


TEXT MINING

... In a Nutshell

Mengqian LU

Text Mining in a Nutshell

- Documents categorization
 1. Text retrieval
 2. Classification [Supervised] ←
 3. Clustering [Unsupervised]
 - Corpora contrast
 - ❑ Reverse logic of classification
 - ❑ Metrics chosen upon the data and objective
- 

Text Mining in a Nutshell

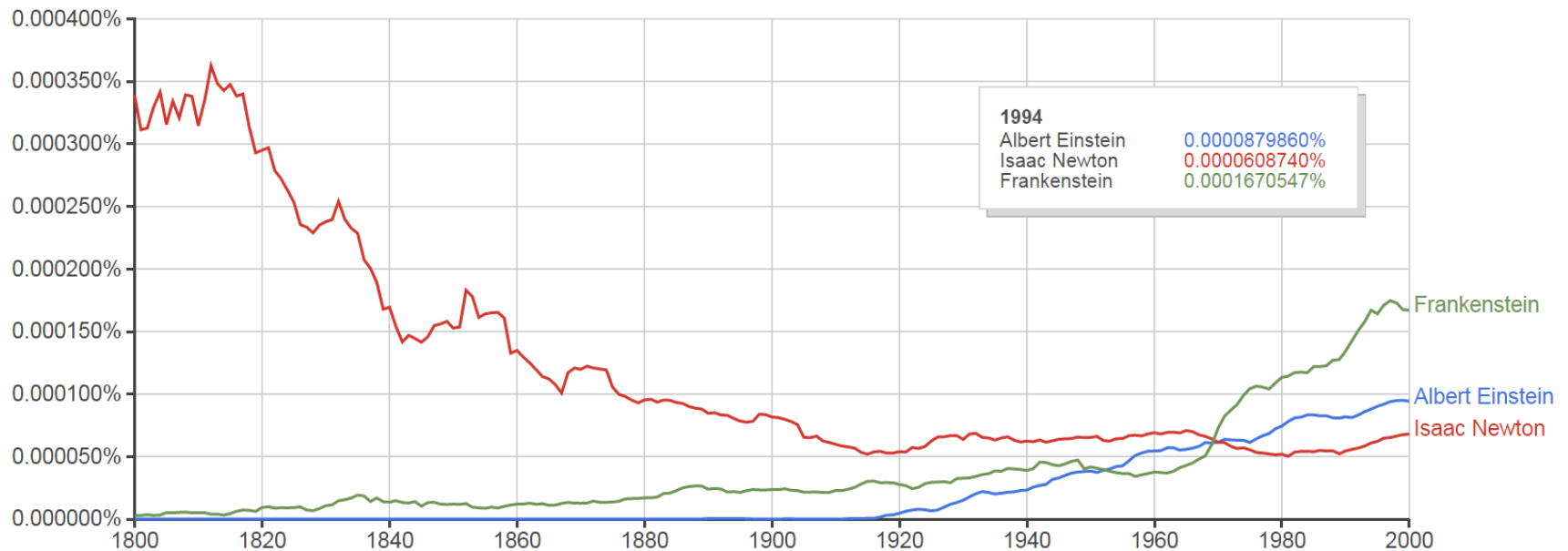
- Temporal structure extraction – “The history of a word”

Google books Ngram Viewer

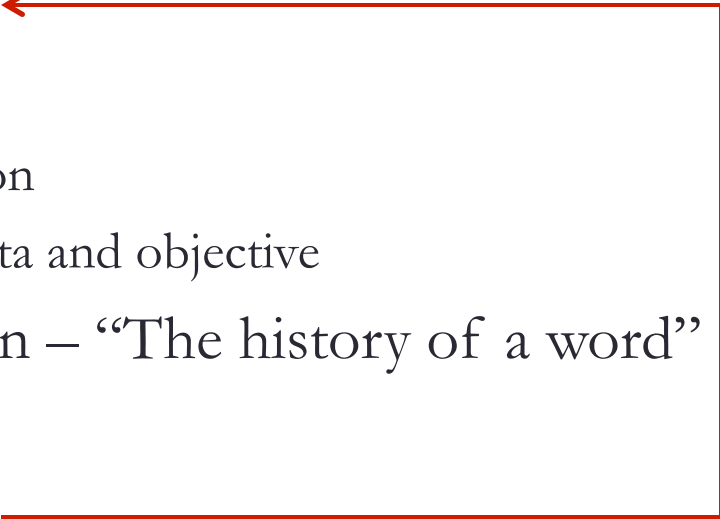
Graph these comma-separated phrases: Albert Einstein,Isaac Newton,Frankenstein ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3

