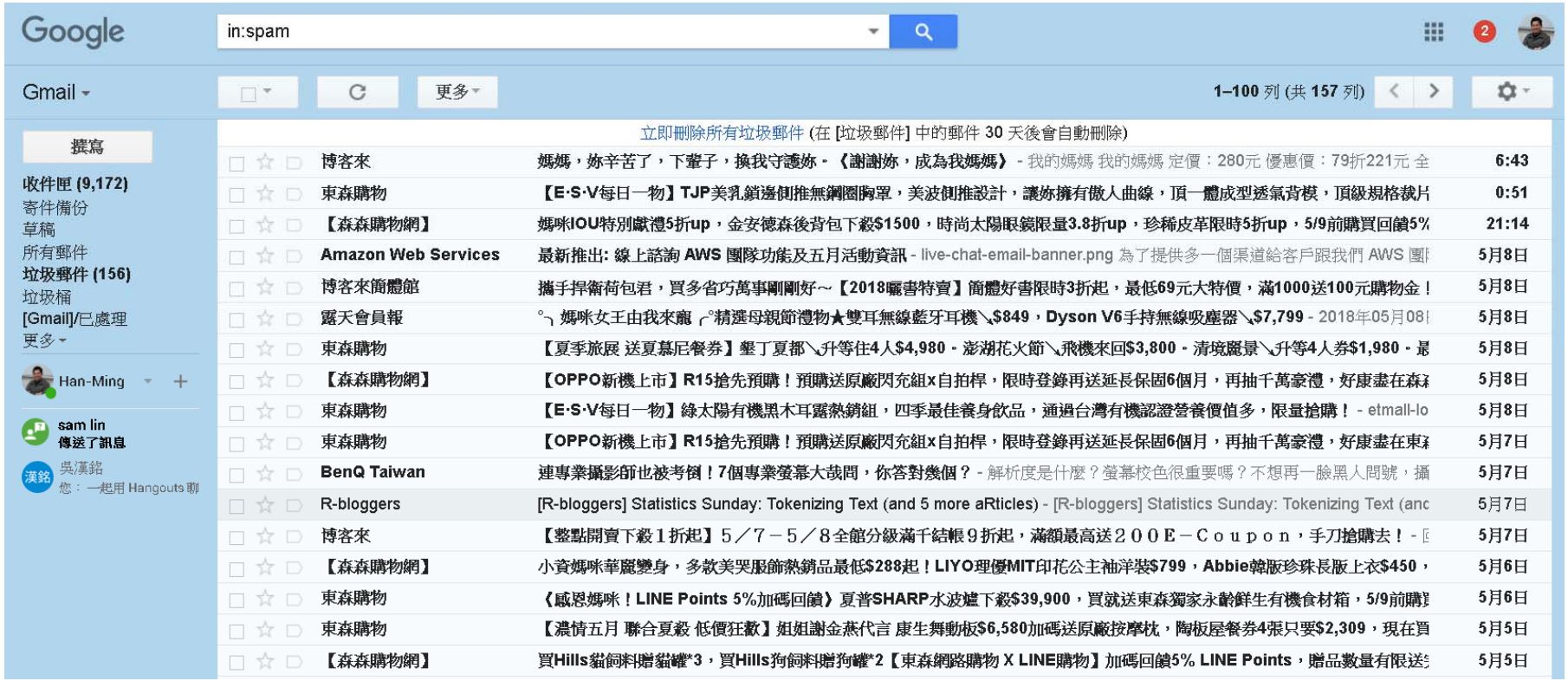
TEXT MINING IN PRACTICE WITH R
TED KWARTLER
WILEY

文字探勘

- 文字探勘: 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潛在資訊

- 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
命名實體辨識
情感分析
文件摘要

N元語法(N-gram)

- 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
- 語句的表示方法如下表。
- 單位可以是字或詞
- 可估計語句的機率，用在不同應用上。

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, …

```
> # R ngram packages
> install.packages("ngram")
> library(ngram)
> words <- "to be or not to be"
> ng <- ngram(words, n=3)
```

```
> get.phrasetable(ng)
      ngrams freq prop
1 not to be     1 0.25
2 be or not     1 0.25
3  to be or     1 0.25
4 or not to     1 0.25
```

需以空白隔開詞彙
不適用於中文

文字處理原則

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

中文處理: 斷詞

- **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，例如: “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**:
  - 中文詞與詞會連在一起，電腦無法直接理解，例如: "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

jiebaR: "結巴"中文分詞套件

- "結巴"中文分詞的R語言版本，支援最大概率法(Maximum Probability)，隱式瑪律科夫模型(Hidden Markov Model)，索引模型(QuerySegment)，混合模型(MixSegment)，共四種分詞模式，同時有詞性標注，關鍵字提取，文本Simhash相似度比較等功能。專案使用了Rcpp和CppJieba進行開發。

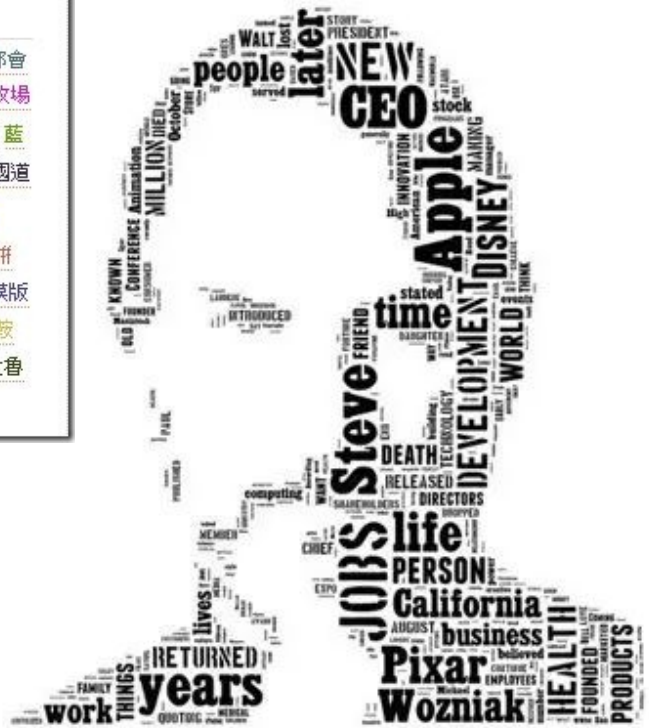
https://www.r-project.org/nosvn/pandoc/jiebaR.html

```
> # install.packages("jiebaR")
> library(jiebaR)
> words <- "大家一起來學文字探勘及文字視覺化"
> tagger <- worker()
> tagger <= words
 [1] "大家" "一"   "起來" "學文" "字"   "探勘" "及"   "文字" "視覺" "化"
> # 標記詞性
> tagger <- worker("tag")
> tagger <= words
     n      m      v      n      n      v      c      n      n      n
"大家"   "一" "起來" "學文"   "字" "探勘"   "及" "文字" "視覺"   "化"
```

文字和文件視覺化

- 文件視覺化的應用: 標籤雲、資訊文字圖等等。
  - 標籤雲或文字雲是**關鍵詞**的視覺化描述，
  - 照片分享社區Flickr是首個使用標籤雲的知名網站，其標籤雲由網站共同創立者、界面互動設計師斯圖爾特·巴特菲爾德（Stewart Butterfield）設計創造(2006)。

Figure 2-18. Tag cloud with random sort and hot/cold coloring



為什麼要文字資料視覺化

- **文字視覺化的作用有以下四點:**
  - 理解 – 理解主旨
  - 組織 – 組織、分類資訊
  - 比較 – 對比文件資訊
  - 關聯 – 關聯文字的 pattern 和其他資訊

- **文字資訊的層級:**
  - 詞彙級
  - 語法級
  - 語義級

文字資訊的層級:詞彙級

- 詞彙級(Lexical Level): 資訊指從一連串的文字中分析出的語義單元資訊。
  - 語義單元 (Token): 由一個或多個字元組成的詞元，是文字資訊的最小單元。
  - 語義單元通常透過基於規則分割文字的分詞技術。
- 詞彙級可分析的資訊包含文字有關的字、詞、子句，以及它們在文章內的分佈統計、詞根詞位等相關資訊。
- 常見的之「關鍵字」即是詞彙等級。

文字資訊的層級: 語法級

- 語法級 (Syntactic Level)資訊指基於文字的語言結構對詞彙級的語義單元進一步分析和解釋而分析得到的資訊。

- 語義單元的語法屬性屬於語法級資訊，例如: 詞性、單複數、詞與詞之間的相關性，以及地點、時間、日期、人名等實體資訊。
- 這些屬性可以透過語法分析器識別。

文字資訊的層級: 語義級

- 語義級(Semantic Level) 資訊是研究文字整體所表達的語義內容資訊和語義關係。
- 包含深入分析詞彙級和語法級所分析的知識在文字中的含義。
- 也包括作者透過文字所傳達的資訊。

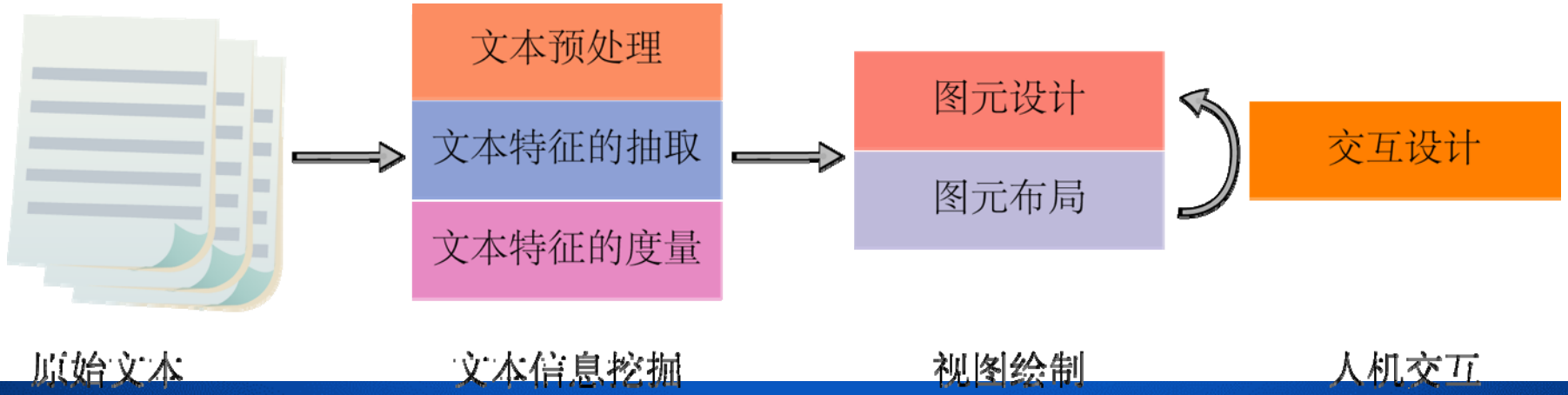
文字視覺化的研究內容

- 文字文件類型: 單文字、文件集合、時序文字資訊。
- 文字視覺化可以幫助使用者快速地了解一個文件的內容、特徵等資訊。
- 視覺化的分類: 文字內容視覺化，文字關係視覺化、文字多層面資訊視覺化。

文字視覺化的流程

- **文字資訊採擷**
  - 文字資料的前置處理: 過濾無效資料。
  - 文字特徵的取出:
    - 詞彙級: 用各類分詞演算法，提取文字的關鍵詞、詞頻分佈。
    - 語法級: 用一些句法分析演算法，提取文字實體資訊。
    - 語義級: 用主題抽取演算法，提取文字的主題。
  - 文字特徵的度量:在多種環境或多個資料來源所抽取的文字特徵進行相似性分析、文字分類等。
- **視圖繪製**
- **人機互動**

SOURCE: https://op46.com/#/technique/7a2ace8a539b7270be54bf9cfcc93071.html



文字資訊分析基礎:
(1) 分詞技術和詞幹分析

- 分析基礎: (1) 分詞技術和詞幹分析、(2) 向量空間模型、
  (3) 主題取出

- **分詞技術和詞幹分析**
  - **分詞(tokenization)**: 將一段文字劃分為多個詞項，剔除停詞，從文字中分析出有意義的詞項。
  - **詞幹分析(stemming)**: 去除詞綴獲得詞根，獲得單字最一般寫法的技術，避免同一個詞的不同表現形式對文字分析帶來干擾。
  - Martin Luther King, Jr. "I Have a Dream", delivered 28 August 1963, at the Lincoln Memorial, Washington D.C.

