Lesson 9: Text Data Visualisation and Analysis: Concepts and Methods

Dr. Kam Tin Seong
Assoc. Professor of Information Systems

School of Computing and Information Systems,
Singapore Management University

2021-07-04 (updated: 2022-01-03)

Content

- Introduction Text Visualisation
  - Text data
- Text Visualisation Methods
  - Tag Cloud
  - Wordle
  - Word Tree
  - Phrase Nets
- R Packages for Text Visualisation

Introduction to Text Visualisation

History of text

Chauvet cave proto-writing

~20,000 years ago

Sumerian cuneiform logographic

~5,000 years ago

Phoenician abjad predecessor of alphabet

~3,000 years ago

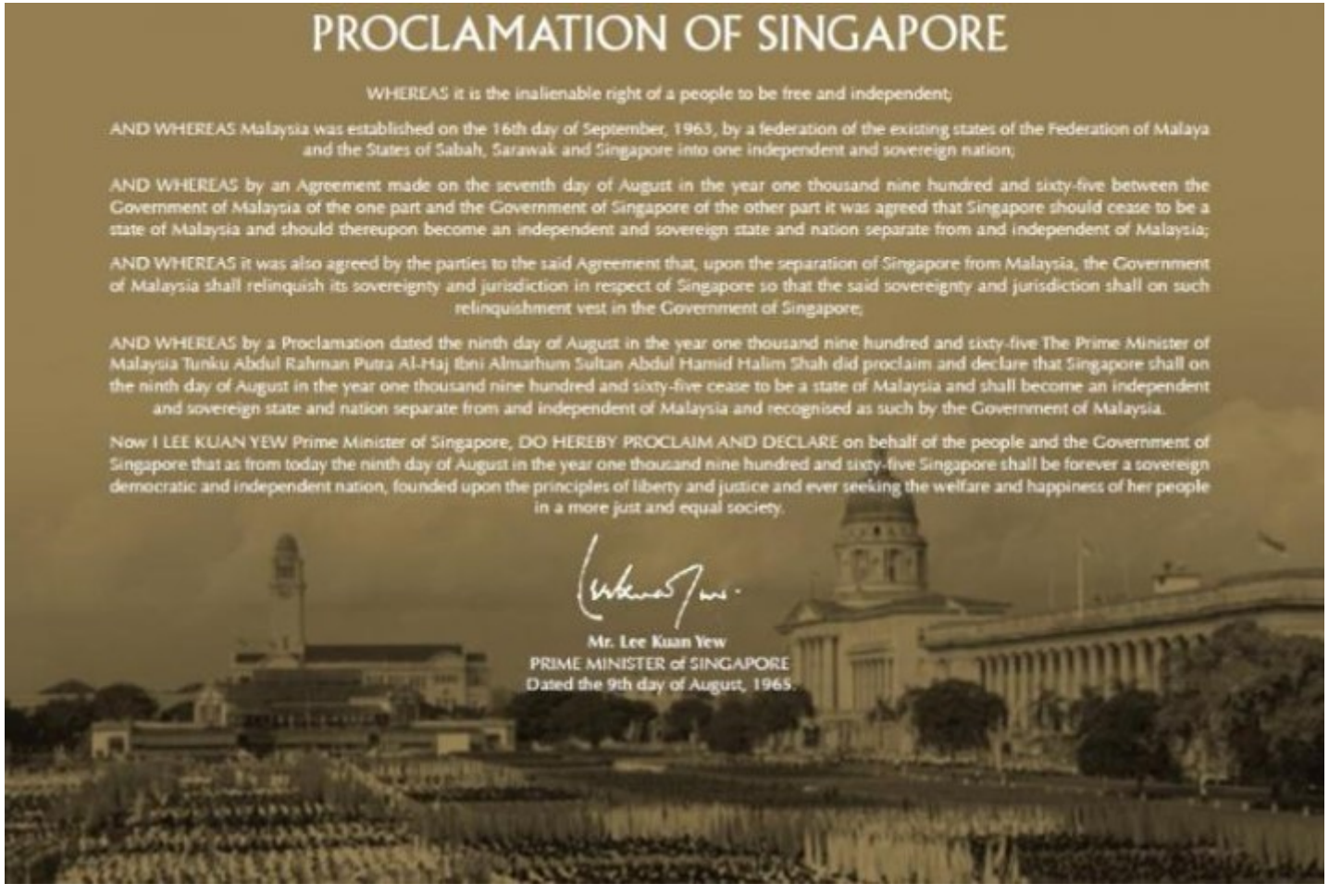
Latin letters

~2,500 years ago

ABCDEF
GHIJKLM
NOPQRS
TUVWXYZ

Introduction to Text Visualisation

Text as historical archive



PROCLAMATION OF SINGAPORE

WHEREAS it is the inalienable right of a people to be free and independent;

AND WHEREAS Malaysia was established on the 16th day of September, 1963, by a federation of the existing states of the Federation of Malaya and the States of Sabah, Sarawak and Singapore into one independent and sovereign nation;

AND WHEREAS by an Agreement made on the seventh day of August in the year one thousand nine hundred and sixty-five between the Government of Malaysia of the one part and the Government of Singapore of the other part it was agreed that Singapore should cease to be a state of Malaysia and should thereupon become an independent and sovereign state and nation separate from and independent of Malaysia;

AND WHEREAS it was also agreed by the parties to the said Agreement that, upon the separation of Singapore from Malaysia, the Government of Malaysia shall relinquish its sovereignty and jurisdiction in respect of Singapore so that the said sovereignty and jurisdiction shall on such relinquishment vest in the Government of Singapore;

AND WHEREAS by a Proclamation dated the ninth day of August in the year one thousand nine hundred and sixty-five The Prime Minister of Malaysia Tunku Abdul Rahman Putra Al-Haj Ibni Almarhum Sultan Abdul Hamid Halim Shah did proclaim and declare that Singapore shall on the ninth day of August in the year one thousand nine hundred and sixty-five cease to be a state of Malaysia and shall become an independent and sovereign state and nation separate from and independent of Malaysia and recognised as such by the Government of Malaysia.

