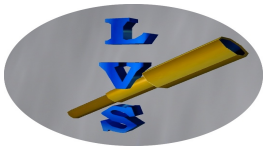
Interactive Visual Data Analysis
Part Two
Interactive Text Mining Suite

Olga Scrivner

Indiana University



Workshop in Methods

Outline

1. Introduce a web application for text processing and mining
2. Learn about natural language processing techniques
3. Develop practical skills

Data Mining

"As our collective knowledge continues to be digitized and stored (...) it becomes more difficult to find and discover what we are looking for." (Blei 2012)

New Ways of Exploring Data Collections

- Word clouds (Vuillemot et al., 2009)

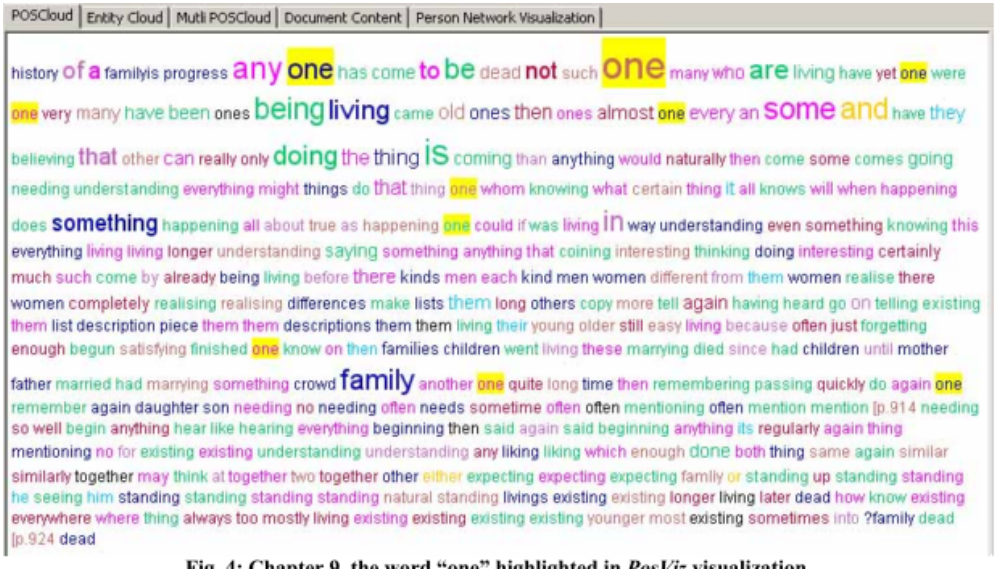
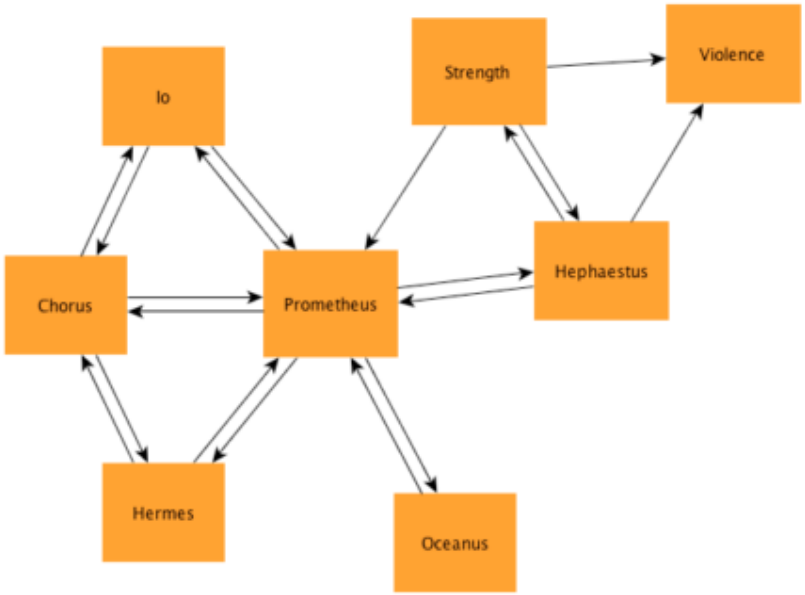