  "I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, that all men are created equal."" (20/34)
  - (x) a, the, that
  - men => man, truths => truth

文字資訊分析基礎:
(2) 向量空間模型 (vector space model)

- 利用向量符號對文字進行度量的代數模型。
  - **詞袋模型** (bag-of-words model), 用來分析詞彙級文字資訊。
- 過濾掉無效詞後，詞袋模型將一個文件的內容歸納為在由**關鍵片語**組成的集合上的**加權分佈向量**。
- 基於詞袋模型計算的一維詞頻向量中，每個維度代表一個單字; 每個維度的值等於單字在文字中出現的統計資訊，可引申為重要性，單字間沒有順序關係。
- 詞袋模型沒有考慮語法、詞序資訊，較容易直觀。
- 詞袋模型的詞頻向量提供更高層文字分析的基礎。

詞袋模型的詞頻向量

I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, that all men are created equal."

I have a dream that one day on the red hills of Georgia, the sons of former slaves and the sons of former slave owners will be able to sit down together at the table of brotherhood.

I have a dream that one day even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression, will be transformed into an oasis of freedom and justice.

I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character.

- 142 words => 78 words
- 以下為此段文件之詞頻向量的一部份:

詞	I	dream	color	skin	nation	slave	injustice	owner
詞頻	4	4	1	1	2	2	1	1

文字的相似性度量: 餘弦相似性

- 採用詞項-文件矩陣來建構多個文件的數學模型。
- 度量文字語義的相似度: **夾角餘弦值**
- 餘弦相似性通過測量兩個向量的夾角的餘弦值來度量它們之間的相似性。
  - 0度角的餘弦值是1 (兩個向量有相同的指向時)。
  - 兩個向量夾角為90°時，餘弦相似度的值為0；
  - 兩個向量指向完全相反的方向時，餘弦相似度的值為-1。
  - 其他任何角度的餘弦值都不大於1；並且其最小值是-1。
  - 兩個向量之間的角度的餘弦值可確定兩個向量是否大致指向相同的方向。

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

，這裡的A_i和B_i分別代表向量A和B的各分量。

範例: 文件餘弦相似性

- **準備文件:**
  - 句子A：我喜歡看電視，不喜歡看電影。
  - 句子B：我不喜歡看電視，也不喜歡看電影。
- **第一步：分詞**
  - 句子A：我/喜歡/看/電視，不/喜歡/看/電影。
  - 句子B：我/不/喜歡/看/電視，也/不/喜歡/看/電影。
- **第二步：列出所有的詞**
  - 我，喜歡，看，電視，電影，不，也
- **第三步：計算詞頻**
  - 句子A：我 1，喜歡 2，看 2，電視 1，電影 1，不 1，也 0
  - 句子B：我 1，喜歡 2，看 2，電視 1，電影 1，不 2，也 1
- **第四步：寫出詞頻向量**
  - 句子A：[1, 2, 2, 1, 1, 1, 0]
  - 句子B：[1, 2, 2, 1, 1, 2, 1]
- **第五步：計算餘弦值**

$$\cos\theta = \frac{1\times1+2\times2+2\times2+1\times1+1\times1+1\times2+0\times1}{\sqrt{1^2+2^2+2^2+1^2+1^2+1^2+0^2}\times\sqrt{1^2+2^2+2^2+1^2+1^2+2^2+1^2}}$$

$$= \frac{13}{\sqrt{12}\times\sqrt{16}}$$

$$= 0.938$$

餘弦值越接近1，就表明夾角越接近0度，也就是兩個向量越相似。

原文網址：https://kknews.cc/tech/ovayjvq.html

Term Frequency-Inverse Document Frequency TF-IDF

- 對向量空間模型，最常用的**多詞權數分配模型**是TF-IDF。
- TF-IDF用以評估一個單字或字對於一個文件集合或一個語料庫中的其中一份文件的重要程度。
- ***概念***: 字詞對於某個文件的重要性隨著它在這個文件中出現的次數成正相關(正比)增加，但同時會隨著它在文件集中出現的頻率而負相關(反比)下降。
- 在一份給定的檔案裡，詞頻 (term frequency, tf)指的是某一個給定的詞語在該檔案中出現的頻率。

TF-IDF 的計算

- 對於在某一特定檔案裡的詞語t_i，它的重要性可表示為:

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

$n_{i,j}$: 是該詞在檔案d_j中的出現次數，而分母則是在檔案d_j中所有字詞的出現次數之和。

- 逆向檔案頻率（inverse document frequency，idf）是一個詞語普遍重要性的度量。某一特定詞語的idf，可以由總檔案數目除以包含該詞語之檔案的數目，再將得到的商取以10為底的對數得到:

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

- |D|：語料庫中的文件總數
- $|\{j : t_i \in d_j\}|$：包含詞語t_i的文件數目（即$n_{i,j} \neq 0$的文件數目）如果詞語不在資料中，就導致分母為零，因此一般情況下使用 $1 + |\{j : t_i \in d_j\}|$

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

某一特定文件內的高詞語頻率，以及該詞語在整個文件集合中的低文件頻率，可以產生出高權重的tf-idf。因此，tf-idf傾向於過濾掉常見的詞語，保留重要的詞語。

Variants of TF/IDF weight

Variants of term frequency (TF) weight

weighting scheme	TF weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Variants of inverse document frequency (IDF) weight

weighting scheme	IDF weight ($n_t = \lvert\{d \in D : t \in d\}\rvert$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(1 + \frac{N}{n_t}\right)$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

Recommended TF-IDF weighting schemes

weighting scheme	document term weight	query term weight
1	$f_{t,d} \cdot \log \frac{N}{n_t}$	$\left(0.5 + 0.5 \frac{f_{t,q}}{\max_t f_{t,q}}\right) \cdot \log \frac{N}{n_t}$
2	$1 + \log f_{t,d}$	$\log(1 + \frac{N}{n_t})$
3	$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$	$(1 + \log f_{t,q}) \cdot \log \frac{N}{n_t}$

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

文字資訊分析基礎:
(3) 主題取出

文件

主題分佈

主題	分佈
T1	0.6
…	…
Tn	0.05

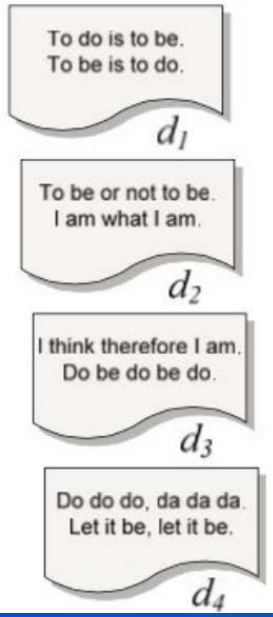
詞頻分佈

詞	word 1	…	word n
分佈	0.45	…	0.002

詞頻分佈

詞	word 1	…	word n
分佈	0.55	…	0.017

- 主題模型是指從**語義等級**描述文字集合內各個文件的語義內容。
- 一個文件的語義內容可描述為多個主題的組合。
- 一個主題有認為是一系列詞的機率分佈或權重分佈。
- 演算法: (1) 基於矩陣分解的非機率模型。(2) 基於貝氏的機率模型。



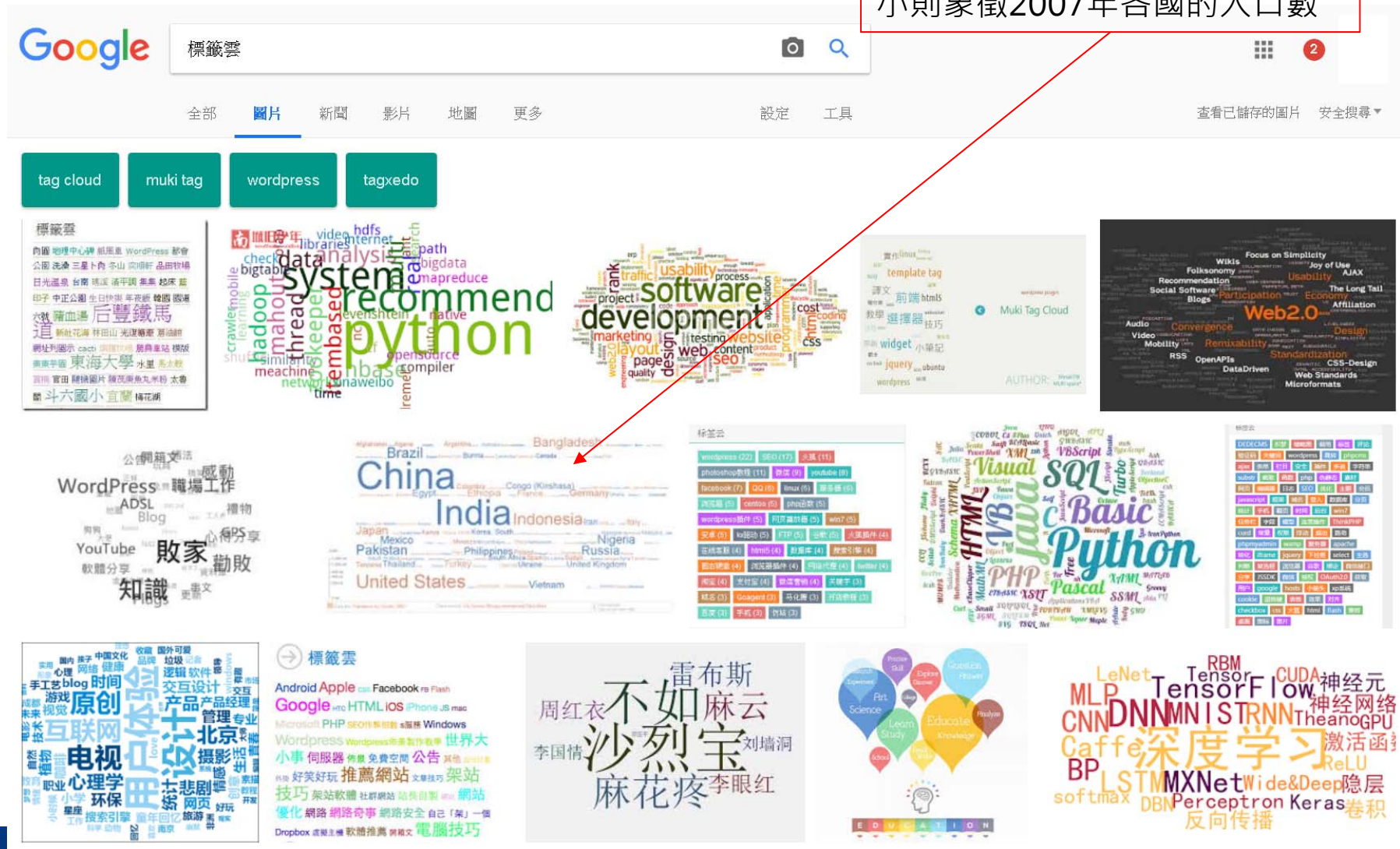
		d_1	d_2	d_3	d_4
1	to	3	2	-	-
2	do	0.830	-	1.073	1.073
3	is	4	-	-	-
4	be	-	-	-	-
5	or	-	2	-	-
6	not	-	2	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	2	-	-
10	think	-	-	2	-
11	therefore	-	-	2	-
12	da	-	-	-	5.170
13	let	-	-	-	4
14	it	-	-	-	4