Now I LEE KUAN YEW Prime Minister of Singapore, DO HEREBY PROCLAIM AND DECLARE on behalf of the people and the Government of Singapore that as from today the ninth day of August in the year one thousand nine hundred and sixty-five Singapore shall be forever a sovereign democratic and independent nation, founded upon the principles of liberty and justice and ever seeking the welfare and happiness of her people in a more just and equal society.

Mr. Lee Kuan Yew
PRIME MINISTER of SINGAPORE
Dated the 9th day of August, 1965.

Introduction to Text Visualisation

Text as a mode of communication

Introduction to Text Visualisation

Textual data for business intelligence analytics

By Channel NewsAsia, Updated: 08/10/2011

CHANNEL NEWSASIA

Market for new flats likely to stabilise in four years: PM Lee

Market for new flats likely to stabilise in four years: PM Lee

33% 67% Shared 5 times

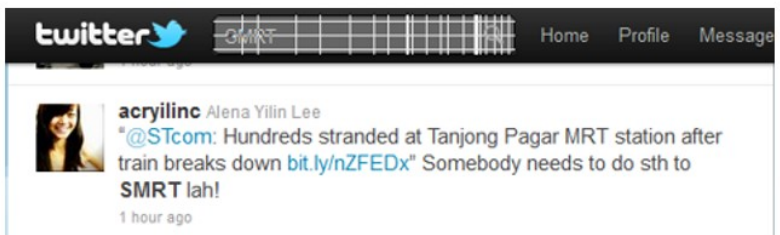
Tweet 2 Recommend 22

SINGAPORE: Prime Minister Lee Hsien Loong said he expects the recent slew of housing measures to stabilise the market for new flats in four years. He also gave his assurance that the prices of new flats will remain affordable.

Mr Lee was speaking at a ceremony to mark the completion of upgrading works in a precinct in Ang Mo Kio.

In his speech, Mr Lee noted the concerns many Singaporeans have over buying new flats and he outlined several ways in which housing authorities are working to meet demand.

Some of these measures have started to take effect, and Mr Lee said he expects the market to stabilise in about four years' time.



Sarah

Spacious and in a very quiet area. Very pleasant. Barbara and Nicolas are welcoming and kind. My daughter and I are delighted to stay everything was perfect. And the place is close to everything really perfect plan.

October 2014 Translated from Français

Vicky

Great location and cozy and clean room. Nicolas and Barbara were great hosts. We had a wonderful experience to stay in Nicolas and Barbara's house. And we would like to stay longer when we come to Paris again. Thank you very much, Barbara and Nicolas!

October 2014

Coralie

Second stay with Nicolas and Barbara, always nice!
The room is spacious, airy and comfortable. Welcoming Nicolas and Barbara was very attentive. I thoroughly recommend it!

October 2014 Translated from Français

Why Visualise Text?

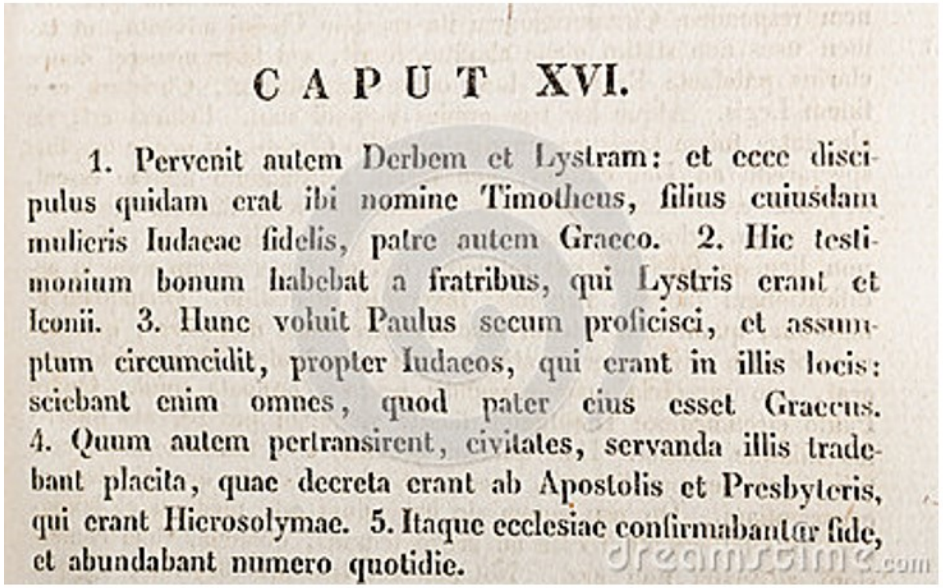
- Understanding – get the “gist” of a document
- Grouping – cluster for overview or classification
- Compare – compare document collections, or
- Inspect evolution of collection over time
- Correlate – compare patterns in text to those in other data, e.g., correlate with social network

Levels of Text Representation

- Lexical level, transforming a string of characters into a sequence of atomic entities, called tokens.
- Syntactic level, identifying and tagging (anotating) each token’s functions.
- Semantic level, extracting of meaning and relationships between pieces of knowledge derived from the structures identified in the syntactical level.

Fundamental of Text Visualisation

Be warn, not all text are written in English and in digital forms!



CAPUT XVI.

1. Pervenit autem Derbem et Lystram: et ecce discipulus quidam erat ibi nomine Timotheus, filius cuiusdam mulieris Iudaeae fidelis, patre autem Graeco. 2. Hic testimonium bonum habebat a fratribus, qui Lystris erant et Iconii. 3. Hunc voluit Paulus secum proficisci, et assumptum circumcidit, propter Iudaeos, qui erant in illis locis: sciebant enim omnes, quod pater eius esset Graecus. 4. Quum autem pertransirent, civitates, servanda illis tradebant placita, quae decreta erant ab Apostolis et Presbyteris, qui erant Hierosolymae. 5. Itaque ecclesiae confirmabantur fide, et abundabant numero quotidie.