Fig. 4: Chapter 9, the word "one" highlighted in *PosViz* visualization

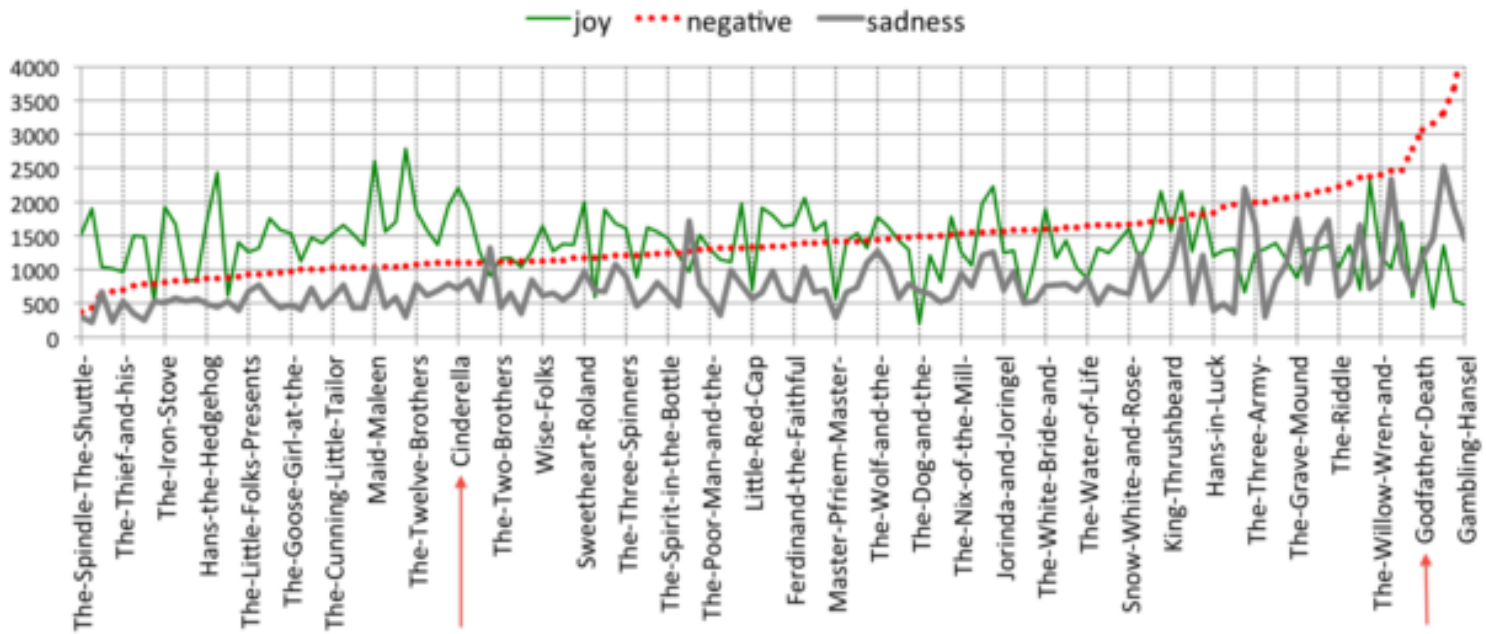
Visualization Methods

- Social network graphs (Rydberg-Cox, 2011)



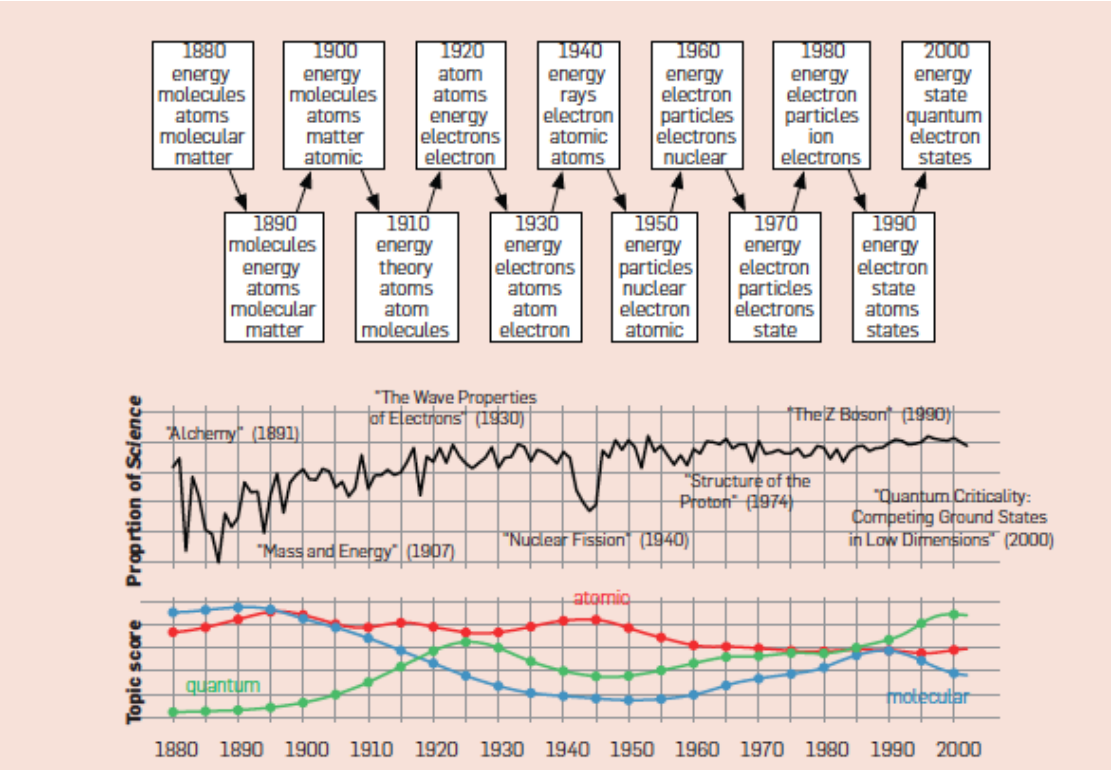
Visualization Methods

- Tracking emotion and sentiment in fairy tales (Mohammad, 2012)