文件視覺化

- **文字內容視覺化**
  - 基於關鍵字的之文字內容視覺化
  - 時序性的文字內容視覺化
  - 文字特徵的分佈模式視覺化
  - 情感分析視覺化
  - 文件資訊檢索視覺化
  - 軟體視覺化
- **文字關係視覺化**
  - 以圖為基礎的文字關係視覺化
  - 文件集合關係視覺化
- **文字多層面資訊視覺化**

基於關鍵字的之文字內容視覺化: 標籤雲

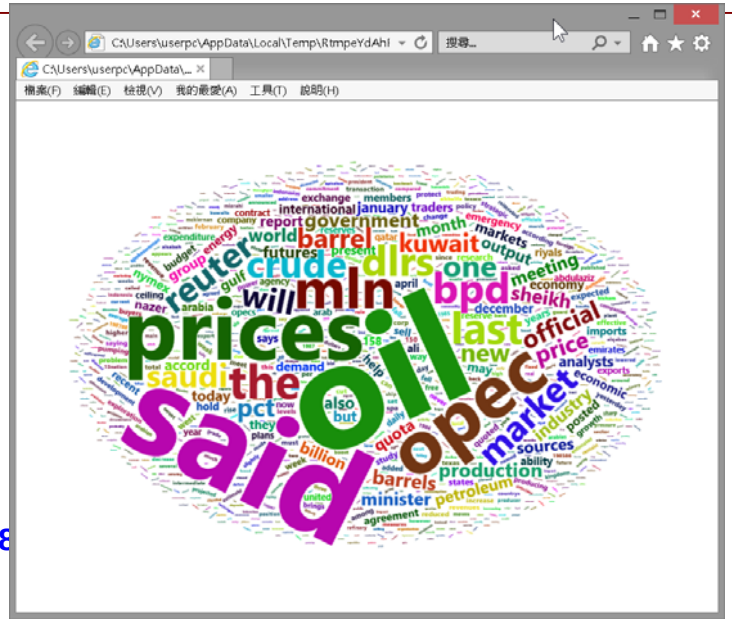
顯示世界各國人口數量的數據雲，顏色用以區分國家，而字體大小則象徵2007年各國的人口數



Wordcloud2

- Wordcloud2 provides an HTML5 interface to wordcloud for data visualization. Timdream's wordcloud2.js is used in this package.
- 繪製出來的結果有交互效果，滑鼠移到一個詞上可以看到詞頻。
- Main R functions:
  - **wordcloud2**: provide traditional wordcloud with HTML5
  - **letterCloud**: provide wordcloud with selected word(letters).

```
> require(devtools)
> install_github("lchiffon/wordcloud2")
> library(wordcloud2)
> head(demoFreq)
         word freq
oil       oil   85
said     said   73
prices prices   48
opec     opec   42
mln       mln   31
the       the   26
> str(demoFreq)
'data.frame':   1011 obs. of  2 variables:
 $ word: Factor w/ 1011 levels "100000","108",..: 613 8
...
 $ freq: num  85 73 48 42 31 26 24 23 23 21 ...
> wordcloud2(data = demoFreq)
> wordcloud2(demoFreq, color = "random-light", backgroundColor = "grey")
> ? wordcloud2
```

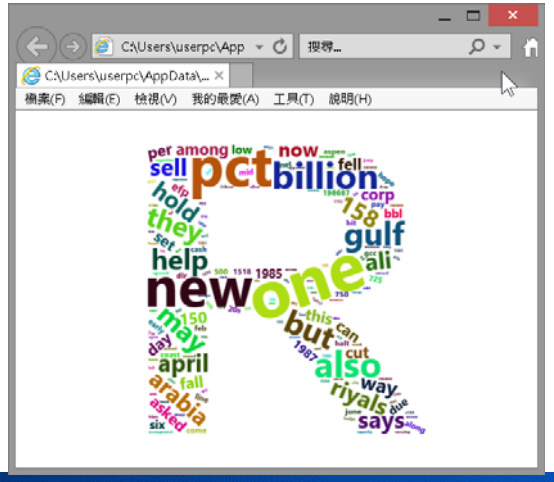

oil
said
prices
opec
mln
the
reuter
crude
dlrs
one
bpd
market
last
official
price
kuwait
barrel
government
production
minister
petroleum
barrels
saudi
pct
billion
quota
sources
industry
analysts
meeting
sheikh
new
will
world
futures
january
traders
month
markets
output
report
international
exchange
members
contract
company
energy
group
budget
expenditure
emergency
december
economy
imports
exports
emirates
economic
posted
ability
agreement
accord
demand
today
hold
year
plans
also
but
recent
arabia
nazer
ceiling
nymex
gulf
says
158
help
ali
may
april
riyals
indonesia
expected
years
members
transaction
contract
company
report
government
month
markets
output
meeting
economy
sheikh
december
official
imports
price
analysts
exports
economic
posted
industry
sources
ability
production
petroleum
minister
barrels
billion
quota
pct
saudi
the
also
but
today
hold
year
plans
they

More examples

```
> # figPath" <- "t.png"
> figPath <- system.file("examples/t.png",package = "wordcloud2")
> wordcloud2(demoFreq, figPath = figPath, size = 1.5, color = "skyblue")
```

```
> letterCloud(demoFreq, word = "R", size = 2)
> letterCloud(demoFreq, word = "NTPU", wordSize = 1)
```

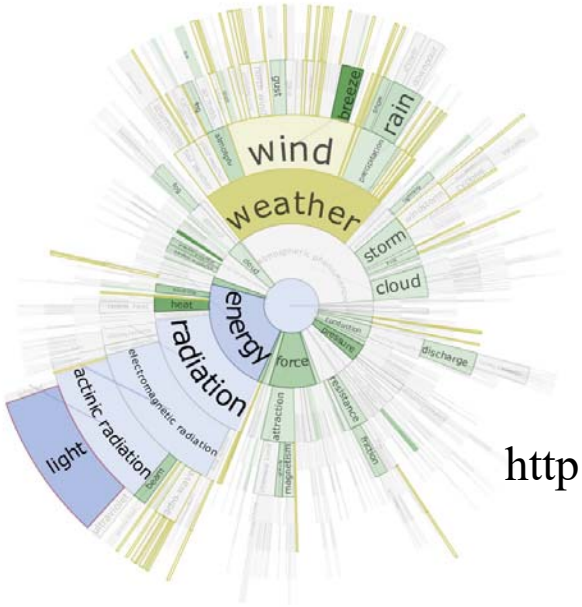
打你的中文名字
試試看



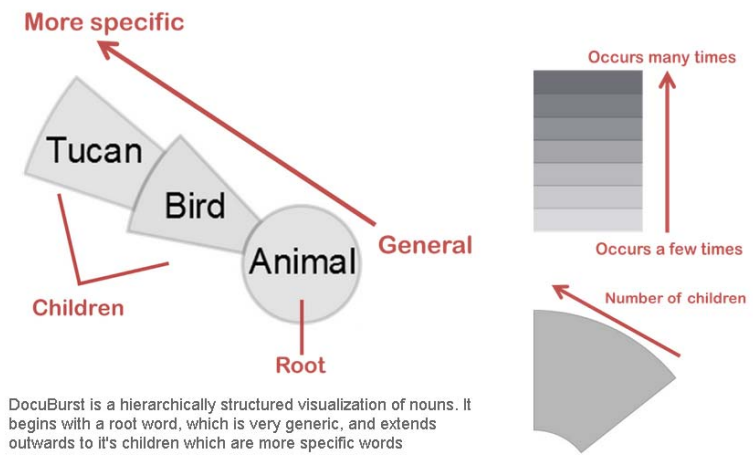
(1) 中文亂碼問題: windows OS，需要Sys.setlocale("LCCTYPE","eng"); 輸入的文字需要是中文 UTF-8。
(2) RStdio問題: 可能不顯示，需要手動刷新或者在流覽器中打開。

