
Basic Text Analysis using R
Justin Chun-ting Ho
Materials: https://github.com/justinchuntingho/

Text Analysis: The Basics

What is Text Analysis?

“Systematic, objective, quantitative analysis of message characteristics"

Kimberly A. Neuendorf, *The Content Analysis Guidebook*

Types of Text Analysis

Degree of human involvement:

- Human coding (100%)
- Supervised
- Unsupervised (0%)

Type of Objective:

- Scaling
- Classification

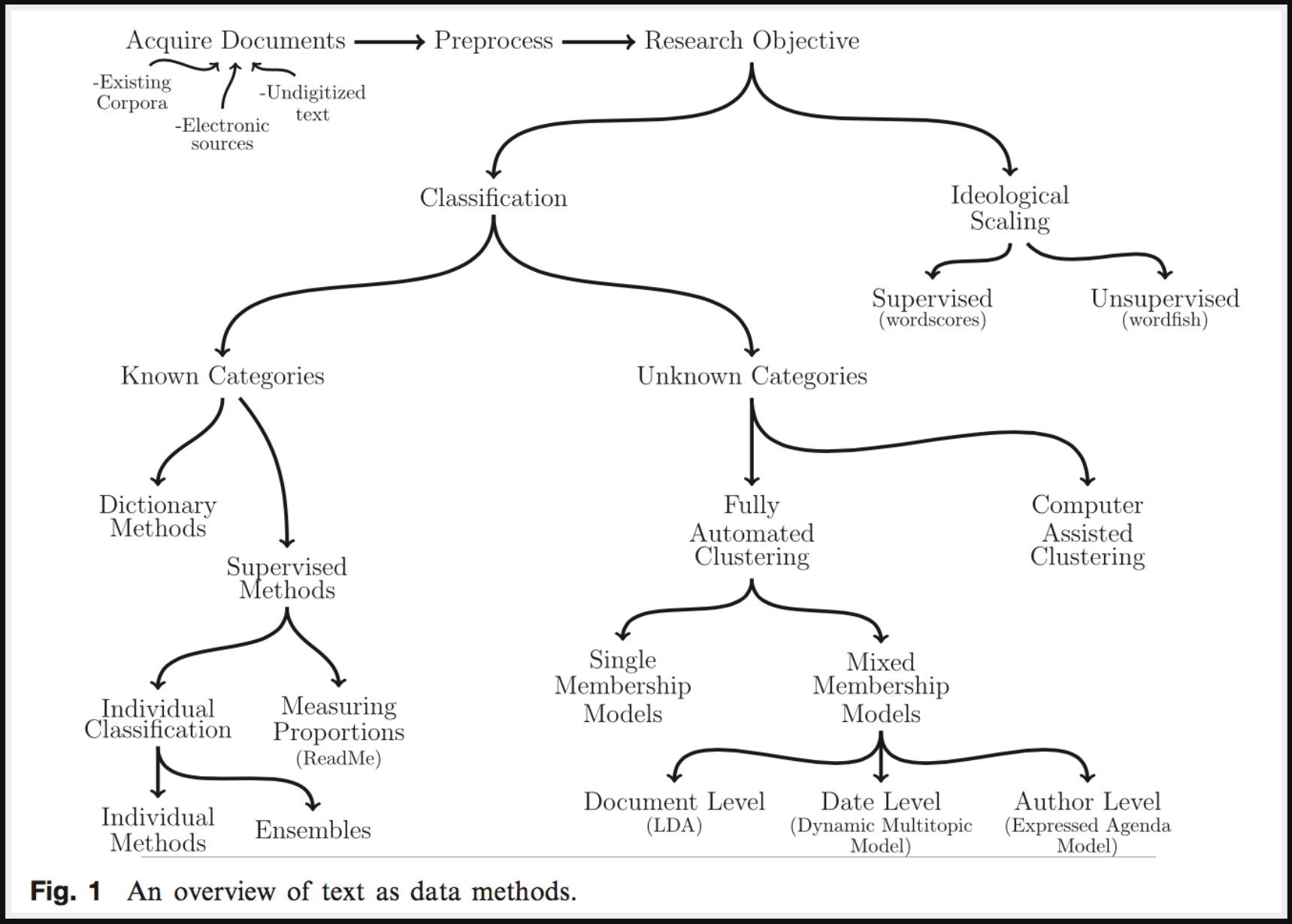


Fig. 1 An overview of text as data methods.