字彙補序
近人論字學莫不推
宣城字彙綜而不遺
條分而不紊爲書學

Text Visualisation Methods

- Tag Cloud
- Wordle
- Word Tree
- Phrase Nets

Tag Cloud

- A tag cloud (word cloud, or weighted list in visual design) is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text.
- 'Tags' are usually single words, normally listed alphabetically, and the importance of each tag is shown with font size or color.

Source: Tag cloud

Application of Tag Cloud I: Branding

- One-word tag cloud of DBS's corporate values statement created using Many Eyes.

Application of Tag Cloud I: Branding

- Two-word tag cloud of DBS's corporate values statement created using Many Eyes.

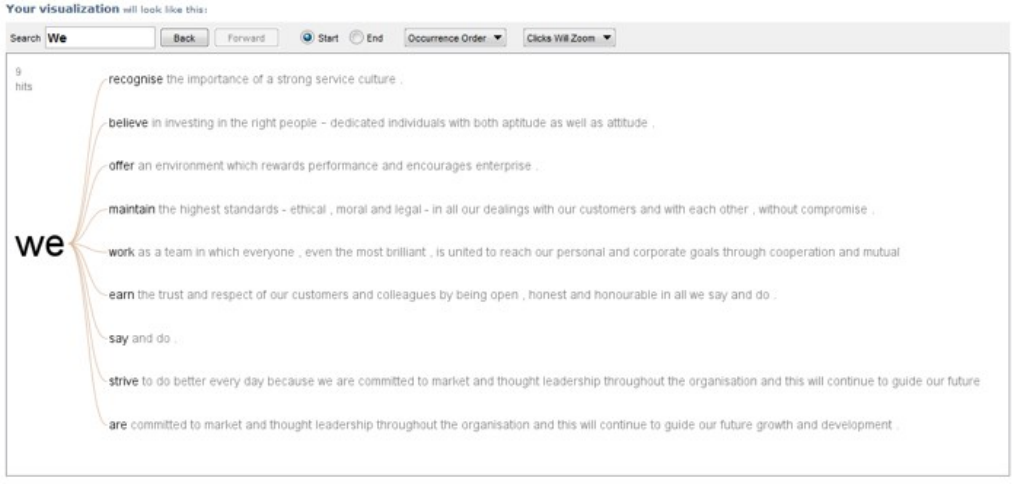
Wordle

- A toy for generating “word clouds” from text that you provide

Word Clouds of Corporate Values Statements

Word Tree

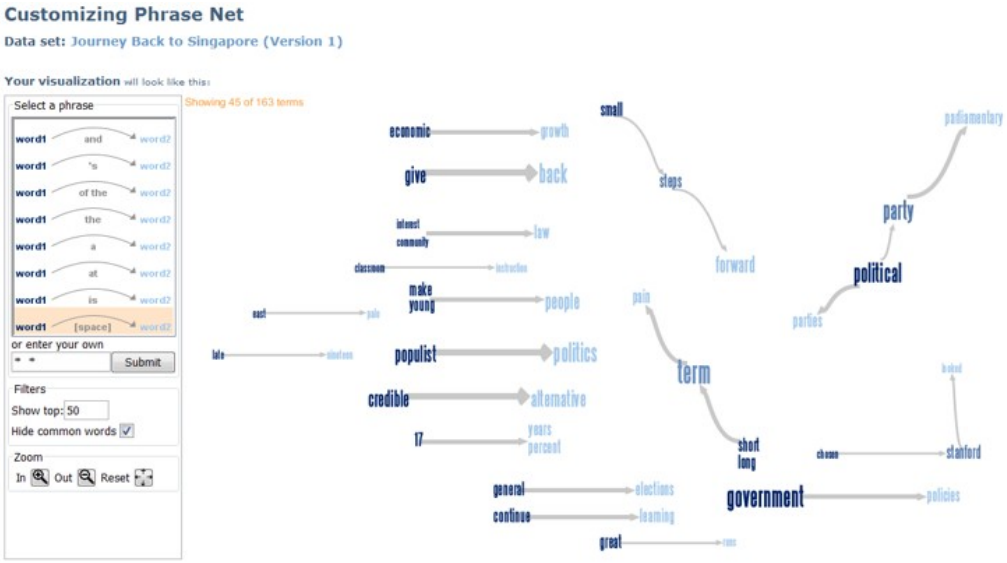
- A visual search tool for unstructured text, such as a book, article, speech or poem. It lets you pick a word or phrase and shows you all the different contexts in which the word or phrase appears.
- The contexts are arranged in a tree-like branching structure to reveal recurrent themes and phrases.



Link: https://www.jasondavies.com/wordtree/

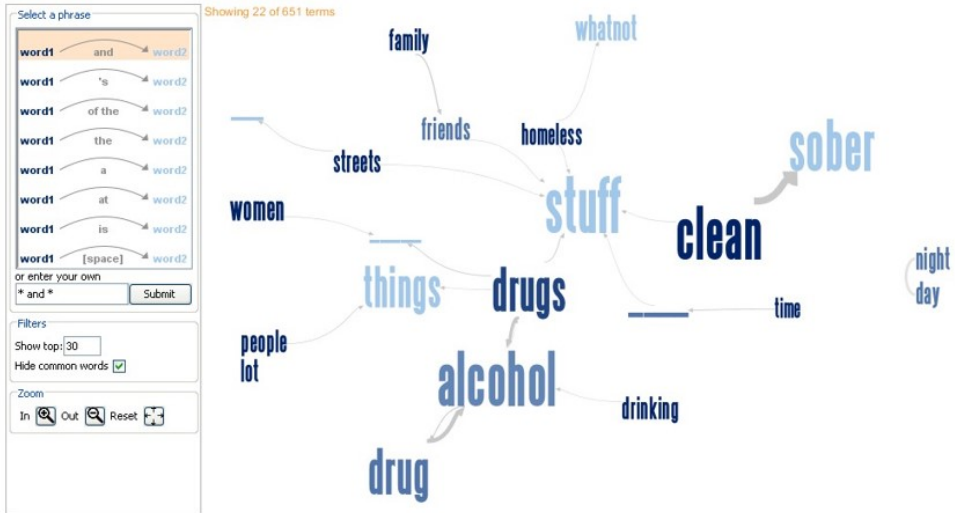
Phrase Net

- A phrase net diagrams the relationships between different words used in a text. It uses a simple form of pattern matching to provide multiple views of the concepts contained in a book, speech, or poem.