基於關鍵字的之文字內容視覺化: 文件散

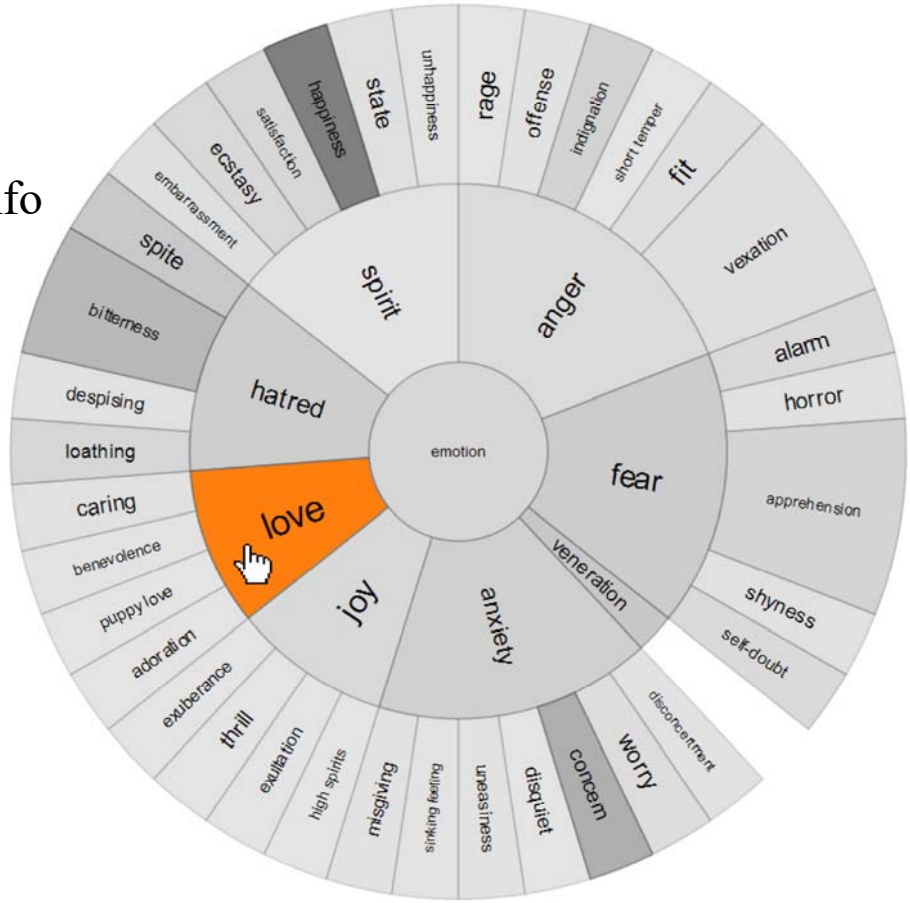
文件散(DocuBurst)也是基於關鍵詞的文字視覺化，不過它還通過徑向佈局體現了詞的**語義等級**。外層的詞是內層詞的下義詞，顏色飽和度的深淺用來體現詞頻的高低。



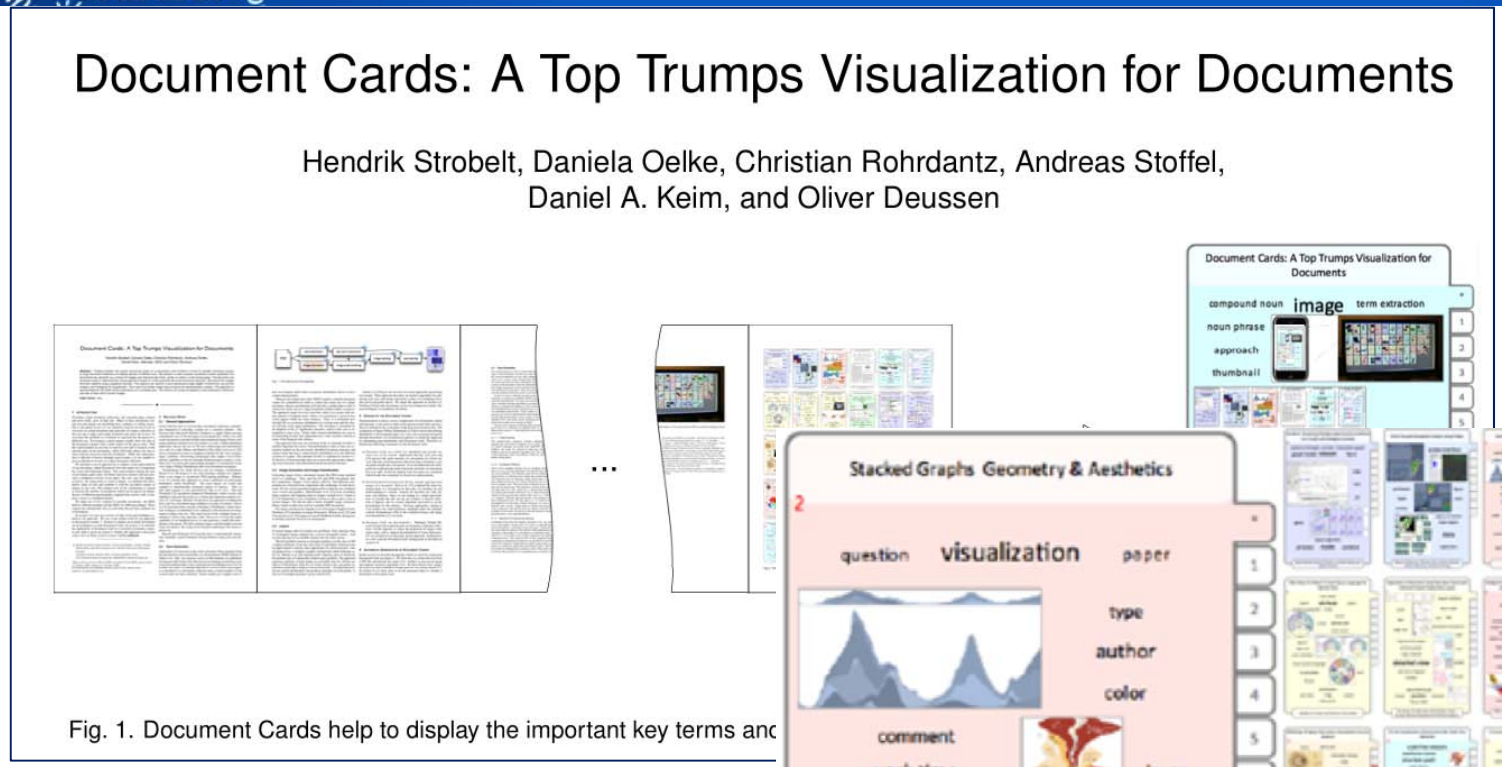
http://www.infovis.info

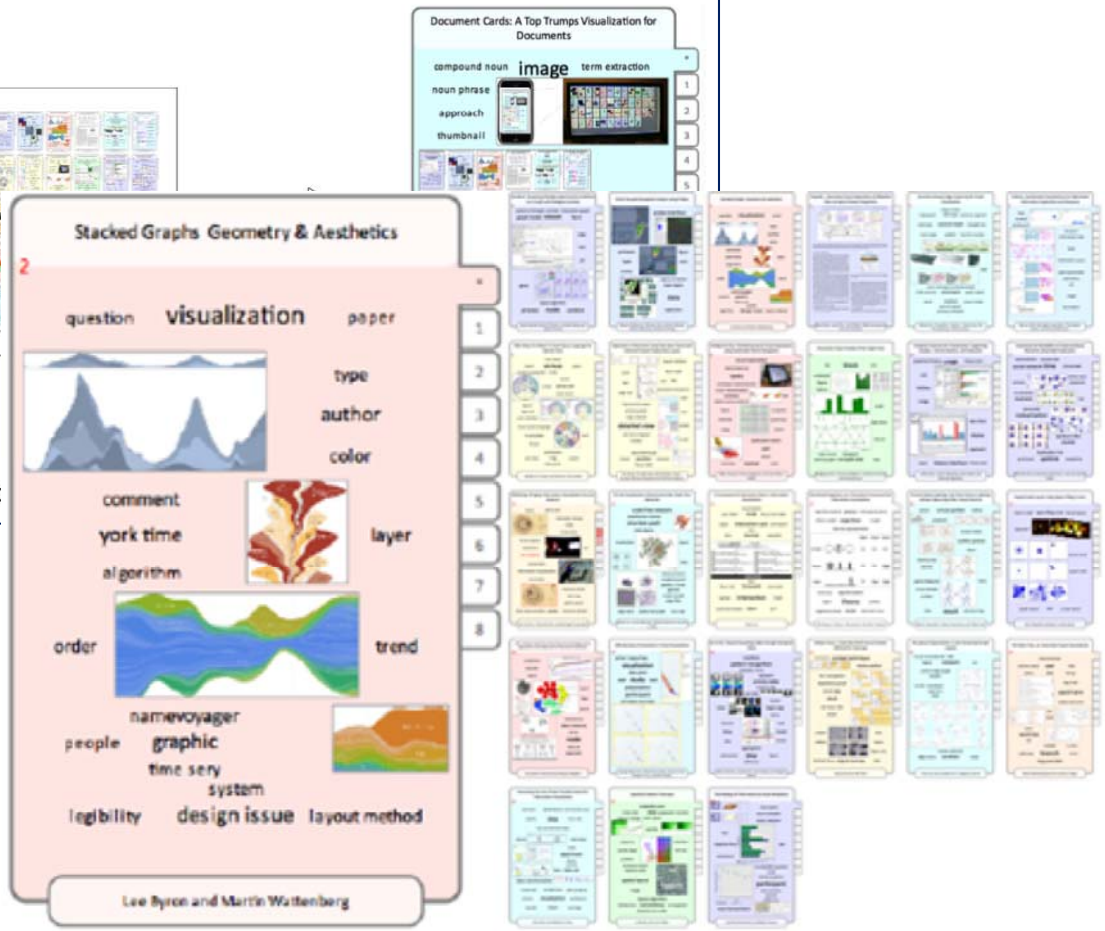


http://vialab.science.uoit.ca/docuburst/help.php



基於關鍵字的之文字內容視覺化: 文件卡





文件卡片(Document Cards)是結合了文件中的**關鍵詞**和**關鍵圖片**進行視覺化，佈局在一張小卡片中。其中的關鍵圖片是指採用智慧演算法抽取並根據顏色分類後的代表性圖片。

時序性的文字內容視覺化: 主題河流

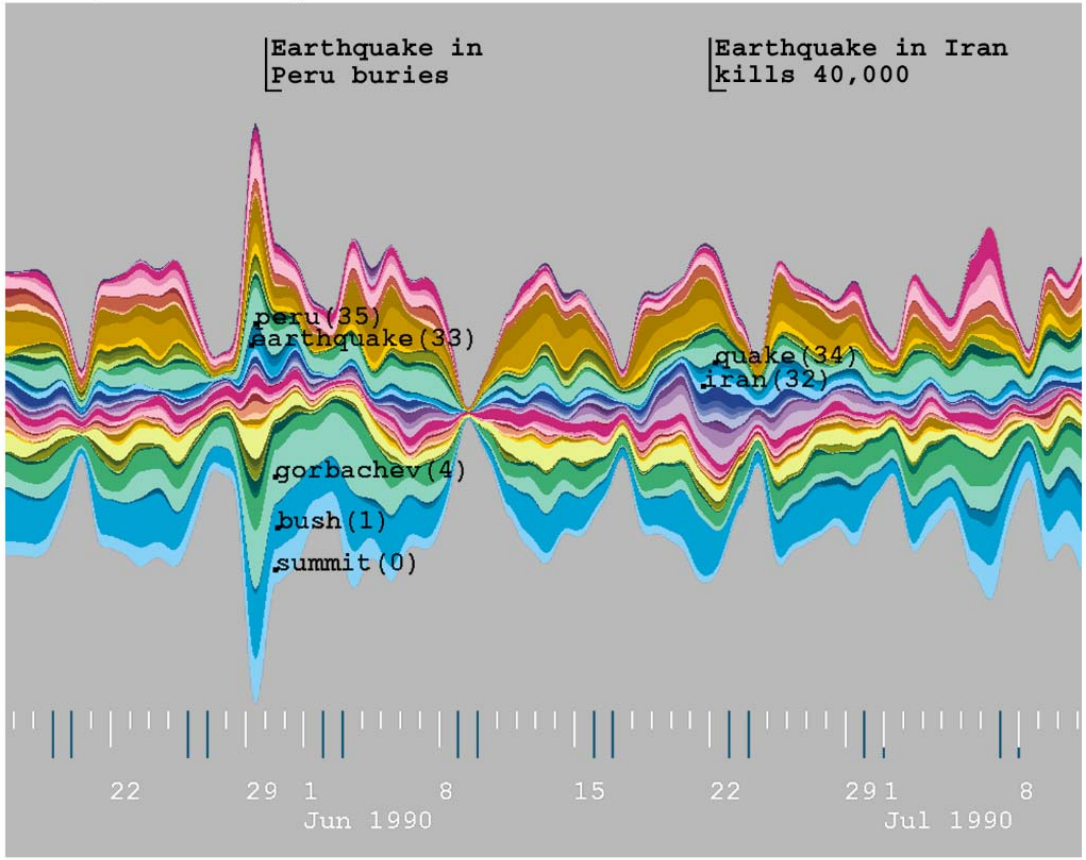


Figure 4: ThemeRiver™ of AP data from June - July 1990 identifies very different events from those revealed immediately afterwards (Figure 3).