(Source: Justin Grimmer and Brandon Stewart, 2013)

Caution!

"All Quantitative Models of Language Are Wrong — But Some Are Useful"

Justin Grimmer and Brandon Stewart, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*

Bag of Words Assumption

- **Word order doesn’t matter**
- **The followings are exactly the same:**

I enjoy eating food and being with my family
I enjoy eating my family and being with food
and being eating enjoy family food I my with

The Pre-Processing

Original Text

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 181
##
## Article 1. All human beings are born free and equal in dignity
```

Remove Punctuation

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 178
##
## Article 1 All human beings are born free and equal in dignity
```

To Lower Case

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 178
##
## article 1 all human beings are born free and equal in dignity
```

Remove Numbers

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 177
##
## article  all human beings are born free and equal in dignity a
```

Remove Stopwords

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 135
##
## article   human beings  born free  equal  dignity  rights   en
```

Stemming

```
## <<PlainTextDocument>>
## Metadata:  7
## Content:  chars: 108
##
## articl human be born free equal digniti right endow reason con
```

Optional: Create N-Gram

```
##
## article_1 1_all all_human human_beings beings_are are_born bor
```

Create DFM

```
##      Terms
## Docs act anoth articl born brotherhood conscienc digniti endow
##    1   1     1      1    1           1         1       1     1
##    2   0     0      1    0           0         0       0     0
```

The Output