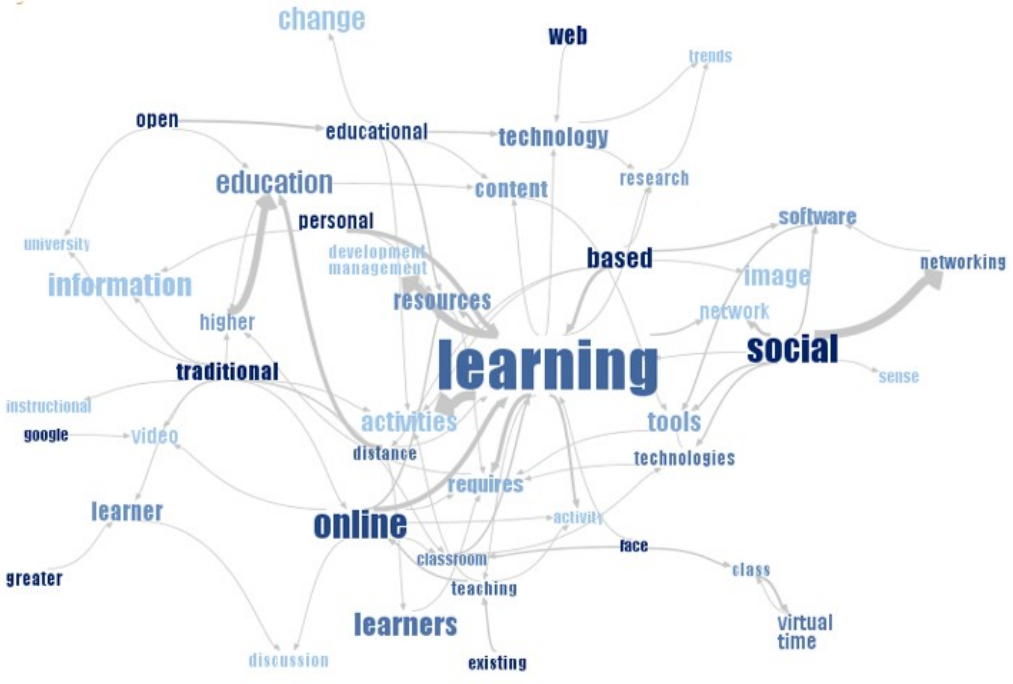


Phrase Net

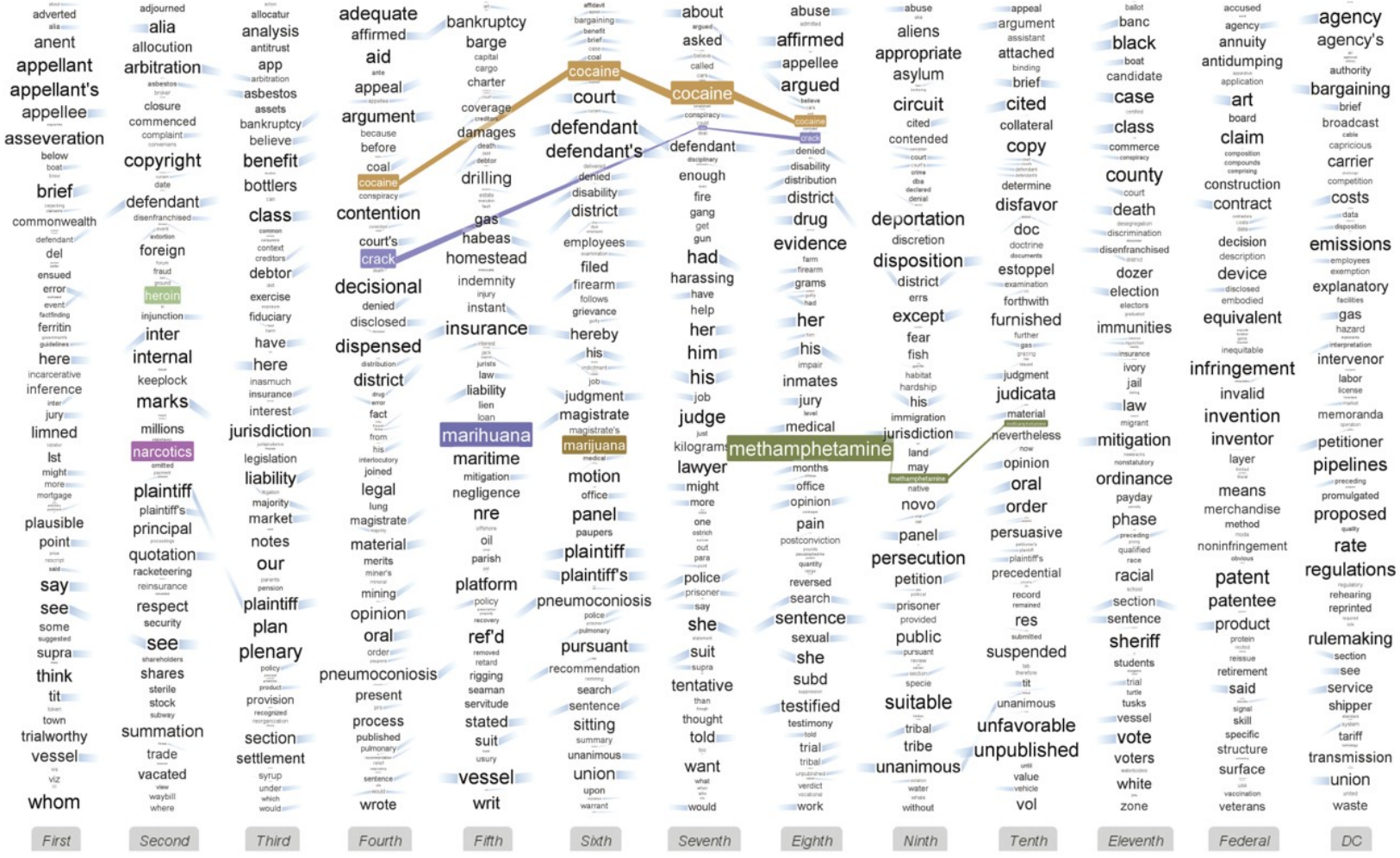
Words separate by the keyword "and"