Search lots of books

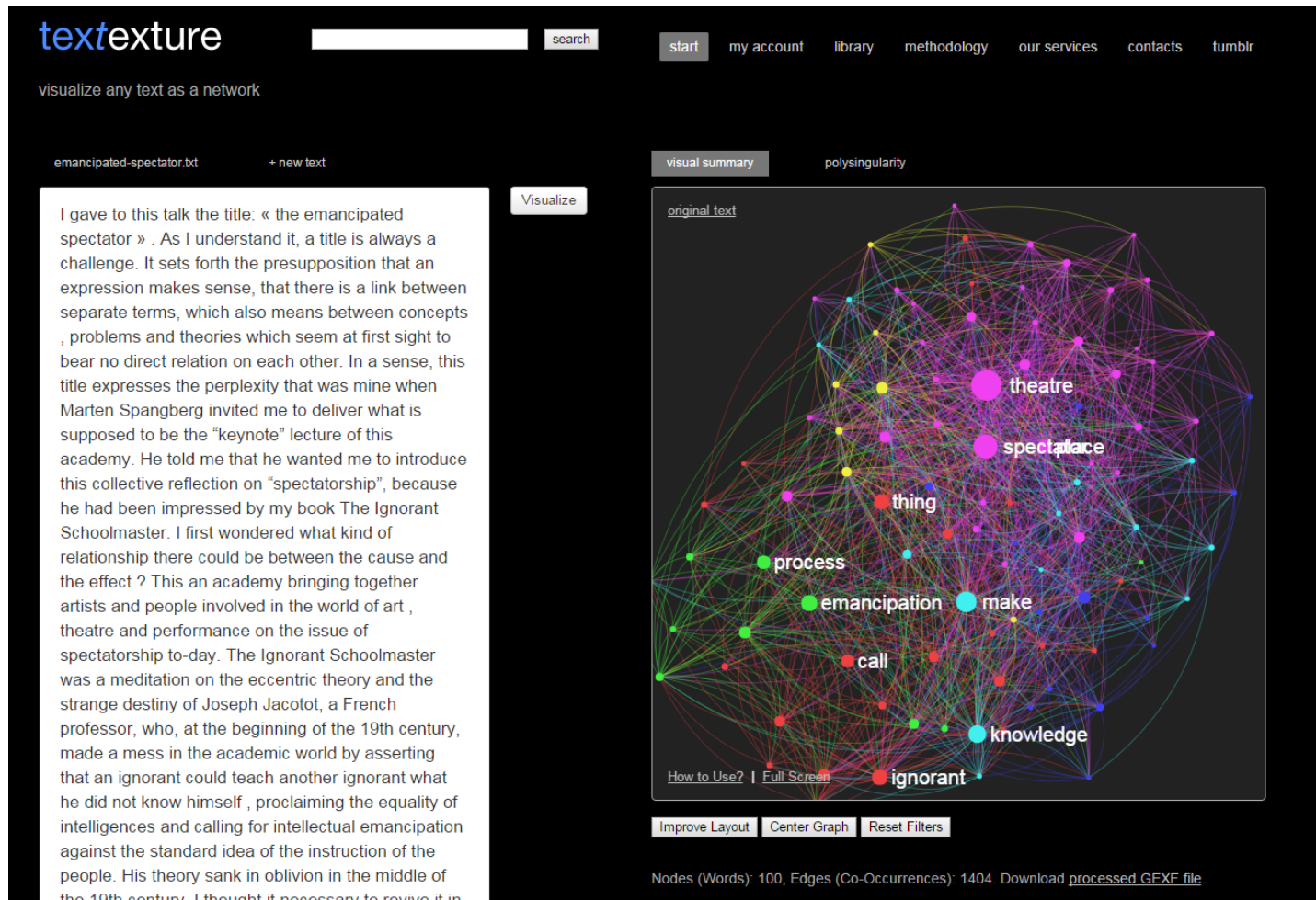


Text Mining in a Nutshell

- Documents categorization
 1. Text retrieval
 2. Classification [Supervised]
 3. Clustering [Unsupervised] ←
 - Corpora contrast
 - ❑ Reverse logic of classification
 - ❑ Metrics chosen upon the data and objective
 - Temporal structure extraction – “The history of a word”
 - Topic modeling
 - ❑ Reverse logic of clustering
 - ❑ Identify a featured semantic map
 - ❑ From “Latent Semantic Analysis” to “Bayesian Topic Modeling”
- 