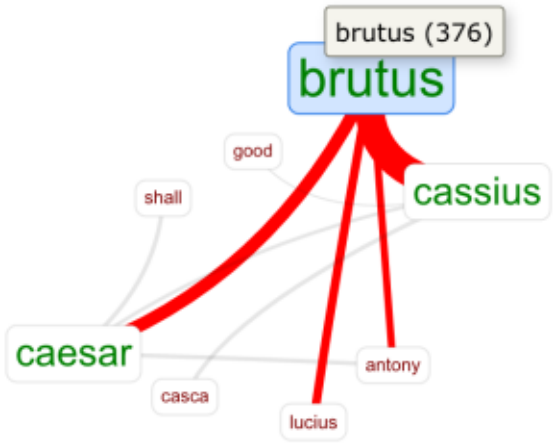
Topic Modeling

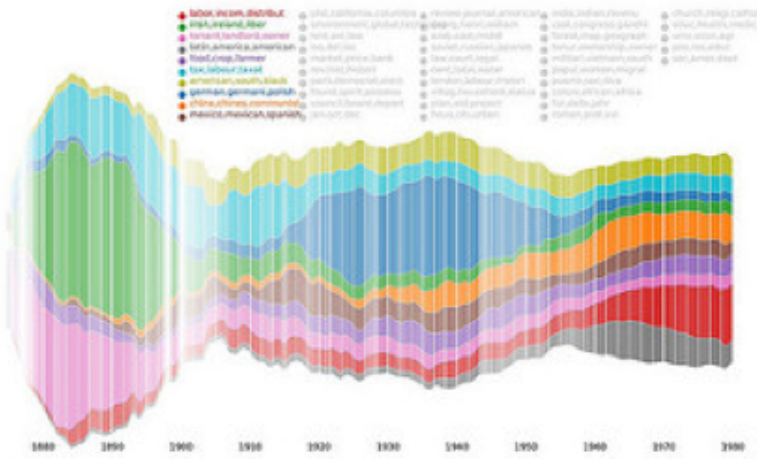
Discovering underlying theme of collection from *Science* magazine 1990-2000 (Blei 2012)



Technological and Methodological Obstacles

- Many tools require some programming skills (Mallet, Meta, R and Python libraries)
- GUI tools are limited to certain formats and functions (Voyant, PaperMachine)
- Lack of active control by users

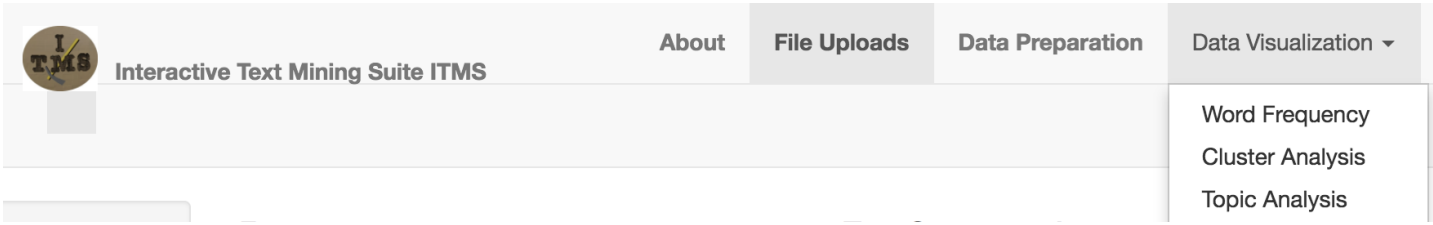




Interactive Text Mining Suite

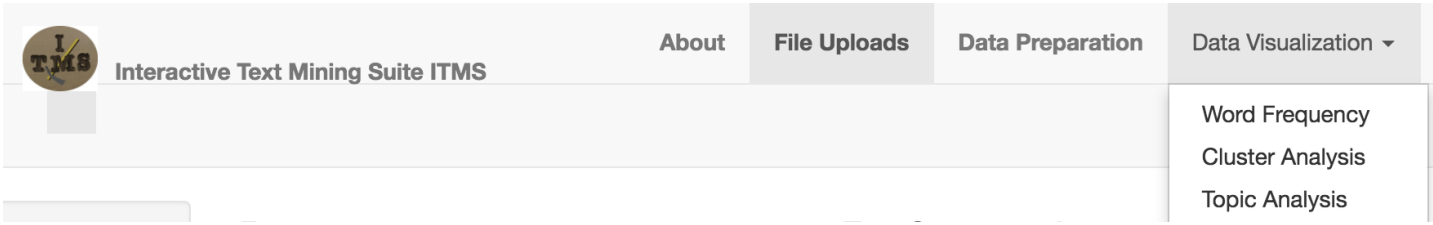
- A user-friendly tool for quantitative analysis and visualization of unstructured data
- Platform-independent
- Interactive

ITMS Structure



1. File Uploads
   - Upload files (txt, pdf, rdf and Google books API)
2. Data Preparation
   - Data preprocessing (stopwords, stemming, metadata)
3. Data Visualization
   - Word frequencies, Cluster analysis and topic modeling

ITMS Structure



1. File Uploads
   - Upload files (txt, pdf, rdf and Google books API)
2. Data Preparation
   - Data preprocessing (stopwords, stemming, metadata)
3. Data Visualization
   - Word frequencies, Cluster analysis and topic modeling