Words that directly follow one another

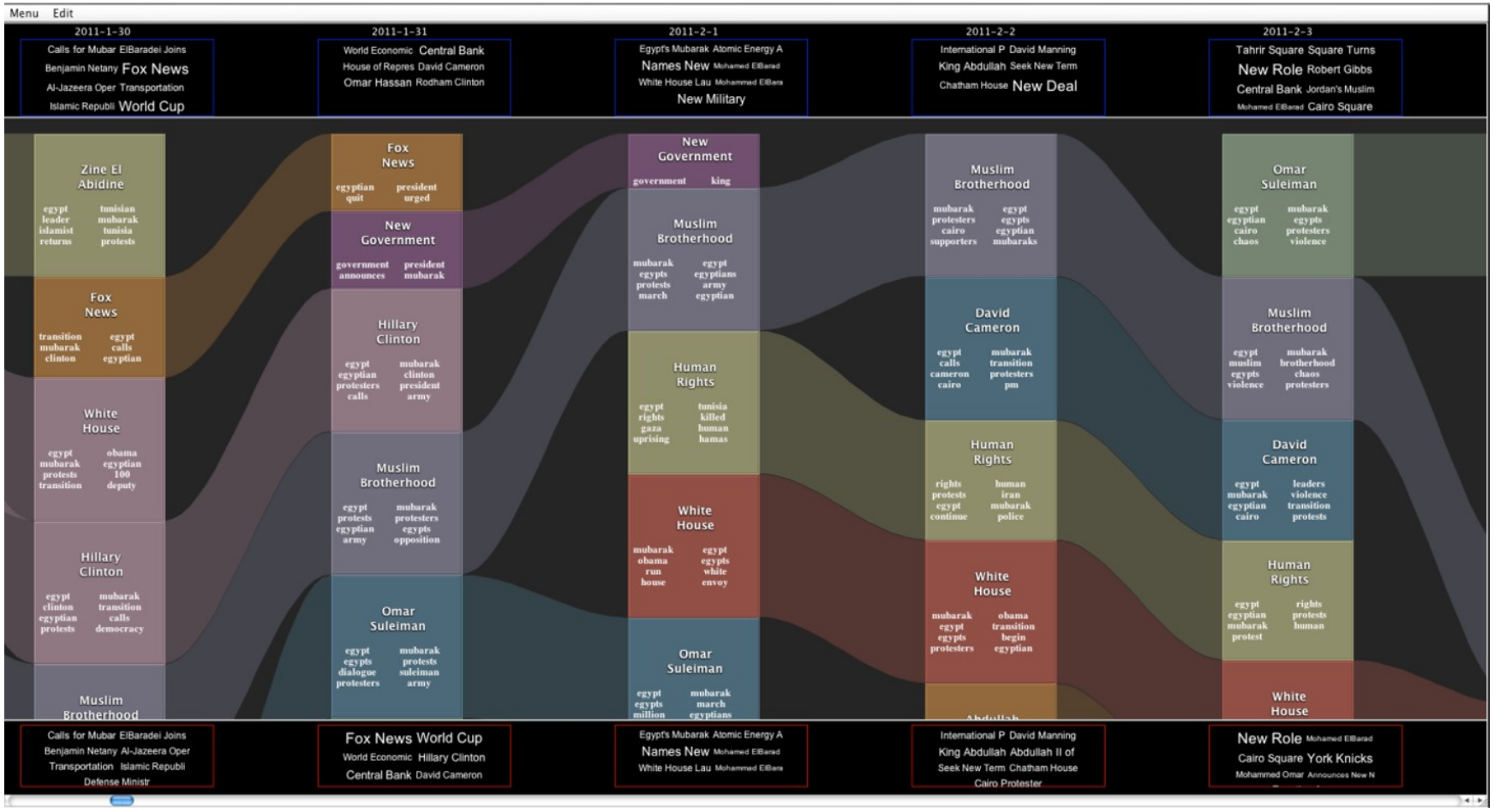


Parallel Tag Cloud



Reference: Parallel Tag Clouds to Explore and Analyze Faceted Text Corpora

Story Tracker: Main View



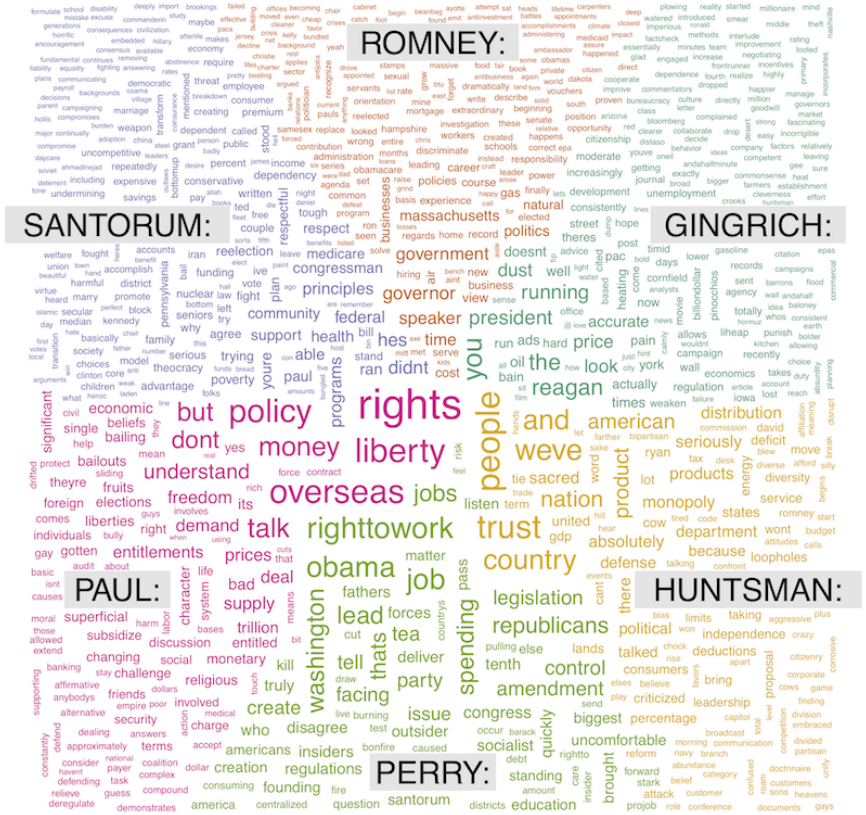
Reference: Story Tracker: Incremental visual textanalytics of news story development

R packages for Text Visualisation

- ggwordcloud: a word cloud geom for ggplot2
- TextPlot: R Library for Visualizing Text Data

wordcloud

- Provides functionality to create pretty word clouds, visualize differences and similarity between documents, and avoid over-plotting in scatter plots with text.
- Visit this link for more information.



wordcloud2: Create Word Cloud by 'htmlwidget'

- A fast visualization tool for creating wordcloud by using wordcloud2.js, ia JavaScript library to create wordle presentation on 2D canvas or HTML.
- It provides Shiny functions.
- Visit this link for more information.

ggwordcloud: a word cloud geom for ggplot2

- ggwordcloud provides a word cloud text geom for ggplot2.
- as an alternative to wordcloud and wordcloud2.

Wordcloud on Shiny

Shiny from RStudio

Back to Gallery Get Code

Word Cloud

Choose a book:

Romeo and Juliet

Change

Minimum Frequency:

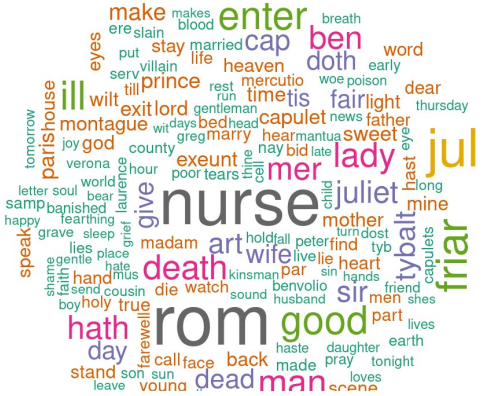
1 12 50

1 6 11 16 21 26 31 36 41 46 50

Maximum Number of Words:

1 300

1 31 61 91 121 151 181 211 241 271 300



global.R server.R ui.R show below

```
fluidPage(
  # Application title
  titlePanel("Word Cloud"),

  sidebarLayout(
    # Sidebar with a slider and selection inputs
    sidebarPanel(
      selectInput("selection", "Choose a book:",
                  choices = books),
      actionButton("update", "Change"),
      hr(),
      sliderInput("freq",
                  "Minimum Frequency:",
                  min = 1,  max = 50, value = 15),
      sliderInput("max",
                  "Maximum Number of Words:",
                  min = 1,  max = 300,  value = 100)
    ),

    # Show Word Cloud
    mainPanel(
      plotOutput("plot")
    )
  )
)
```

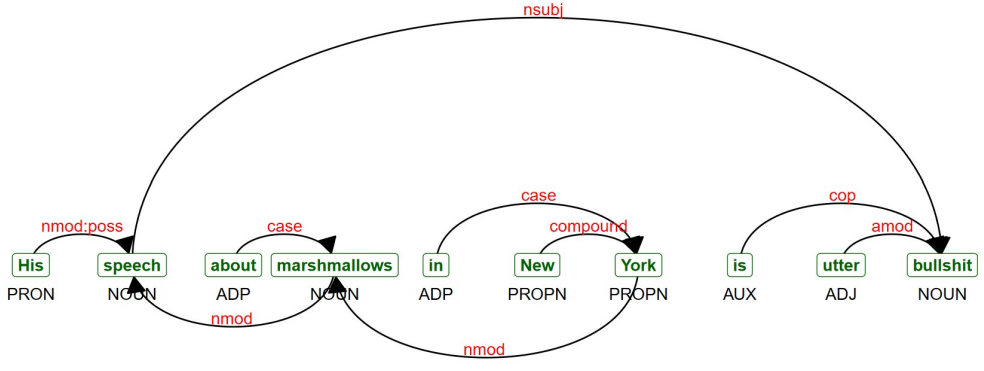
For live demo, visit this link

TextPlot: R Library for Visualizing Text Data

- Aims to visualise complex relations in texts.
- Provides functionalities for displaying text co-occurrence networks, text correlation networks, dependency relationships as well as text clustering.
- Visit this link for more information.

