Keyness

古貿昌、石晴方

Roadmap

- Definition of Keyness
- Qualitative vs. Quantitative Approach
- Statistics for Keyness
- Keyness Analysis in PTT Corpus
- Visualization

Definition of Keyness

- *Aboutness*: the understanding of the main concepts, topics or attitudes discussed in a text or corpus (Philips 1989)
- *Keyword*: a word which occurs with unusual frequency in a given text by comparison with a reference corpus of some kind (Scott 1997)
- *Keyness*: computed for words by comparing their frequencies in the target corpus to the frequencies in a reference corpus.

Qualitative vs. Quantitative Approach

- **Qualitative Approach**: *keywords* as terms presumably carrying socio-cultural meanings characteristic of ideologies (Williams 1976).
    - were determined based on the **subjective** judgement of the socio-cultural meanings of the predefined list of words
- **Quantitative Approach**: a bottom-up **corpus-based** method to discover key terms reflecting the ideological undercurrents of particular text collections (Stubbs 1996, 2003)
    - a data-driven approach sympathetic to the notion of **statistical** keywords

Keyness Analysis

- To evaluate whether the word occurs more frequently in the **target** corpus as compared to its occurrence in the **reference** corpus. If yes, the word may be a key term of the target corpus.
- We can quantify the relative attraction of each word to the target corpus by means of a statistical association metric.
- This kind of analysis can be extended to other linguistic units as well.

  e.g. key phrases, key ngrams ...

Statistics for Keyness

- **Difference Coefficient** (Leech and Fallon 1992): (a-c) / (a+c)
- **Relative Frequency Ratio** (Damerau 1993): (a/c) / [(a+b)/(c+d)]
- **Chi-squared ($\chi 2$)** (Aarts 1971, 2004): $\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
- **Log-likelihood Ratio (G^2)** (Dunning 1993): $G^2 = 2 \sum_i O_{ij} log(\frac{O_{ij}}{E_{ij}})$

$$E_{ij} = \frac{N_i N_j}{N}$$

Corpus A (target corpus) : {橘子、蘋果、葡萄、芭樂、檸檬、番茄 }

Corpus B (reference corpus) : {荔枝、榴槤、番茄、橘子、柳丁、橘子、橘子、香蕉}

Exercise: We want to compute the keyness of '橘子' in the two corpora.

First, we obtain a **contingency table** like below.

	word (橘子)	**other words (~橘子)**	**total**
Corpus A	a = O11 = 1	b = O12 = 5	a + b = 6
Corpus B	c = O21 = 3	d = O22 = 5	c + d = 8
total	a + c = 4	b + d = 10	a + b + c + d = 14

	word	other words	total
target corpus	a = O11 = 1	b = O12 = 5	a + b = 6
reference corpus	c = O21 = 3	d = O22 = 5	c +d = 8
total	a + c = 4	b + d = 10	a + b + c + d = 14

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad E_{ij} = \frac{N_i N_j}{N} \qquad G^2 = 2 \sum_i O_{ij} log(\frac{O_{ij}}{E_{ij}})$$

E11 = 4 × 6 / 14 = 1.714 ; E12 = 10 × 6 / 14 = 4.286 ; E21= 4 × 8 / 14 = 2.286 ; E22 = 10 × 8 / 14 = 5.714

Chi2 = $(O11 - E11)^2 / E11 + (O12 - E12)^2 / E12 + (O21 - E21)^2 / E21 + (O22 - E22)^2 / E22$

= $(1 - 1.714)^2 / 1.714 + (5 - 4.286)^2 / 4.286 + (3 - 2.286)^2 / 2.286 + (5 - 5.714)^2 / 5.714$

= 0.73

Case Study :
Keyness Analysis in PTT Corpus

- A keyness analysis to the “Gossiping Board” and “WomenTalk Board” in PTT.
- Data Pre-Processing
  - Word Segmentation
  - Frequency List