Workshop Files

- Download 3 text files

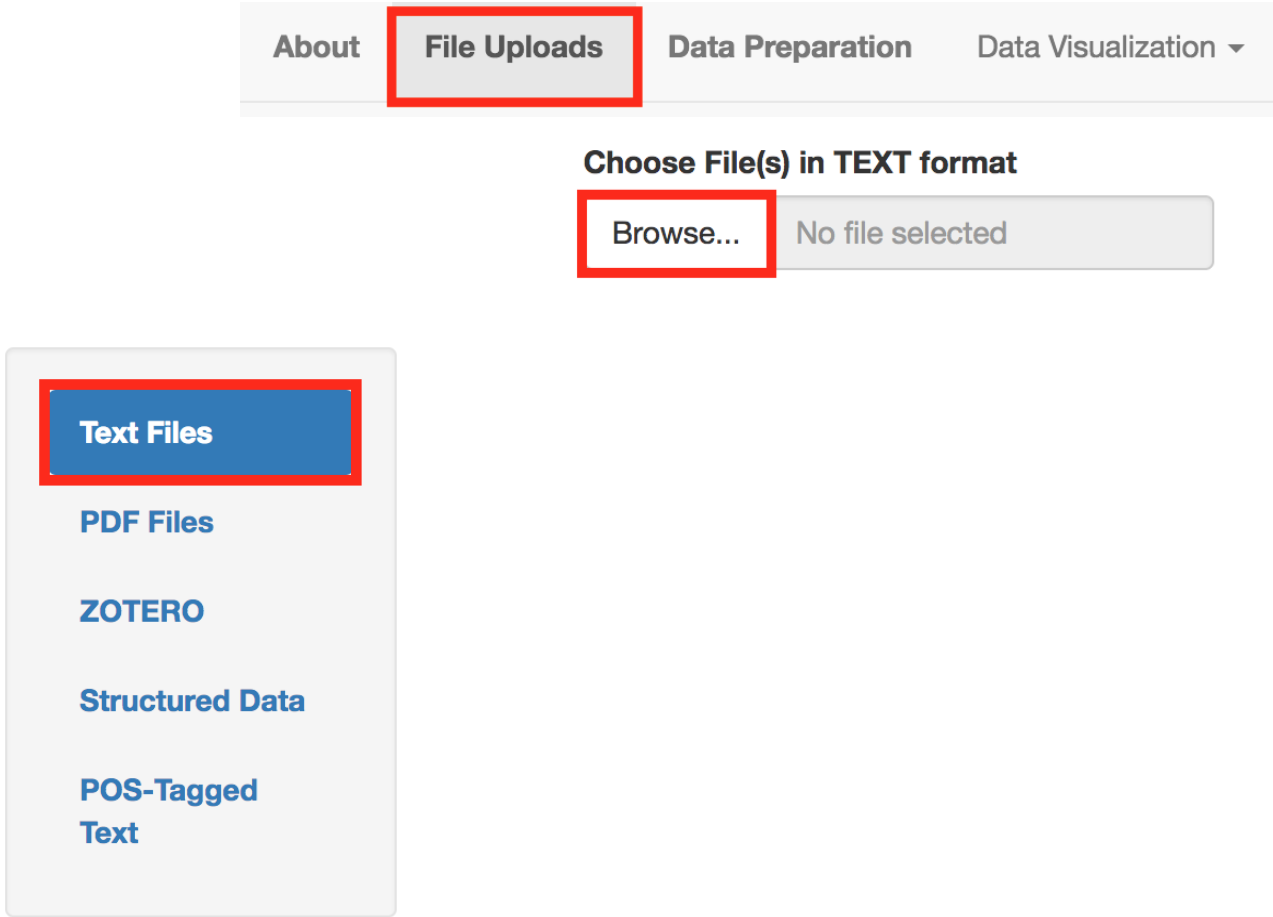
  `http://ssrc.indiana.edu/seminars/wim.shtml`

  - NY Times articles (3 documents in a plain text format)

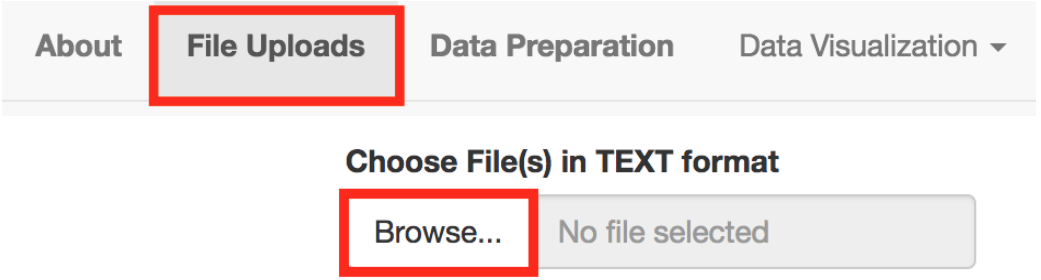
- ITMS Web site:

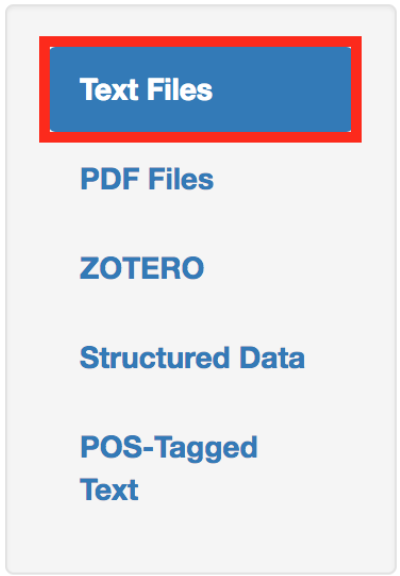
  `http://www.interactivetextminingsuite.com`