Dependency Parser

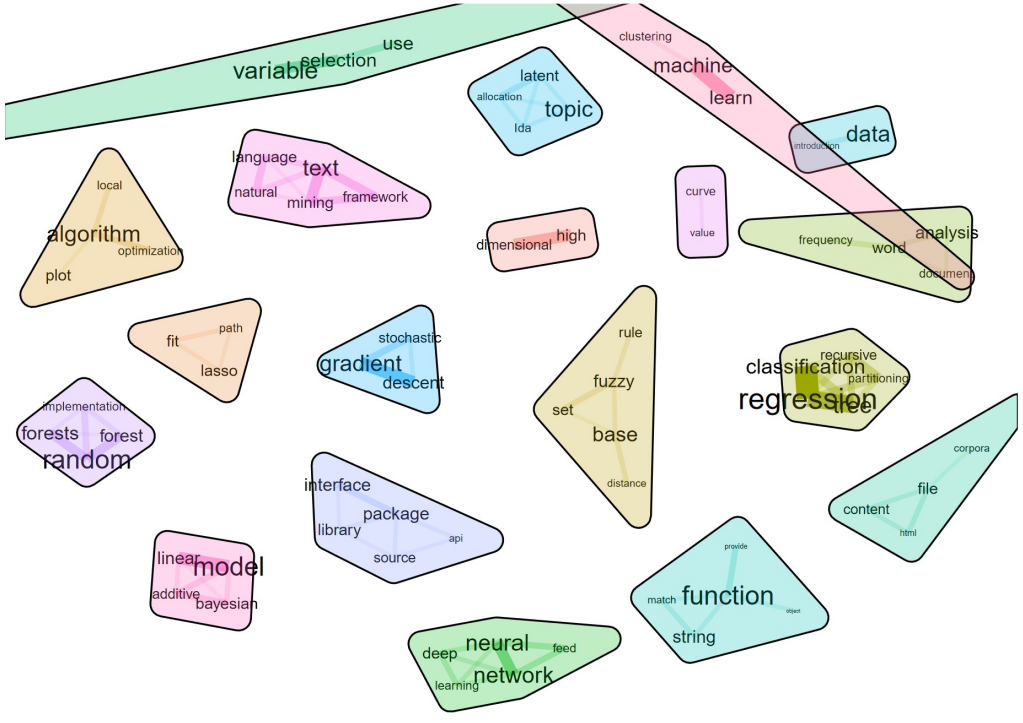
- This example visualises the result of a text annotation which provides parts of speech tags and dependency relationships.



TextPlot

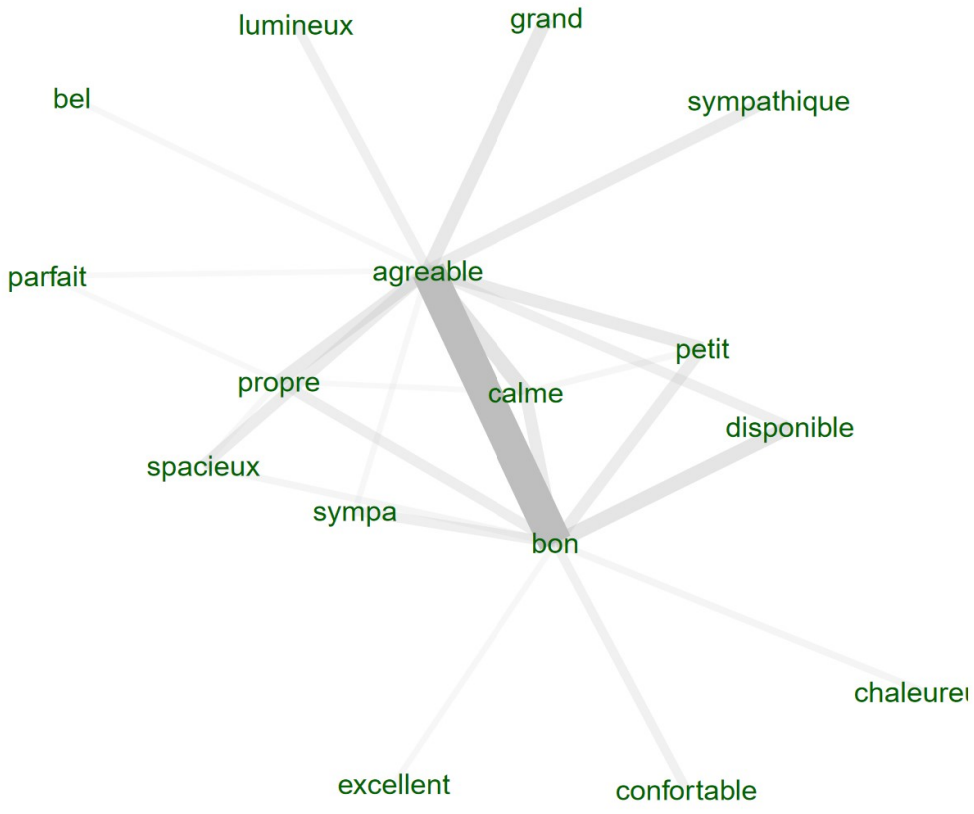
Biterm Topic Model plots

- This example shows plotting a biterm topic model which was pretrained and put in the package as an example.



Co-occurrence of texts

- The following graph shows how frequently adjectives co-occur across all the documents.