Creating Dataframes

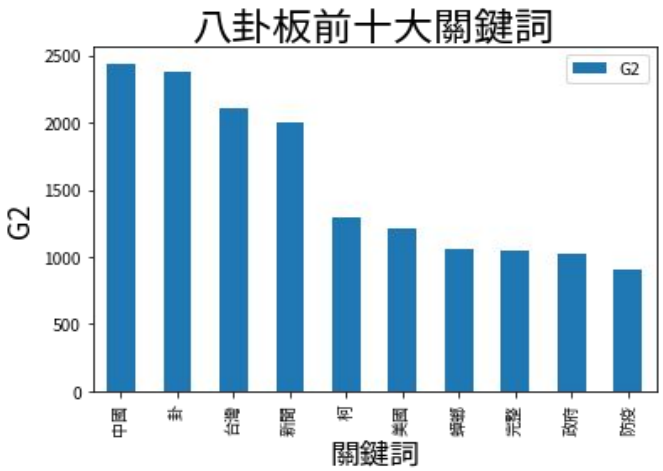
八卦版前十大關鍵詞

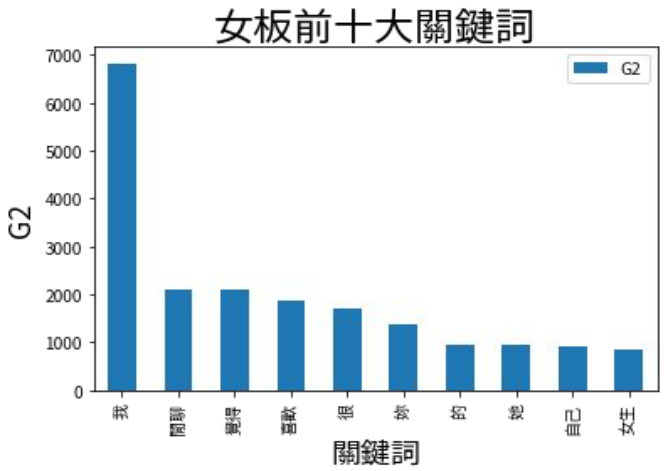
	word	pref	chi2	G2
0	中國	tgt_corpus	2141.280249	2440.416732
1	卦	tgt_corpus	1783.951928	2374.213899
2	台灣	tgt_corpus	1955.260280	2108.527661
3	新聞	tgt_corpus	1740.090568	1998.836604
4	柯	tgt_corpus	1022.848097	1297.490043
5	美國	tgt_corpus	1047.899530	1215.179207
6	蟑螂	tgt_corpus	897.489417	1061.776213
7	完整	tgt_corpus	812.411001	1051.849217
8	政府	tgt_corpus	906.442856	1019.902739
9	防疫	tgt_corpus	744.402694	909.280864

女版前十大關鍵詞

	word	pref	chi2	G2
0	我	ref_corpus	6745.572660	6826.209723
1	閒聊	ref_corpus	1635.144850	2114.888689
2	覺得	ref_corpus	2045.699349	2107.149524
3	喜歡	ref_corpus	1764.370396	1886.710972
4	很	ref_corpus	1700.451545	1703.984082
5	妳	ref_corpus	1251.704906	1365.127481
6	的	ref_corpus	952.375242	949.828993
7	她	ref_corpus	933.917937	948.420217
8	自己	ref_corpus	918.329003	921.477614
9	女生	ref_corpus	789.866815	834.000558

Creating Bar Charts





Creating Wordclouds

八卦版文字雲

女版文字雲

Case Study :
Keyness Analysis in PTT Corpus

- Get to Colab!
  - https://colab.research.google.com/drive/1hK_TfcrKkGFicL1ff2wuNExFwJ6APeYs
  - (complete version)
    https://colab.research.google.com/drive/19ldWiUmlxR_jqGzEFnxr908wX72nyAb9

References

Aarts, F. G. A. M. 1971. On the distribution of noun-phrase types in English clause structure. Lingua, 26, 281-93. Reprinted in G. Sampson & D. McCarthy (eds.) 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum. pp. 35-57.

Damerau, F. J. 1993. “Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts.” *Information Processing & Management* 29 (4): 433-447.

Dunning, T. 1993. “Accurate Methods for the Statistics of Surprise and Coincidence.” *Computational Linguistics* 19 (1): 61-74.

Gabrielatos, C. 2018. “Keyness Analysis: Nature, Metrics and Techniques.” In *Corpus Approaches to Discourse*, 225–58. Routledge.

Gries, S. T. 2018. *Quantitative Corpus Linguistics with R: A Practical Introduction*. 2nd ed. Routledge.

Leech, G., and R. Fallon. 1992. “Computer Corpora–What Do They Tell Us About Culture.” *ICAME Journal* 16.

Phillips, M. 1989. *Lexical Structure of Text*. Discourse Analysis Monograph no. 12. English Language Research, University of Birmingham.

Scott, M. 1997. PC analysis of key words - and key key words. *System* 25 (2): 233-245.

Stubbs, M. 1996. *Text and Corpus Analysis*. Blackwell.

Stubbs, M. 2003. *Words and Phrases*. Blackwell.

Williams, R. 1976. *Keywords*. Oxford University Press.

https://alvinntnu.github.io/NTNU_ENC2036/