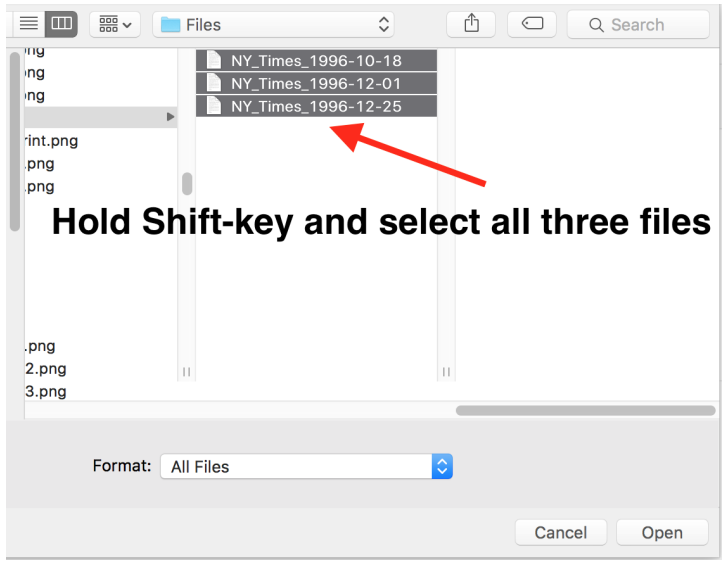
Upload File



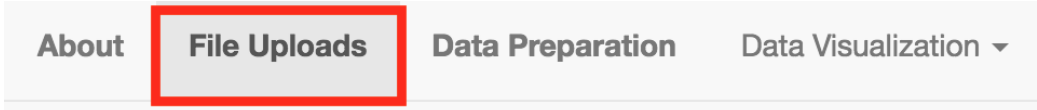
Upload File







Upload File



Choose File(s) in TEXT format

Browse... | No file selected

- **Text Files**
- PDF Files
- ZOTERO
- Structured Data
- POS-Tagged Text

Choose File(s) in TEXT format

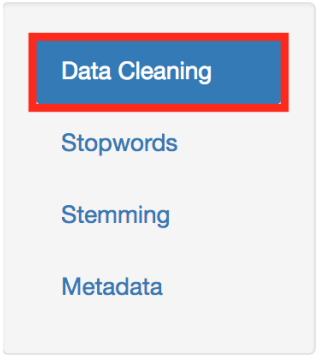
Browse... | 3 files

Upload complete

[1] "NY_Times_1996-10-18.txt"
"NY_Times_1996-12-01.txt" [3]
"NY_Times_1996-12-25.txt"
Corpus Size Total: 1611

Preprocessing Data

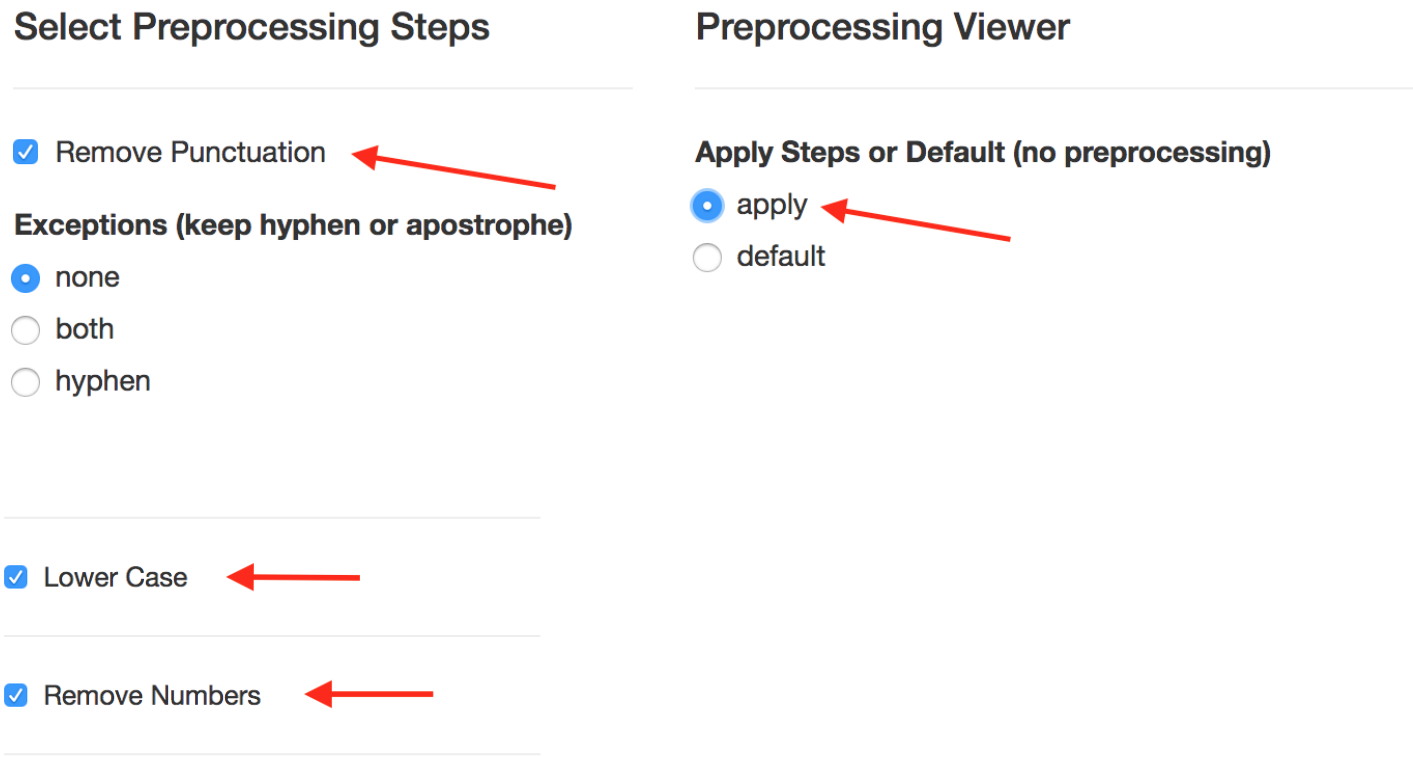
Before performing data analysis we should preprocess data.