主題河流(ThemeRiver)是一種經典的時序文字視覺化方法。

橫軸表示時間，每一條不同顏色線條可視作一條河流，而每條河流則表示一個主題，河流的寬度代表其在當前時間點上的一個度量(如主題的強度)。

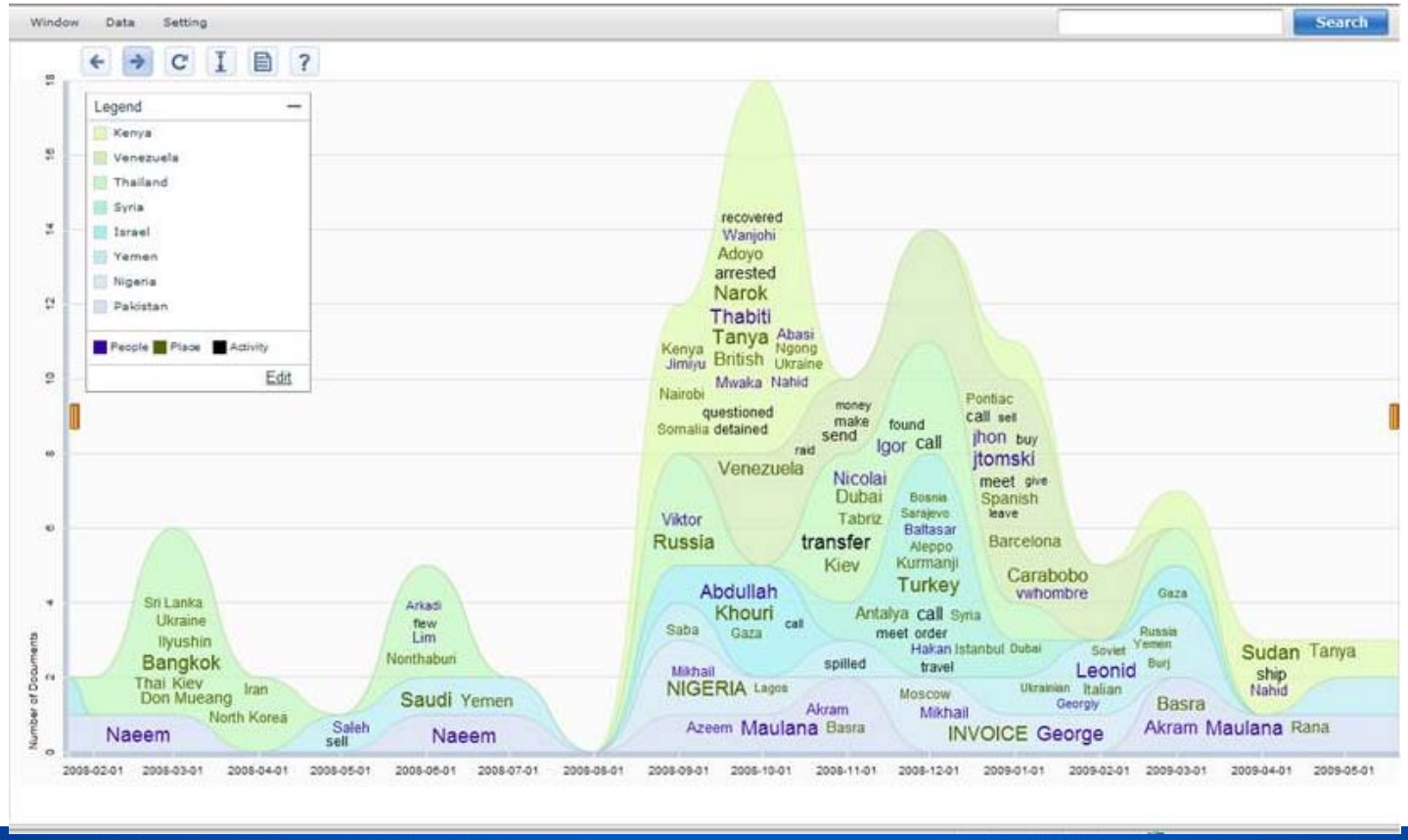
可宏觀上看出多個主題的發展變化，又能看出在特定時間點上主題的分佈。

Susan L. Havre, Elizabeth G. Hetzler, Lucy T. Nowell，ThemeRiver: Visualizing Theme Changes over Time Published 2000 in INFOVIS

時序資料是指具有時間或順序特性的文字，例如一篇小說故事情節的變化，或一個新聞事件隨時間的演化。

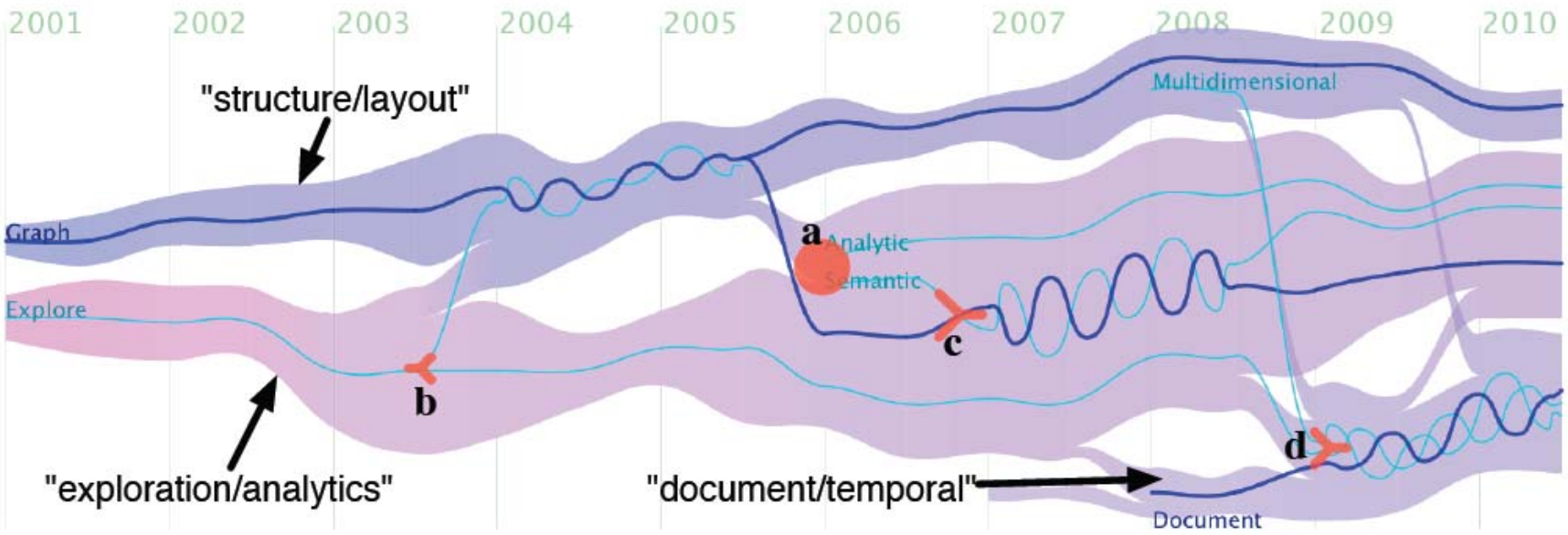
時序性的文字內容視覺化: TIARA

- TIARA (Text Insight via Automated Responsive Analytics)
- TIARA結合了標籤雲，通過主題分析技術(latent dirichlet allocation，LDA)，將文字關鍵詞根據時間點放置在每條色帶上，並用詞的大小來表示關鍵詞在該時刻出現的頻率。
- TIARA可以幫助使用者快速分析文字具體內容隨時間變化的規律，而不是僅僅一個度量帶變化。