1. Document-Feature Matrix

	word1	word2	word3	...	wordN
doc1	0	2	0		3
doc2	1	0	0		0
doc3	0	0	2		1
...					
docN	0	1	0		0

2. Meta-data Martix

	var1	var2	var3	...	varN
doc1	Y	1	0		3
doc2	Y	0	0		2
doc3	N	0	0		1
...					
docN	Y	0	0		3


Time for R

Keyword Analysis

Keyword Analysis

What is a Keyword?

“A keyword may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind”

(Scott, M. (1997). PC analysis of key words - and key key words. System, 25(2), 233-45.)

Keyword Analysis

What is Keyness?

“The keyness of a keyword represents the value of log-likelihood or Chi-square statistics; in other words it provides an indicator of a keyword’s importance as a content descriptor for the appeal.”

(Scott, M. (1997). PC analysis of key words - and key key words. System, 25(2), 233-45.)

Keyword Analysis

What is a Chi-squared test?

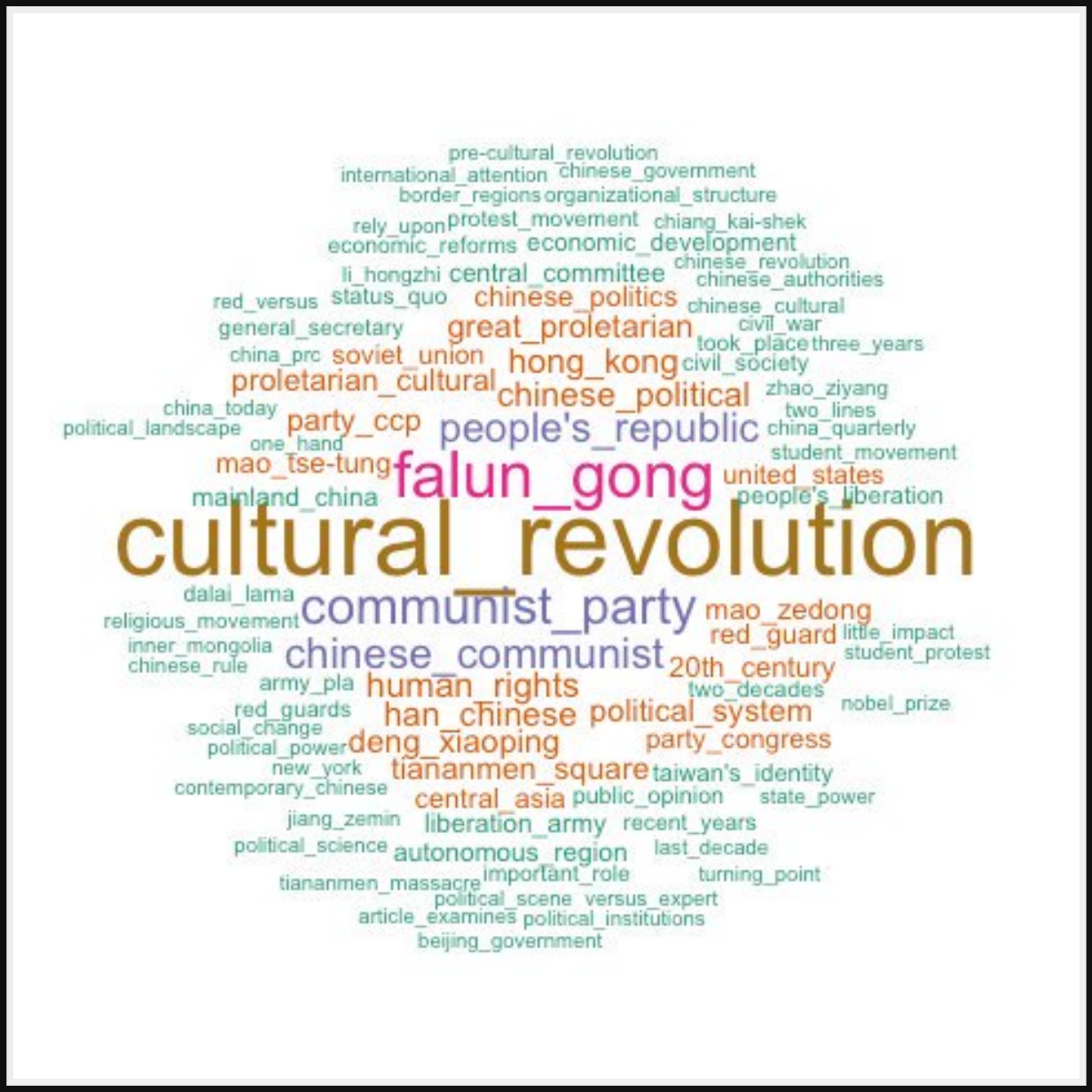