Document:
The New York Times October 18, 1996, Friday, Late Edition Final Two Different Pleas for Change: Excerpts From a Second Senate Debate SECTION: Section B; Page 22; Column 1; Metropolitan Desk LENGTH: 1145 words Following are excerpts from last night's debate in Trenton between Robert G. Torricelli, a Democrat, and Richard A. Zimmer, a Republican, as transcribed by The New York Times. Opening Statements TORRICELLI For several months Dick Zimmer and I have been campaigning around New Jersey, asking for your help to get elected to the United States Senate. I know you're disappointed in the campaign. So am I. It's deteriorated into personal accusations and acrimony. But in truth, this campaign isn't about me. And it's not about Dick Zimmer. It's about you, your families and your future. It's also not about taxes or spending. I voted for a tax cut last year. So did Dick Zimmer. I voted for the

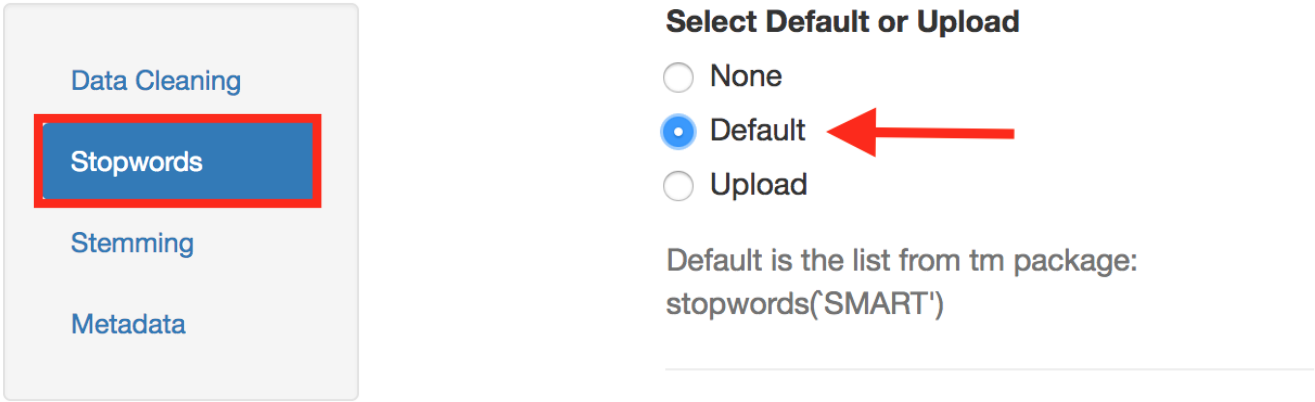
Preprocessing Options

Select preprocessing options and click **apply**.



Stopwords

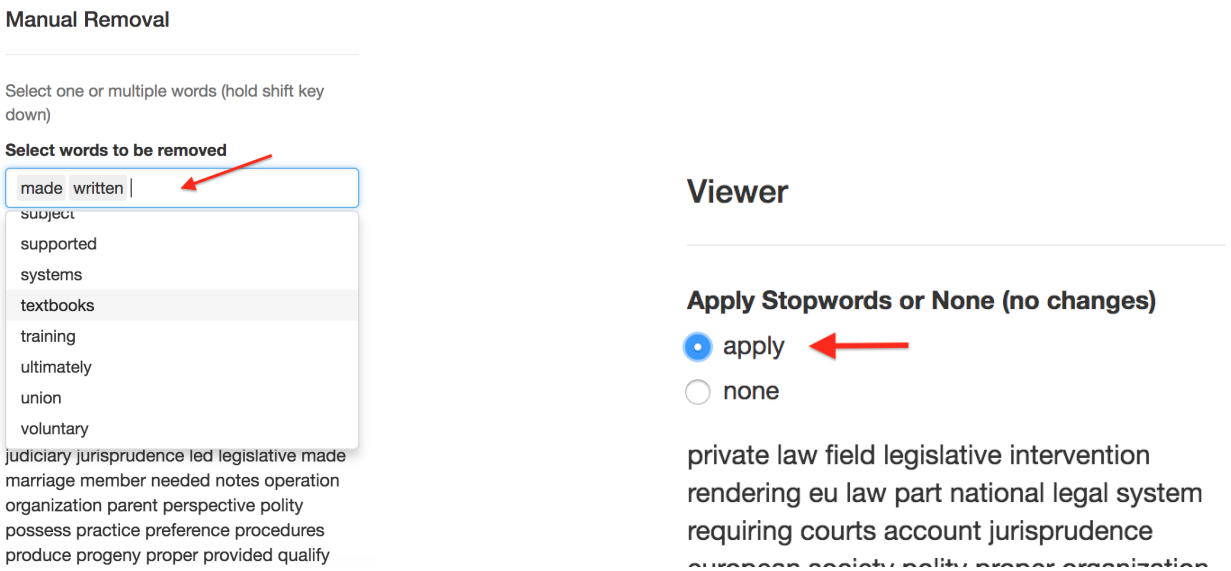
Stopwords (e.g. **the**, **and**): select **Default** for English




```
  [1] "a"           "a's"
  [5] "above"       "accordir
  [9] "actually"    "after"
 [13] "against"     "ain't"
 [17] "allows"      "almost"
 [21] "already"     "also"
 [25] "am"          "among"
 [29] "and"         "another"
 [33] "anyhow"      "anyone"
 [37] "anyways"     "anywhere
 [41] "appreciate"  "appropri
 [45] "around"      "as"
 [49] "asking"      "associat
 [53] "away"        "awfully"
```