corporaexplorer: An R package for dynamic exploration of text collections

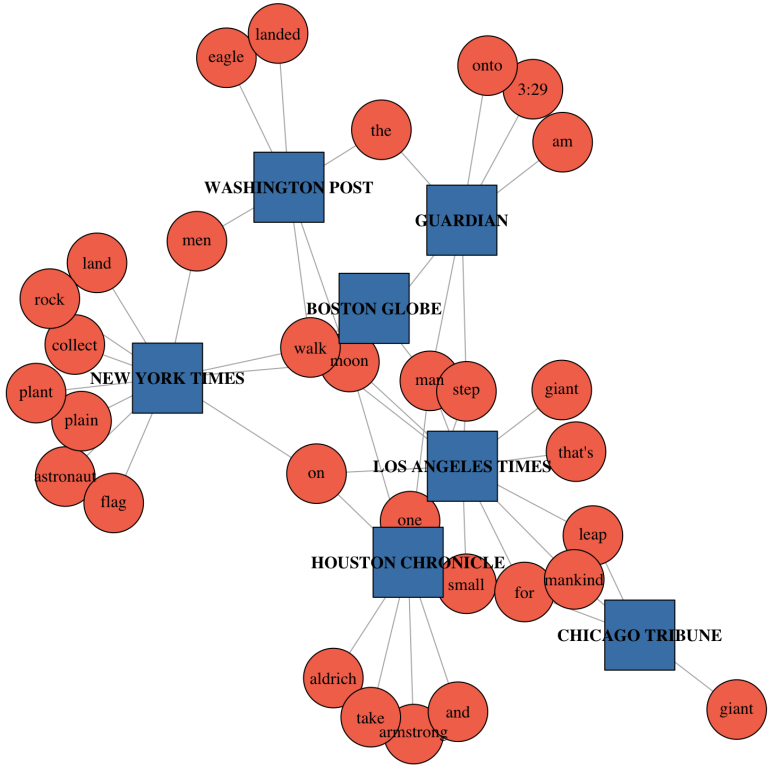
- It is an R package that uses the Shiny graphical user interface framework for dynamic exploration of text collections.
- It's intended primary audience are qualitatively oriented researchers who rely on close reading of textual documents as part of their academic activity.

Visit this link for more information.

textnets

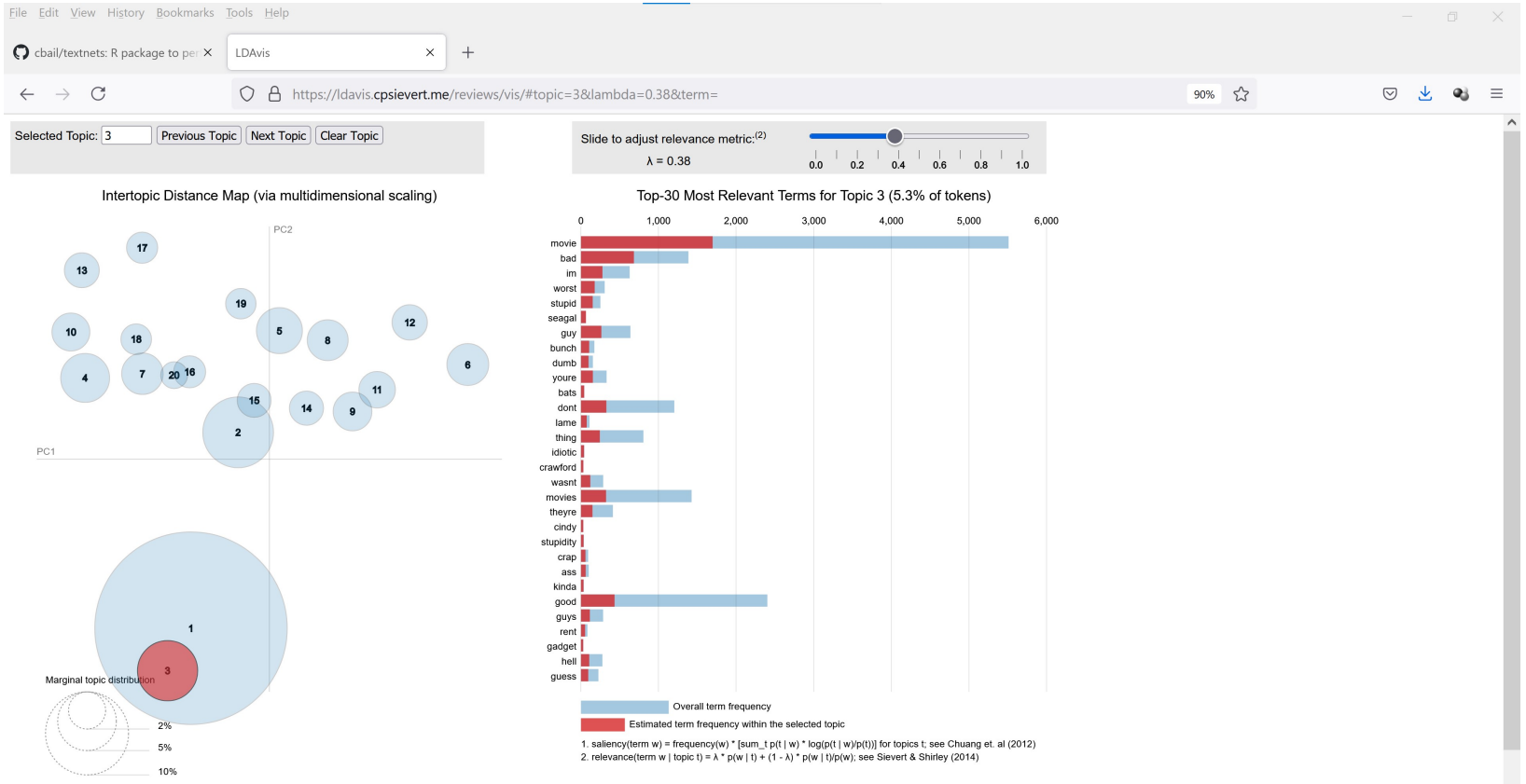
- R package for automated text analysis using network techniques.
- Visit the github repository for more information.
- Notice that this package is not on cran yet. You need to install it by using the code *install_github("cbail/textnets")*.

Reference: Bail, Christopher A. (2016) "Combining Network Analysis and Natural Language Processing to Examine how Advocacy Organizations Stimulate Conversation on Social Media." *Proceedings of the National Academy of Sciences*, 113:42.



LDAvis

- R package for interactive topic model visualization.
- Visit the github repository and cran for more information.



Source: For live demo, visit this link.

References

Cao, Nan and Cui, Weiwei (2016) **Introduction to text visualization**, Springer. This book is available at smu e-collection.