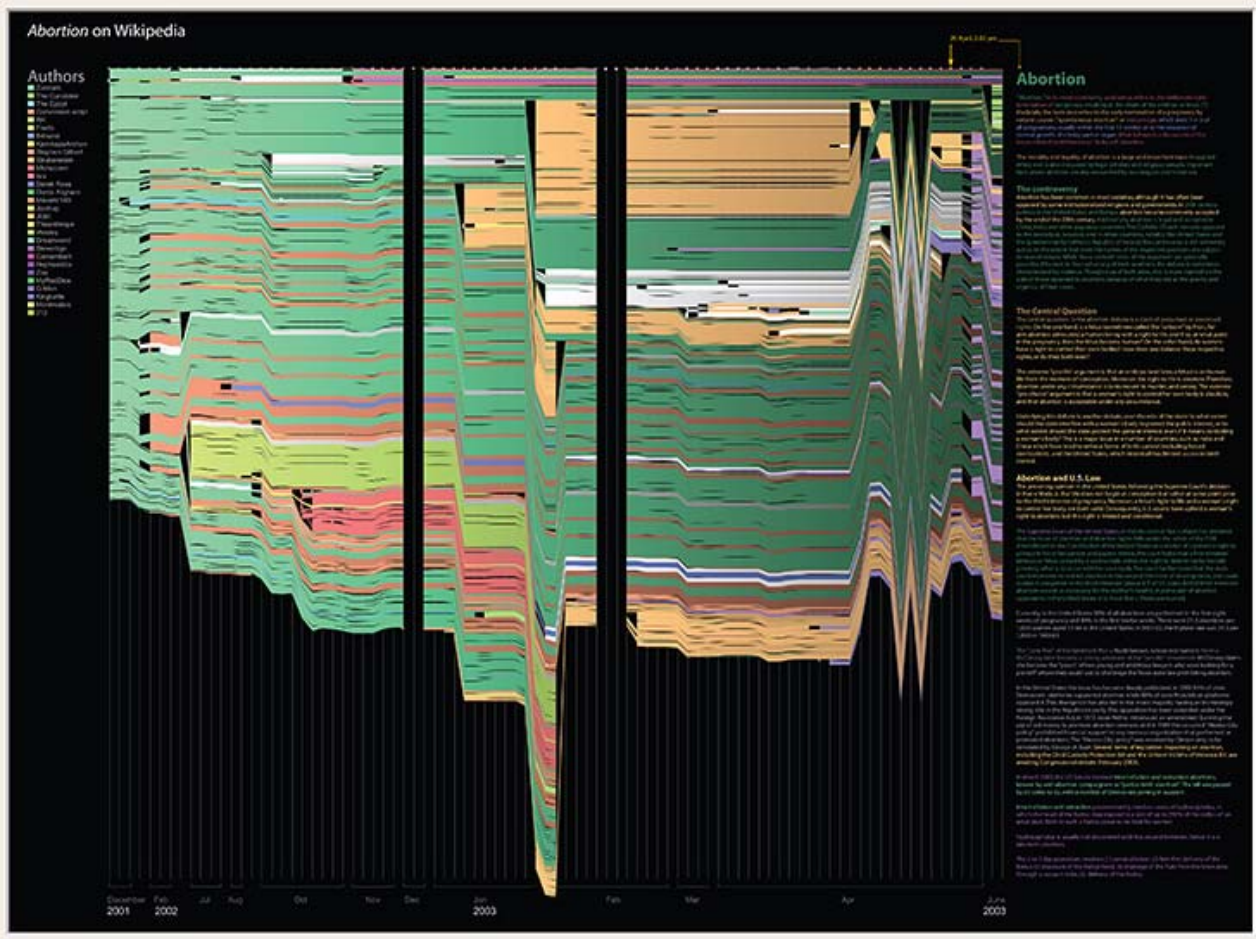
時序性的文字內容視覺化: TextFlow

- TextFlow 是 ThemeRiver 的一種拓展，它不僅表達了主題的變化，還表達了各個主題隨著時間的分裂與合併。如某個主題在某個時間抽成了兩個主題，或多個主題在某個時間合併成了一個主題。



http://www.cad.zju.edu.cn/home/vagblog/?p=360

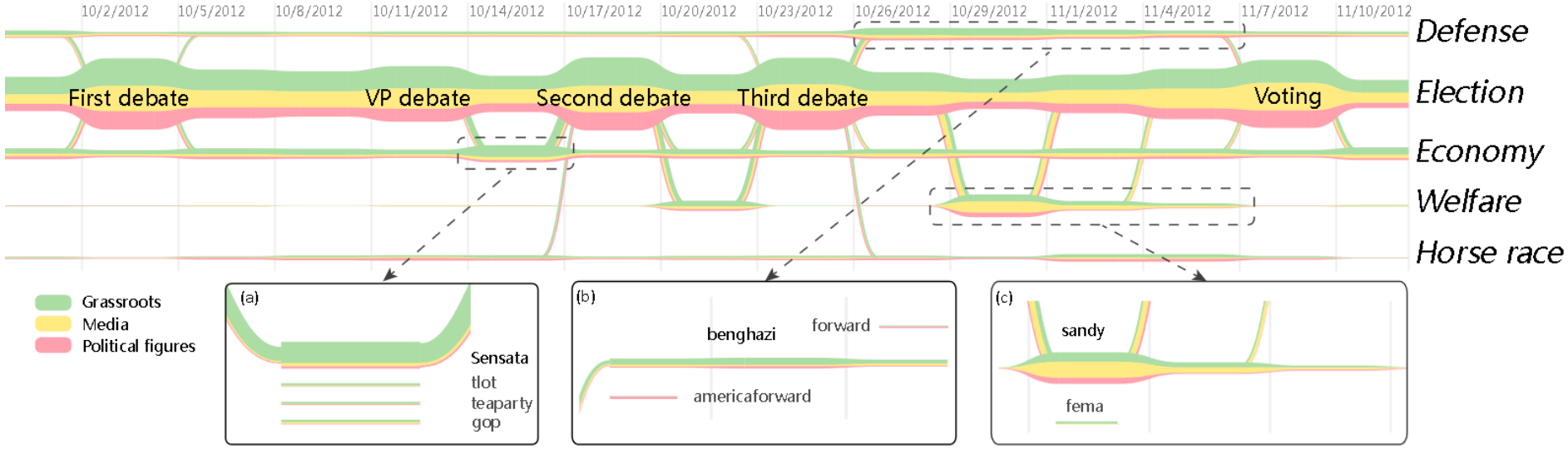
時序性的文字內容視覺化: 歷史流



- HistoryFlow 則主要研究文件內容隨時間的變化。圖中以維基百科一篇詞條的更新為例，縱軸表示文章的版本更新時間點。
- 每一種顏色代表一個作者，在同一個時間軸上色塊代表相應的作者所貢獻的文字塊，並且色塊的位置代表該文字塊在文章中的順序。
- 縱覽全圖就可以輕易看出文章的修改。

http://scimaps.org/mapdetail/history_flow_visuali_56

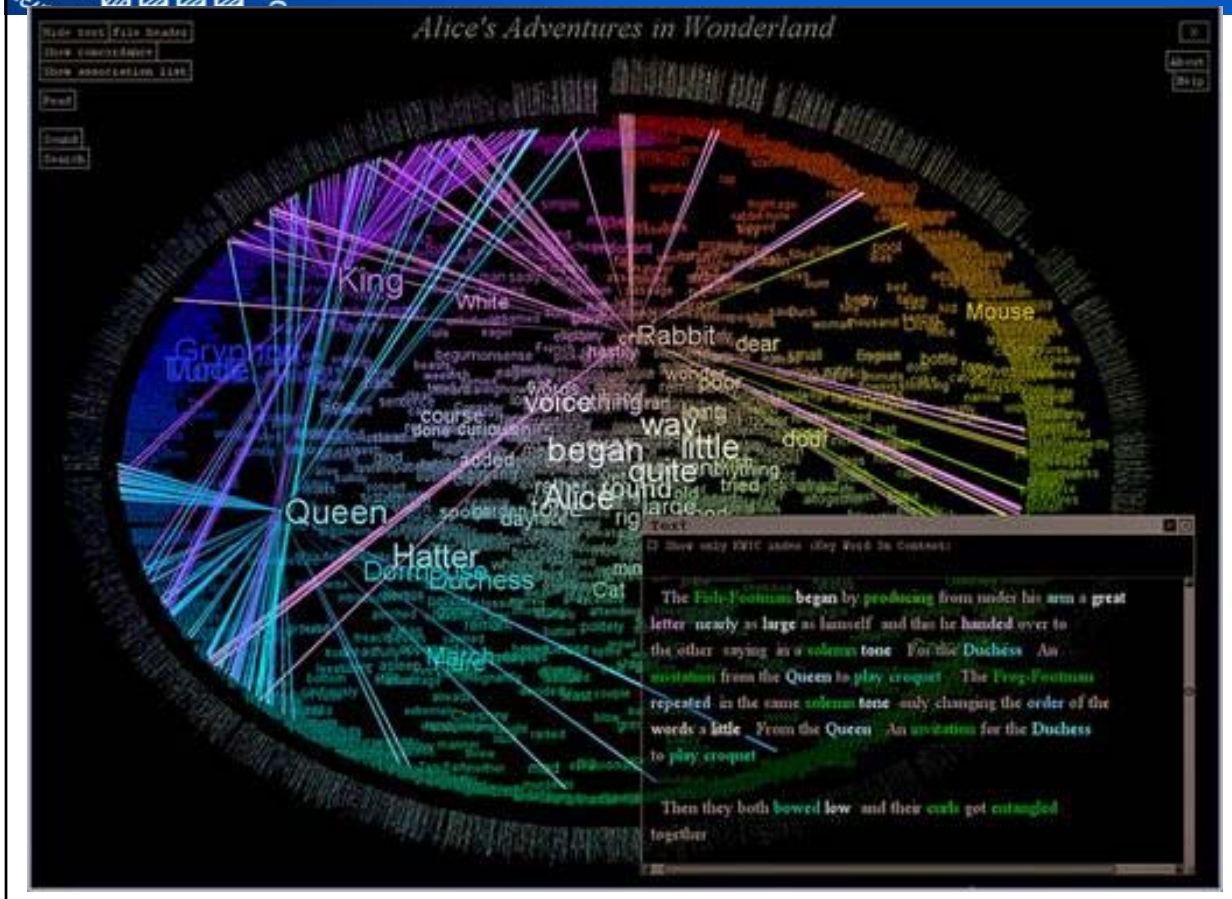
時序性的文字內容視覺化: StoryFlow



Visualization of the 2012 US presidential election. It contains **89,174,308 tweets, 900 opinion leaders, and 63 time frames**. The opinion leaders are organized by a 2-level topic hierarchy and are bundled together by the LOD technique. Each color represents an opinion leader group. Three interesting attention transition patterns related to (a) the Benghazi (班加西(利比亞城市)) issue of the Obama administration, (b) the Sensata scandal of Romney (米特·羅姆尼(前任麻薩諸塞州州長)), and (c) hurricane Sandy (颶風珊迪) have been identified.