Manual Removal of Stopwords

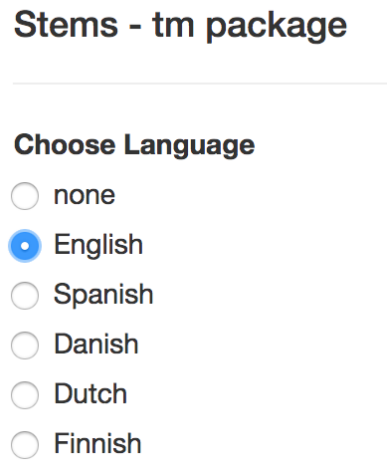
Based on the need, remove any additional stopwords that you may consider a noise, e,g, **paper**, **shows** etc



Select **apply**

Stemming

To improve analytics, you can stem all your tokens, ex. instead of **worked**, **works**, **working**, you will have only one relevant stem **work**

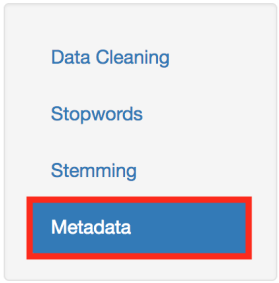


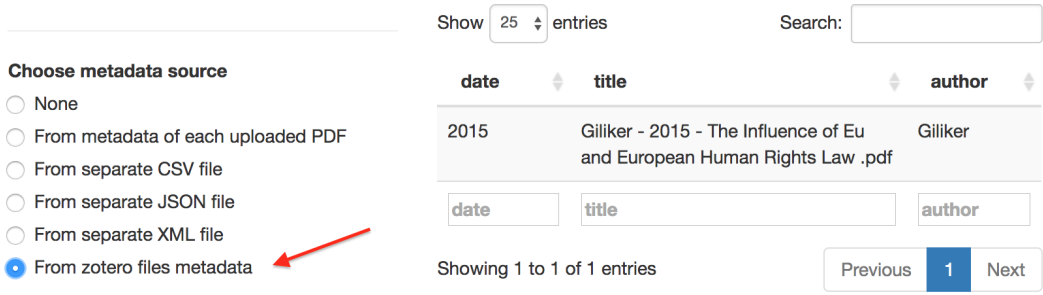
Stem Viewer

privat law field legisl intervent render eu law part nation legal system requir court account jurisprud european societi politi proper organ oper legal system law appli studi sourc led reconfigur common law legal famili parent legal system enter marriag give rise progeni privat european communiti act european union law legal effect nation legal system basi nation court requir appli eu law remain conceptu strike heart domest legal system hold state court subject eu law requir note prefer australia canada zealand legal system analysi support earlier studi year train need qualifi practic favour legal system possess subconsci bias system procedur vis vis member state embed legal system conceiv state complianc eu law voluntari act reli extent echr case law part uk legal system dealt textbook uk academ general access english ultim produc judgment higher qualiti give judiciari perspect legal system court benefit fulli insight provid compar law

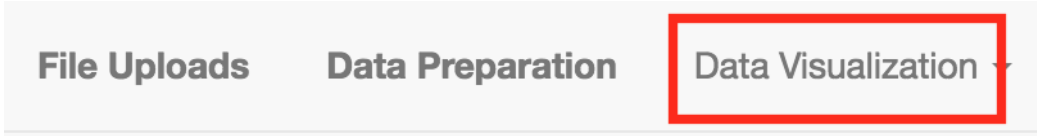
Metadata Extraction

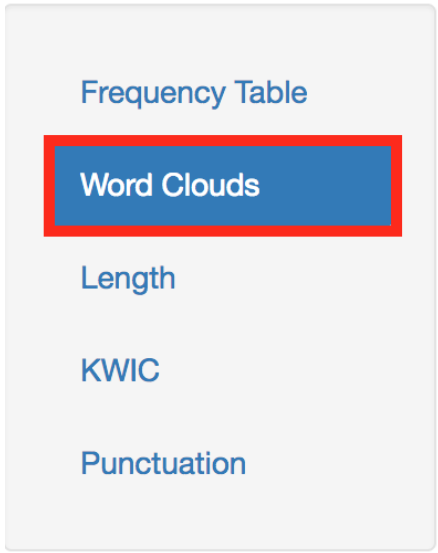
You can extract or upload metadata. You will need datestamp (year) information for chronological topic modeling.