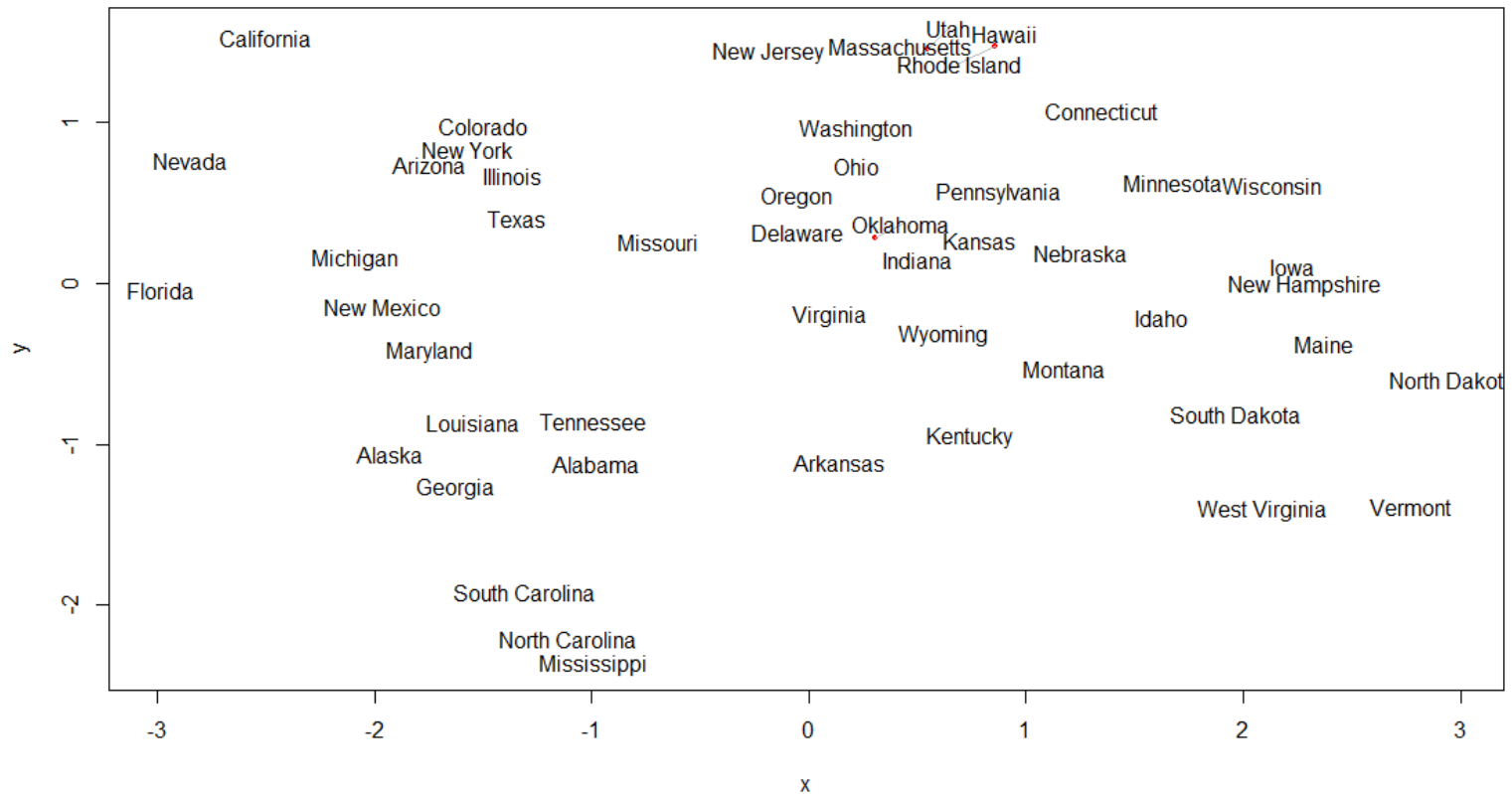
Text Visualization

- Affinity network



Text Visualization

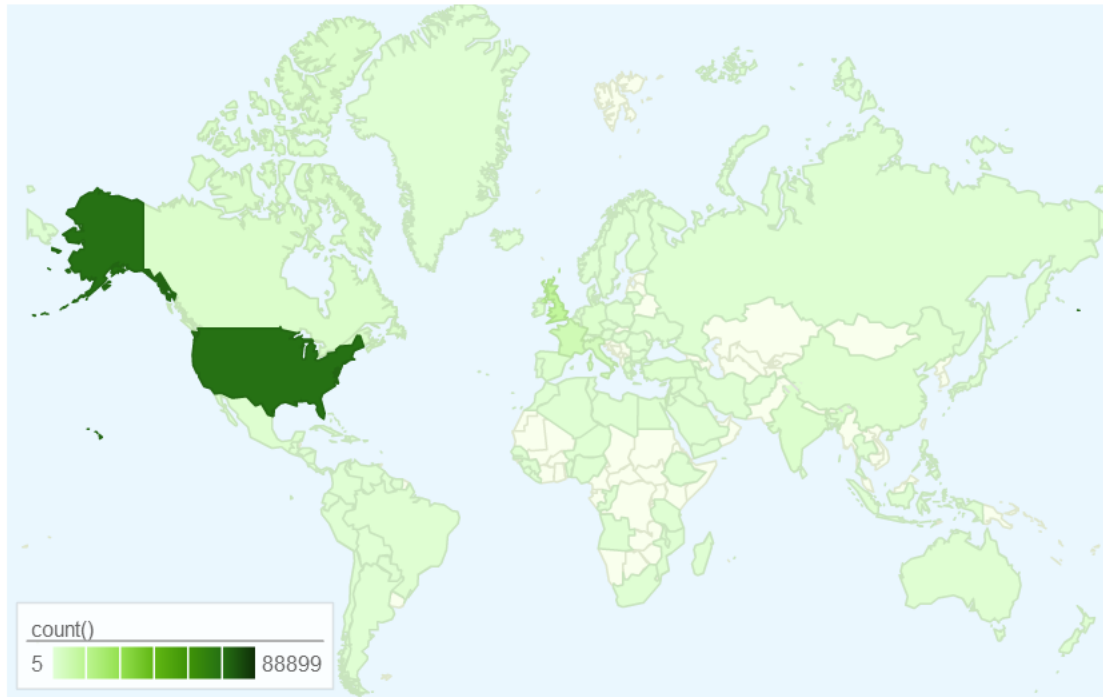
- Dimension reduction – e.g. Principle component analysis



Based on R data “SOTU”

Text Visualization

- Spatial (By Matt Wilken source: <http://www3.nd.edu/~mwilkins/Fusion.html>)



Named locations aggregated by nation, linear density scale.
Mouseover individual nations for raw counts.

Intensity maps of the distribution of named places in 1098 volumes of U.S. fiction dating from 1851-75, aggregated by nation and by U.S. state.



HAVE FUN~