- 電影或小說經常說到的時間線、或劇情線等，都能用 StoryFlow 來表示，它通過層次渲染的方式，生成一個 StoryLine 佈局。每條線是一條人物線，當兩人在劇情中有某種聯絡（同時出場或其他交集）時會在圖中相交，橫軸表示時間。
- StoryFlow 還允許使用者實時互動，包括捆綁操作、刪除、移動以及直線化等等。
- 視訊演示：https://www.youtube.com/watch?v=yoq82mC30Iw

文字特徵的分佈模式視覺化: 文字弧



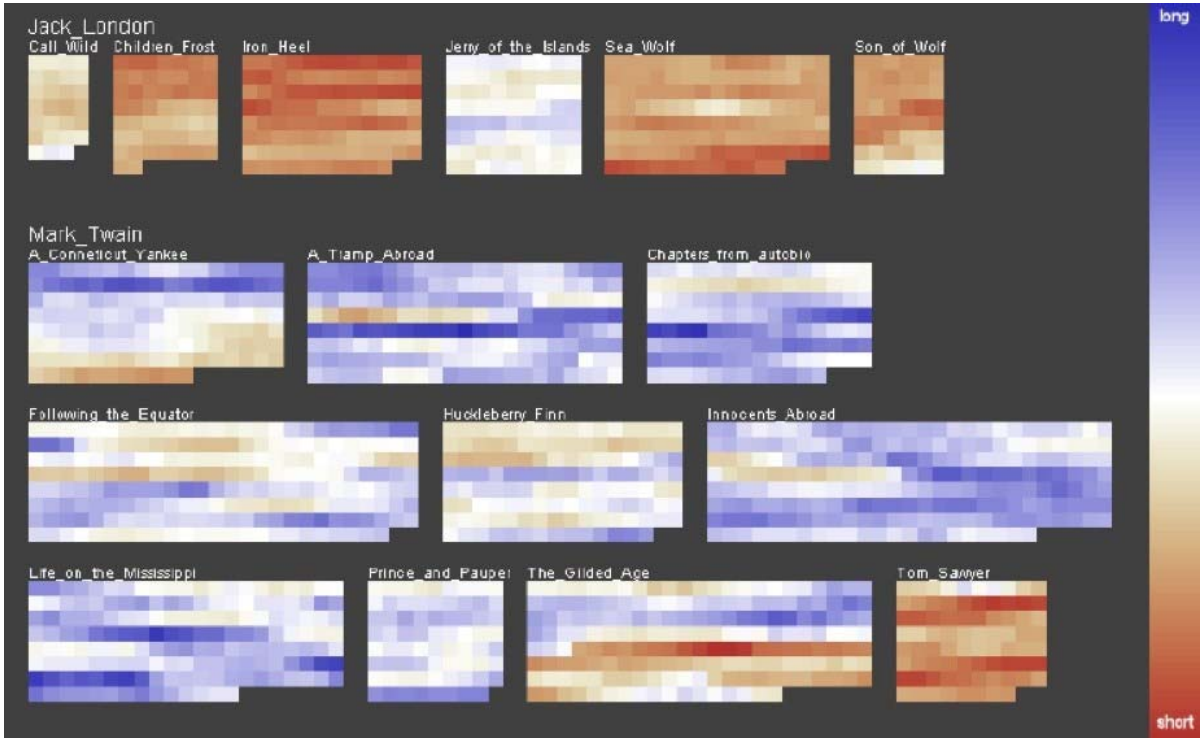
Alice in Wonderland

TextArc 用來視覺化一個文件中的詞頻和詞的分佈情況。

- 整個文件用一條螺線表示，文件的句子按文字的組織順序佈局在螺線上，螺線包圍著的是文件中出現的單詞，每個單詞的位置由其在文字中的頻率和出現位置決定，飽和度用來對映詞頻。
- 全域性出現頻率越高的詞越靠近中心，而區域性出現頻率越高的詞越靠近其相應的螺線區網域。選中某個單詞後，自動用射線關聯到它在文中出現的位置。

http://csis.pace.edu/digitalgallery/Marginalia/index.html

文字特徵的分佈模式視覺化: 文獻指紋



Keim D A, Oelke D. Literature fingerprinting: A new method for visual literary analysis[C]//Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on. IEEE, 2007: 115-122.

- 文獻指紋(Literature Fingerprinting)可用於呈現全文特徵分佈。
- 一個畫素區塊代表一段文字，一組畫素區塊代表一本書。
- 顏色對映的是文字特徵(寫作度量特徵值)，圖中是句子的平均長度。從圖中明顯看出兩人的寫作風格迥異。

http://www.idvbook.com/teaching-aid/figures/chapter-9/

情感分析視覺化

- 情感分析是指從文字中挖掘出心情、喜好、感覺等主觀資訊。
- 可了解人們對於一個事件的觀點或情感的發展。
- 下圖是基於矩陣檢視的客戶反饋資訊的視覺化，其中的行是指文字(使用者觀點)的載體，列是使用者的評價，顏色表達的是使用者評價的傾向程度，紅色代表消極，藍色代表積極，每個方格內的小格子代表使用者評價的人數，評價人數越多小格子越大。
- 文件資訊檢索視覺化:TileBar, Sparkler
- 軟體視覺化: SeeSoft, Code_Swarm,

Oelke D, Hao M, Rohrdantz C, et al. Visual opinion analysis of customer feedback data, Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on. IEEE, 2009: 187-194.

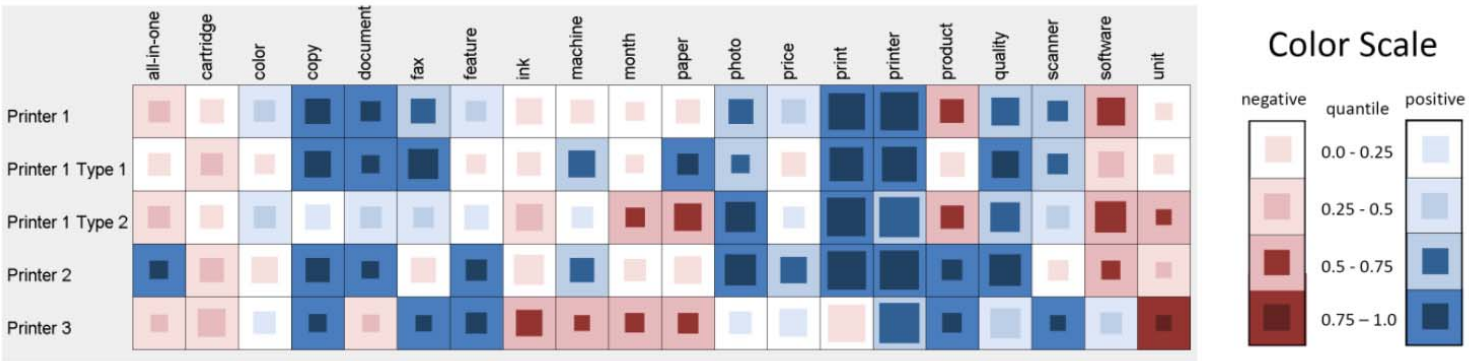
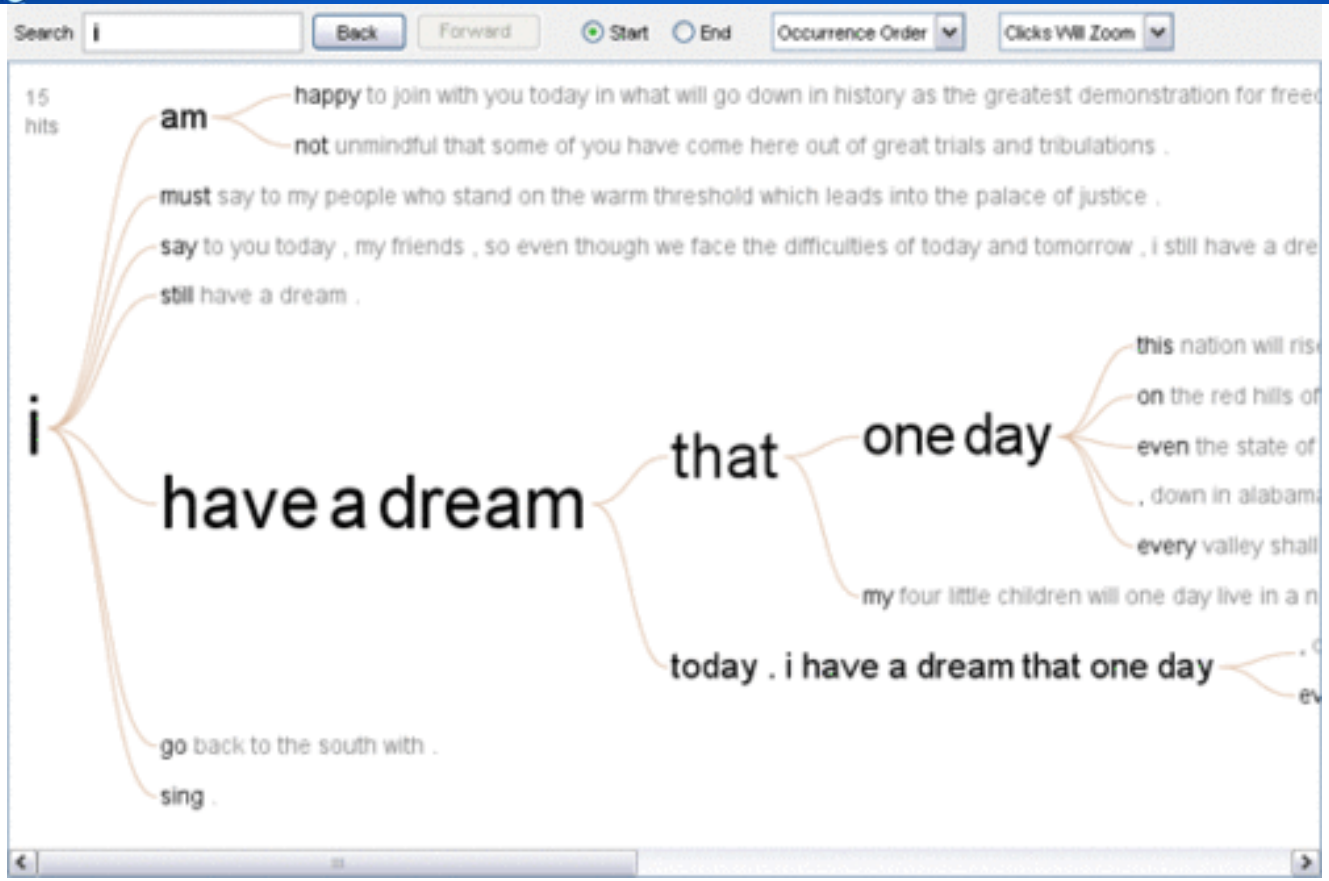


Figure 3. Summary Report of printers: Each row shows the attribute performances of a specific printer. Blue color represents comparatively positive user opinions and red color comparatively negative ones (see color scale). The size of an inner rectangle indicates the amount of customers that commented on an attribute. The larger the rectangle the more comments have been provided by the customers.

