香港浸會大學
HONG KONG BAPTIST UNIVERSITY

新聞系
DEPARTMENT OF JOURNALISM

Text Analysis and Maps

Lectured by: Dr. Xinzhi ZHANG

Research Assistant Professor, Department of Journalism

Hong Kong Baptist University

13 June, 2018

@CUCN Data-driven Journalism Workshop

Agenda

- Text Analysis
  - Issues in text mining
  - Tools
- Plotting Maps
  - Dots
  - Continuous data (choropleth map, heatmap)
  - Tools

Text Mining

- Text mining
- Text analytics

Documents or words

- In text mining, each document is encoded as the **list** of words it contains.
- “Case” = document
- “Variable” = the vocabulary of a document set
- “Value” = the number of times the word appears in the document

Documents or words

- Document 1 (D1): The boy hits the dog.
- Document 2 (D2): The boy hits the dog and the cat.
- Document 3 (D3): The boy likes the cat cat.

	The	boy	hits	dog	and	cat	likes
D1	2	1	1	1	0	0	0
D2	3	1	1	1	1	1	0
D3	2	1	0	0	0	2	1

- Each row = document vector
- All rows = term-document matrix (TDM)
- Individual cell = “term frequency”

”Bag of Words”

- Tokenization
- Tokenization is the process of converting the string into individual words, or “tokens.”

TF-IDF

- There are some “common” words (high frequencies)
- We want to identify the words that can distinguish the document from others.
  - Stopwords: “the” “a” “and”…
  - Common words: if all the documents contain the word “dog” then we cannot tell the differences among these documents based on the word “dog.”

TF-IDF

- Solution: “de-weight” the ”common” words
- Common = appearing in many documents
- DF = “document frequency” = fraction of documents that containing the term
- IDF = “Inverse document frequency” = invert DF and logged

- TF-IDF = TF*IDF

Online tools

- http://www.wordle.net
- Making the word cloud
- Other word cloud tools [Here]

Online tools

- http://voyant-tools.org
- More advanced text analysis
- Tutorial: http://docs.voyant-tools.org/start/

Maps

- The use of map in data-driven journalism is partly driven by the availability of geodata.
- The presentation of maps involves the usage of dots and continuous data on the maps.
- A map can sometimes change the world.

1854, a map produced by Doctor John Snow,.
The map essentially represented each death as a bar. It became apparent that the cases were clustered around the pump in Broad (now Broadwick) street.

Dots on the maps

- Dots can represent occurrences in certain geographical locations.
- Dots can serve like icons (an icon is something looks like another ☺
- Dots can be attached to variables (whose visual attributes can be changed).

Continuous data on the maps

- Choropleth map
  - A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map.
- Heatmap
  - A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors (i.e., the saturation or hue).

Create a map

- Programming languages:
  - R (with the packages: rworldmap)
- Professional tools
  - Carto
  - Geohey (极海)
- A list of online tools can be found here [Link][Resources]