Comparison between the observed frequency and expected frequency.

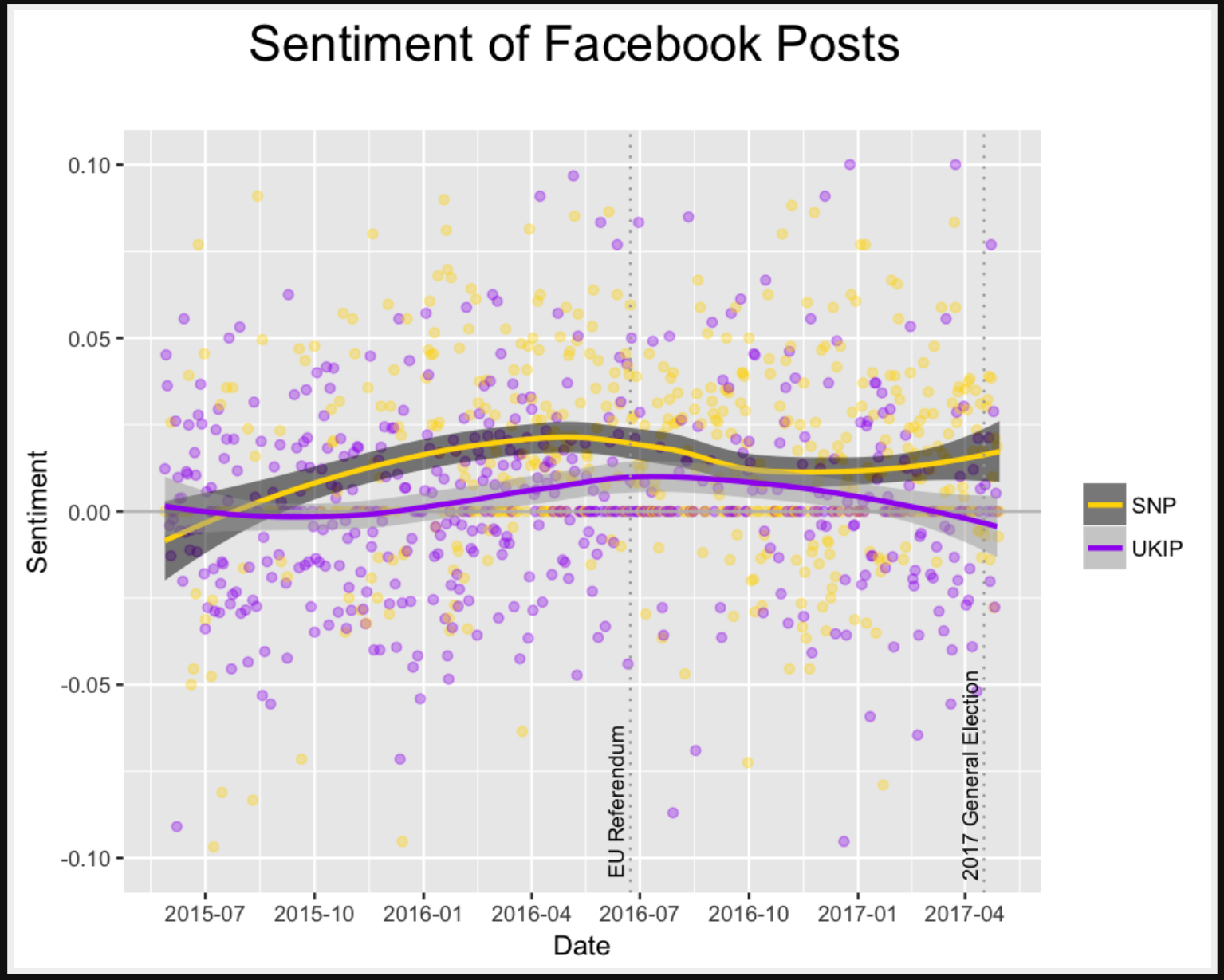
	love	All other words	Total
Male	414	1714029	1714443
Female	1214	2592238	2593452
Total	1628	4306267	4307895

- Expected frequency: **Row total** times **Column total** divided by the **total number of words** in the corpus.
- Plug into this equation: $$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$
- Comparison between the observed frequency and expected frequency of the word in question and all other words.


Time for R

Other Applications


pre-cultural_revolution
international_attention
chinese_government
border_regions
organizational_structure
rely_upon
protest_movement
chiang_kai-shek
economic_reforms
economic_development
chinese_revolution
li_hongzhi
central_committee
chinese_authorities
red_versus
status_quo
chinese_politics
chinese_cultural