以圖為基礎的文字關係視覺化: 單字樹 (word tree)



- 單詞樹(Word Tree)， 可把文字中的句子按樹形結構佈局，可看出一個單詞在文字中出現的頻率和單詞前後的聯絡。

Wattenberg M, Viégas F B. The word tree, an interactive visual concordance. IEEE transactions on visualization and computer graphics, 2008, 14(6).

http://roskylegaled.com/blog/post/martin-luther-king-jr-s-i-have-a-dream-speech-as-a/

以圖為基礎的文字關係視覺化: 子句網路 (Phrase Nets)

1174 IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 15, NO. 6, NOVEMBER/DECEMBER 2009

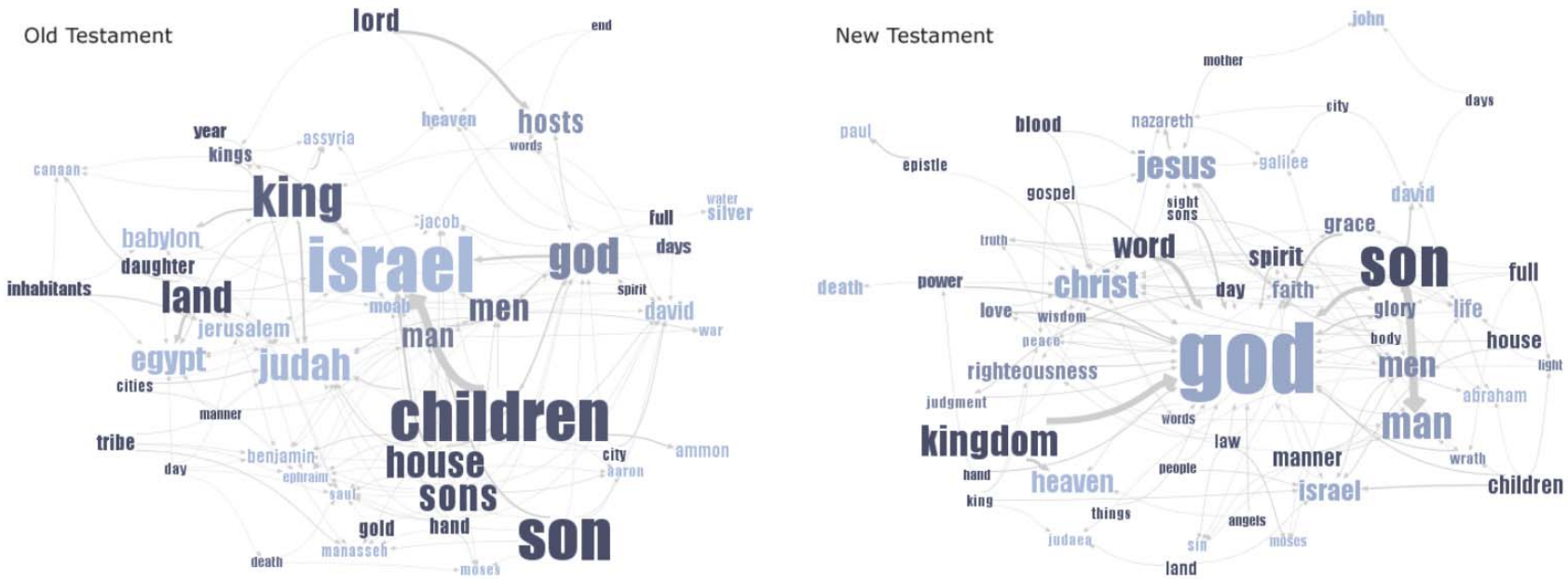
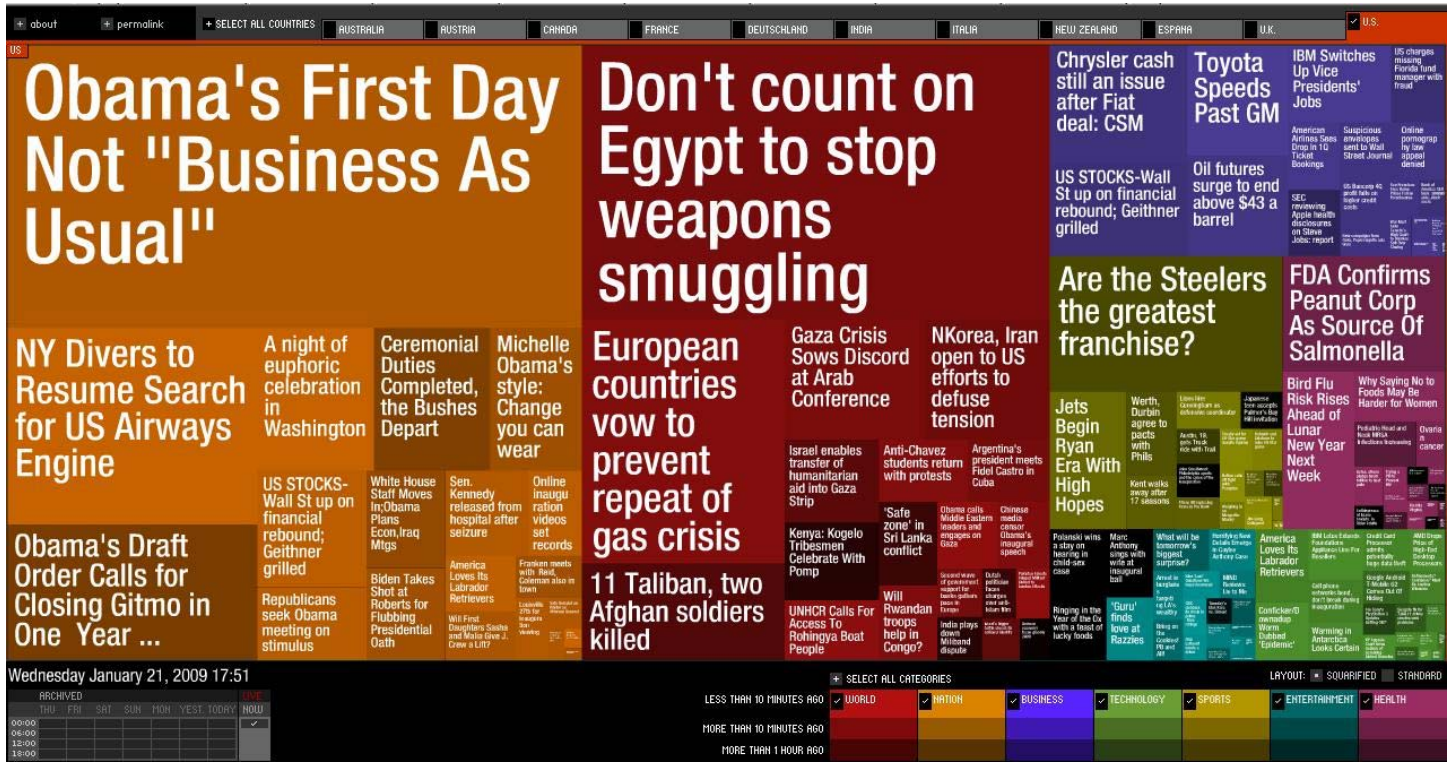


Fig 4. Matching the same pattern on different texts. Here we used the pattern “X of Y” to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

- 短語網路(Phrase Nets)是經典的力導向圖結構，圖中的節點是從文字中挖掘出的詞彙級或語法級的語義單元，邊代表語義單元的聯絡，邊的方向即短語的方向，邊的寬度是短語在文字中出現的頻率。

以圖為基礎的文字關係視覺化: 新聞地圖 (NewsMap)

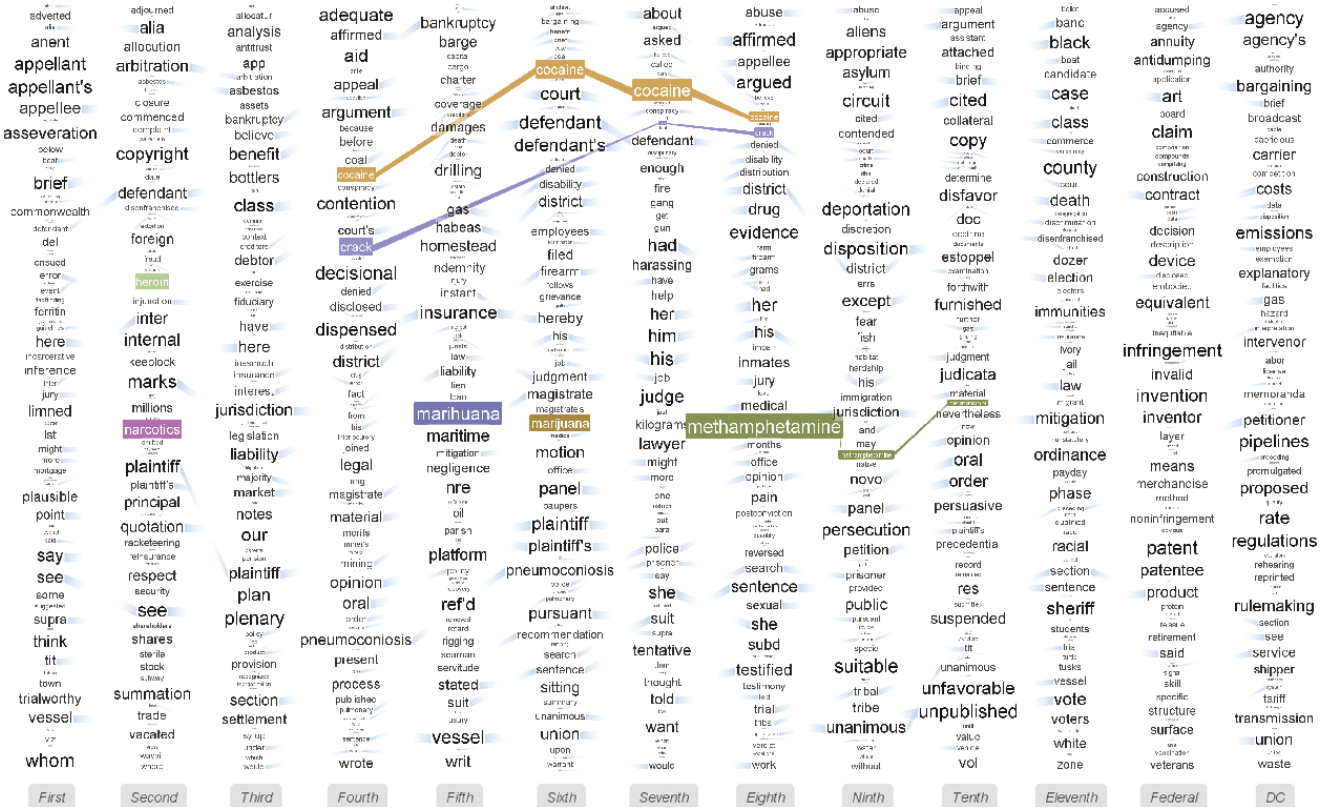


Maramushi's Newsmap takes news feeds from Google News and creates a visualisation based on the **popularity of stories**. It updates on the fly as news breaks with the tiles linking to the unabridged stories. Great for a snap-shot of what is going on in the world, even better for those far from home who want a quick over view of what is happening…about as much as ever it would appear.

- TreeMap 也是一種經典的視覺化關係佈局。
- NewsMap 就是基於 TreeMap 展示新聞，顏色用於區分新聞型別。

http://swei-industries.tumblr.com/post/387546517/maramushi-newsmap

文字多層面資訊視覺化: 平行標籤雲



Collins C, Viegas F B, Wattenberg M. Parallel tag clouds to explore and analyze faceted text corpora[C]//Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on. IEEE, 2009: 91-98.

- 平行標籤雲(Parallel Tag Clouds)結合了平行座標和標籤雲檢視。
- 每一列是一個層面的標籤雲，連線的折線展現了被選標籤在多個層面的分佈。

文件集合關係視覺化

- 文件集合關係視覺化: 星系視圖 (Galaxy View)
- 文件集合關係視覺化: 主題地貌 (ThemeScape)
- 文件集合關係視覺化: 基於範例的大文字集合投影
- 文字多層面資訊視覺化: 文字集點中存在多個層面的資訊和上下文連結資訊，
  - 例如時間，地點。
  - 例如: ContexTour, FacetAtlas, 平行標籤雲