general_secretary
great_proletarian
civil_war
took_place
three_years
china_prc
soviet_union
hong_kong
civil_society
proletarian_cultural
chinese_political
zhao_ziyang
china_today
two_lines
political_landscape
party_ccp
people's_republic
china_quarterly
one_hand
student_movement
mao_tse-tung
falun_gong
united_states
mainland_china
people's_liberation
cultural_revolution
dalai_lama
communist_party
mao_zedong
religious_movement
red_guard
little_impact
inner_mongolia
chinese_communist
student_protest
chinese_rule
20th_century
army_pla
human_rights
two_decades
red_guards
han_chinese
political_system
nobel_prize
social_change
political_power
deng_xiaoping
party_congress
new_york
tiananmen_square
taiwan's_identity
contemporary_chinese
central_asia
public_opinion
state_power
jiang_zemin
liberation_army
recent_years
political_science
autonomous_region
last_decade
important_role
tiananmen_massacre
turning_point
political_scene
versus_expert
article_examines
political_institutions
beijing_government

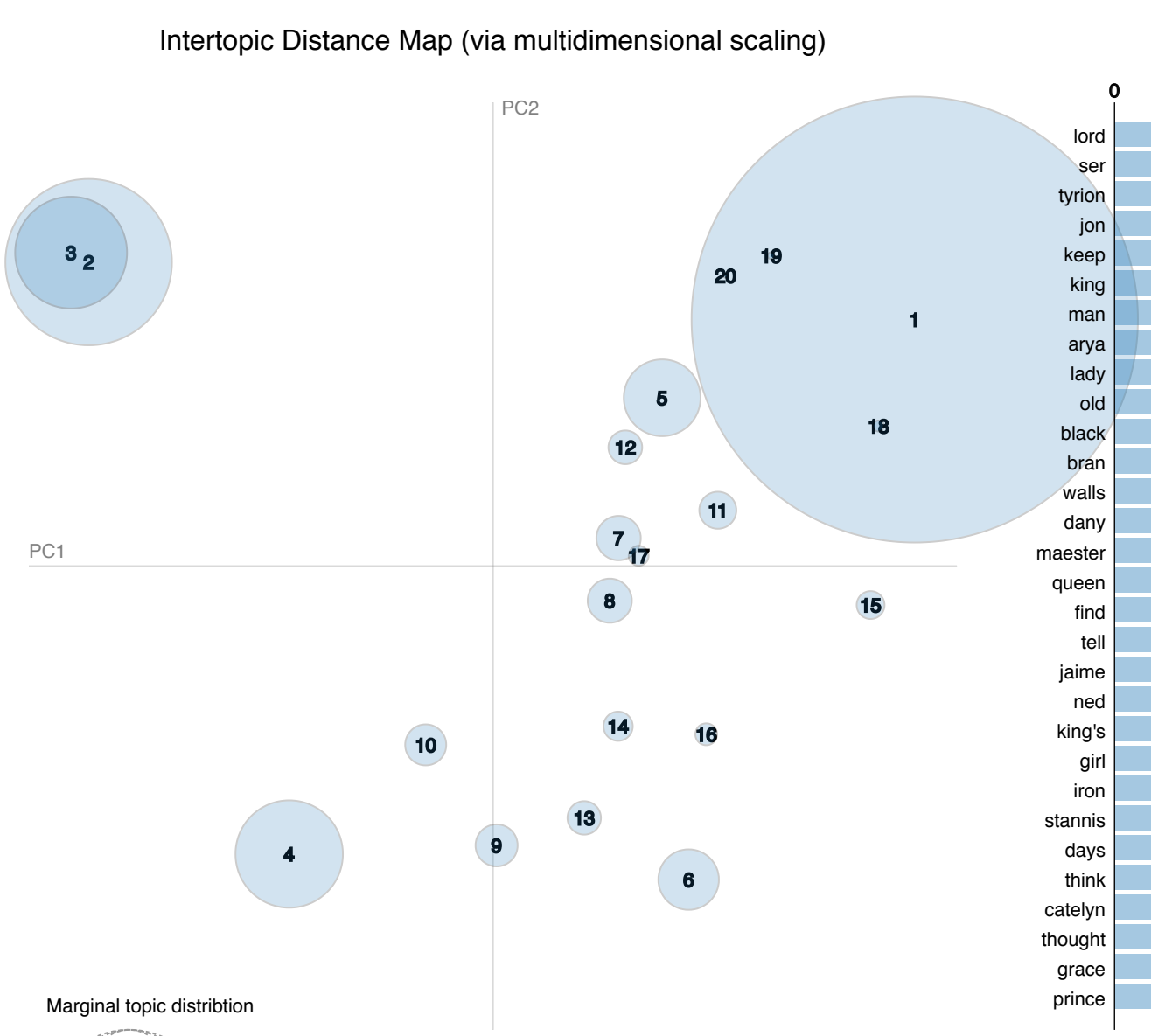

Sentiment of Facebook Posts
0.10
0.05
0.00
-0.05
-0.10
Sentiment
EU Referendum
2017 General Election
2015-07
2015-10
2016-01
2016-04
2016-07
2016-10
2017-01
2017-04
Date
SNP
UKIP

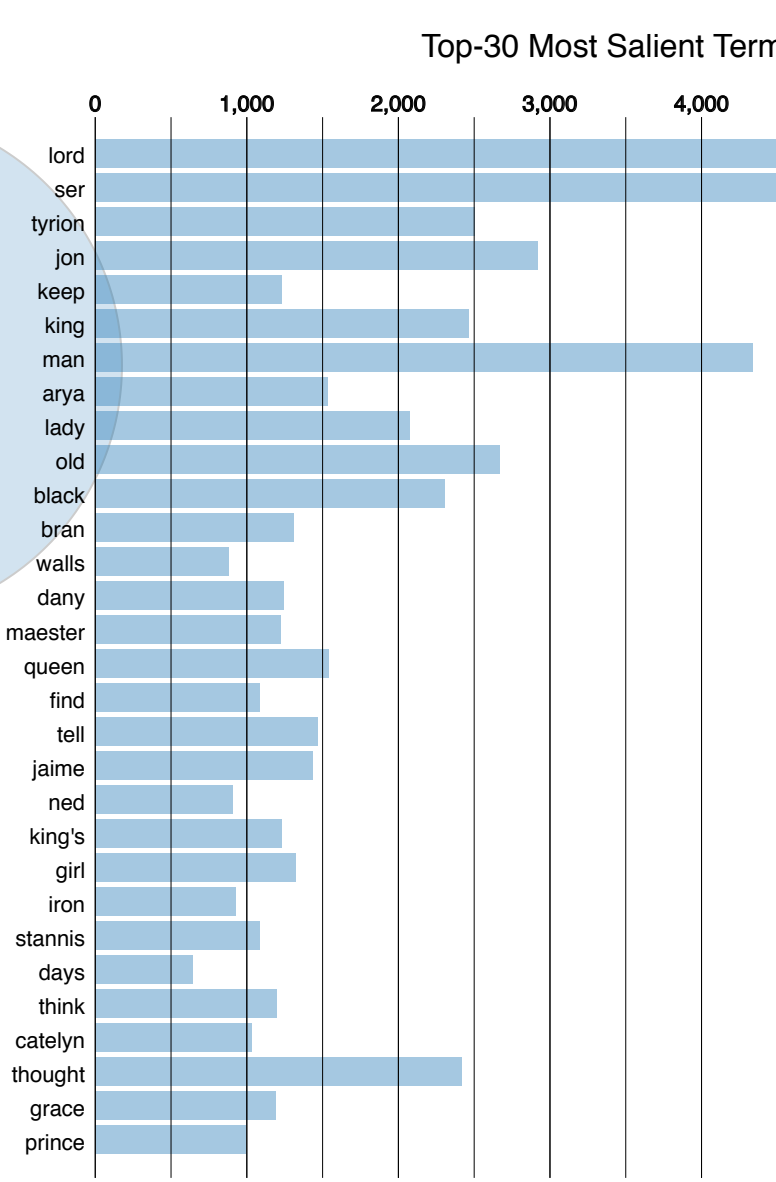
Selected Topic: 0 Previous Topic Next Topic Clear Topic

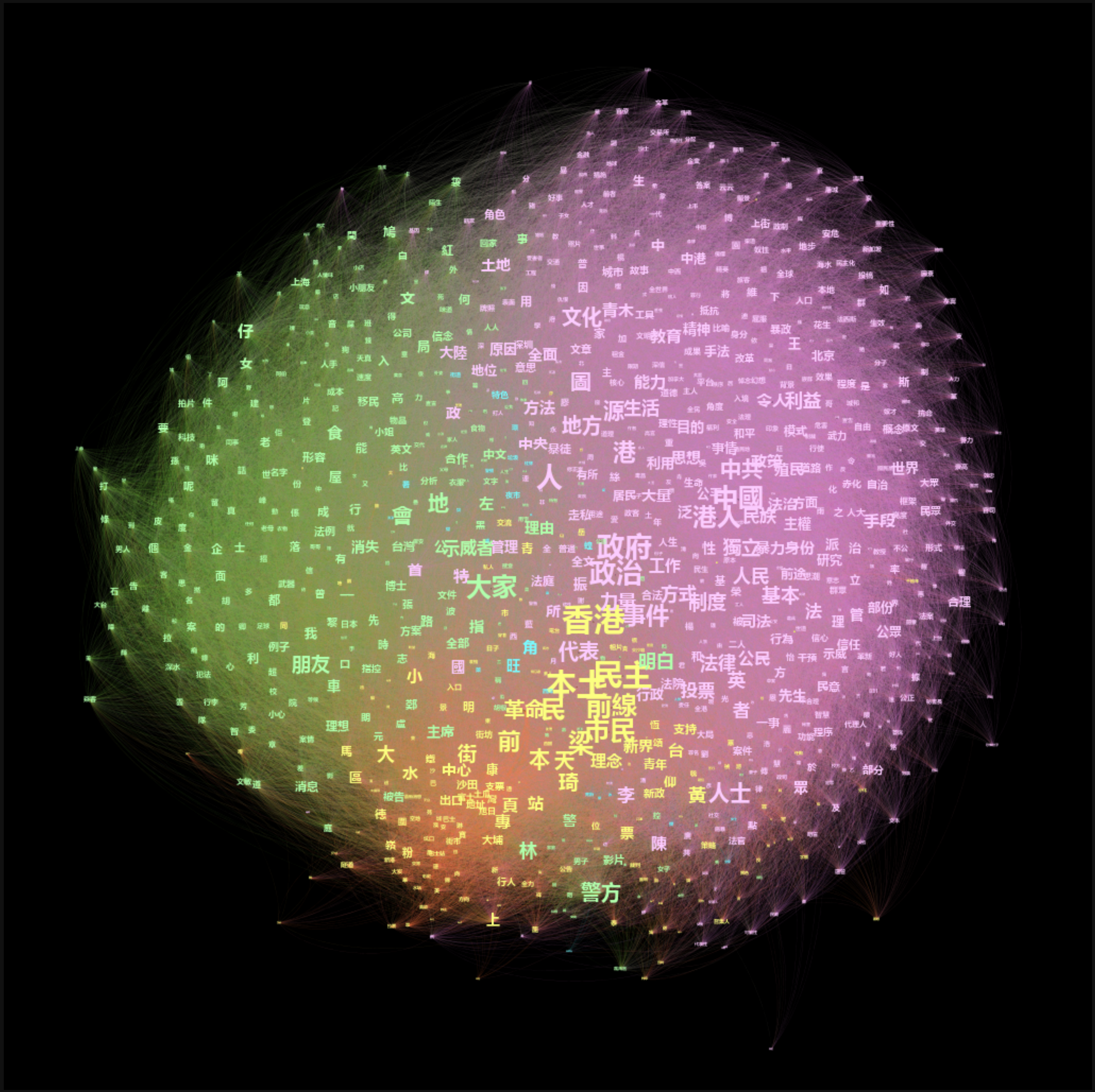
Slide to adjust relevance metric:(2)

λ = 1

0.0 0.2






香港
民主
本土
政府
政治
前線
市民
大家
中國
港人
人
事件
革命
梁
地
會
文化
港
中共
獨立
制度
代表
朋友
警方
人民
基本
法律
投票
黃
人士
地方
方式
明白
生活
利益
示威者
教育
能力
方法
天
琦
林
李
陳
街
前
本
世界
手段
公民

Thank you very much!

Twitter: @justin_ct_ho
Github: justinchuntingho
Email: Jusitn.Ho@ed.ac.uk

Edinburgh Text Analysis Research Group:
https://jiscmail.ac.uk/TEXTANALYSIS