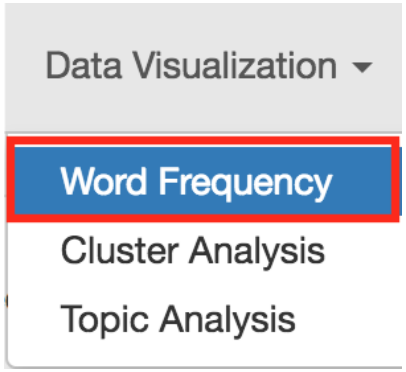




Visualization



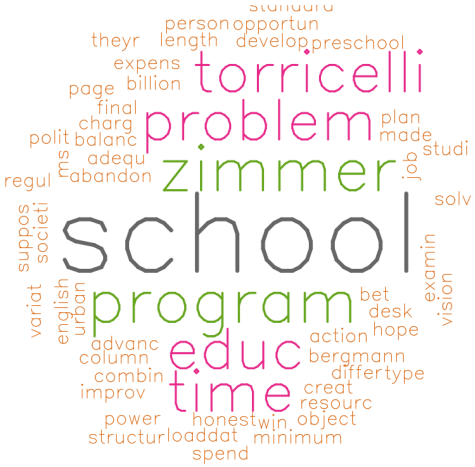




Word Cloud Representation

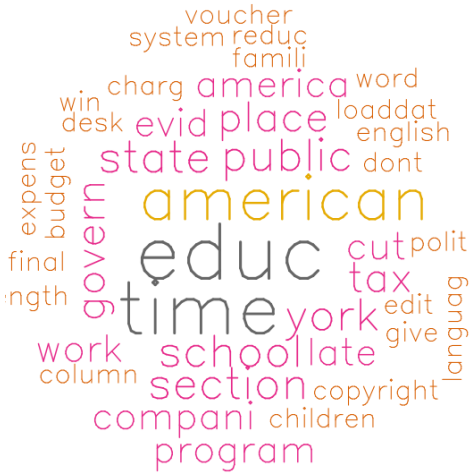
Cloud Type

- (•) Word Cloud
- ( ) Commonality Cloud
- ( ) Comparison of two or more docs



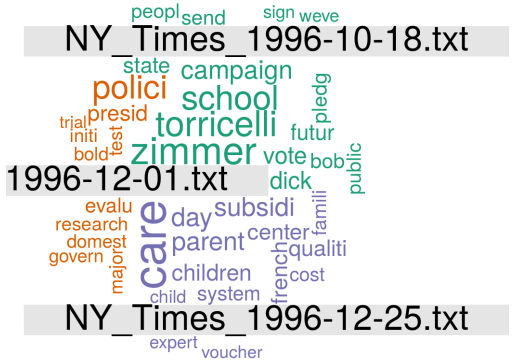
Cloud Type

- ( ) Word Cloud
- (•) Commonality Cloud
- ( ) Comparison of two or more docs



Cloud Type

- ( ) Word Cloud
- ( ) Commonality Cloud
- (•) Comparison of two or more docs



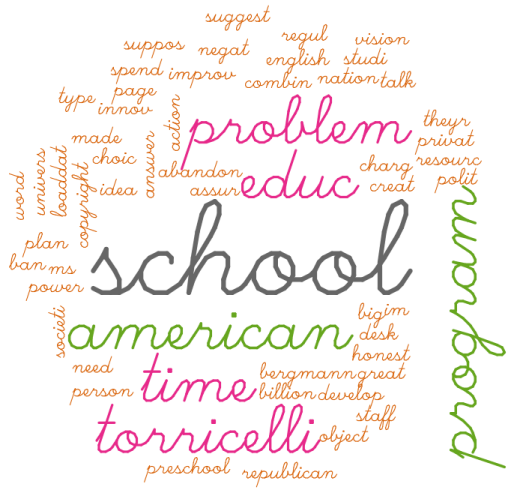
Customization

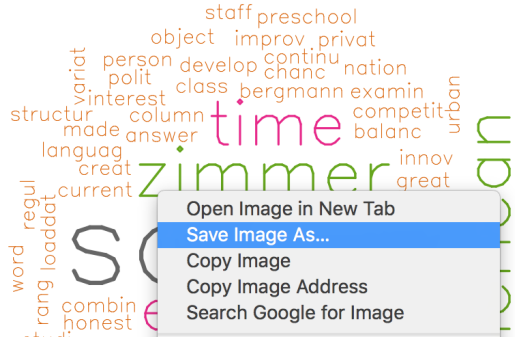
Select Font

- ○ Sans Serif
- ◉ Script
- ○ Gothic

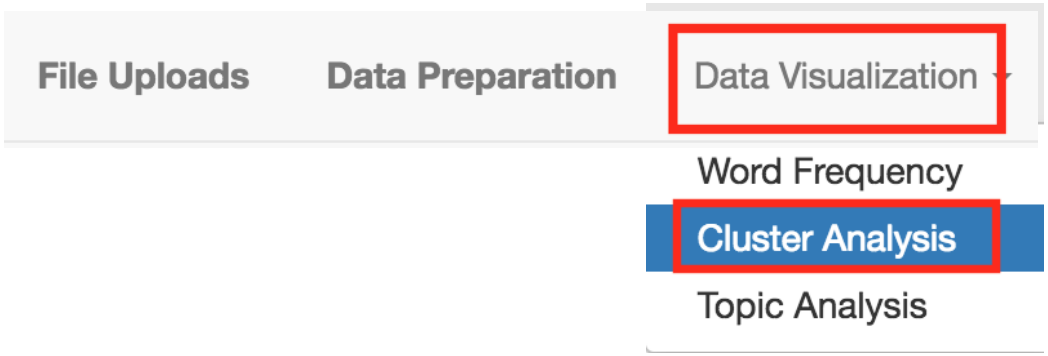
Select Color Palette

- ○ black
- ○ green
- ◉ multi



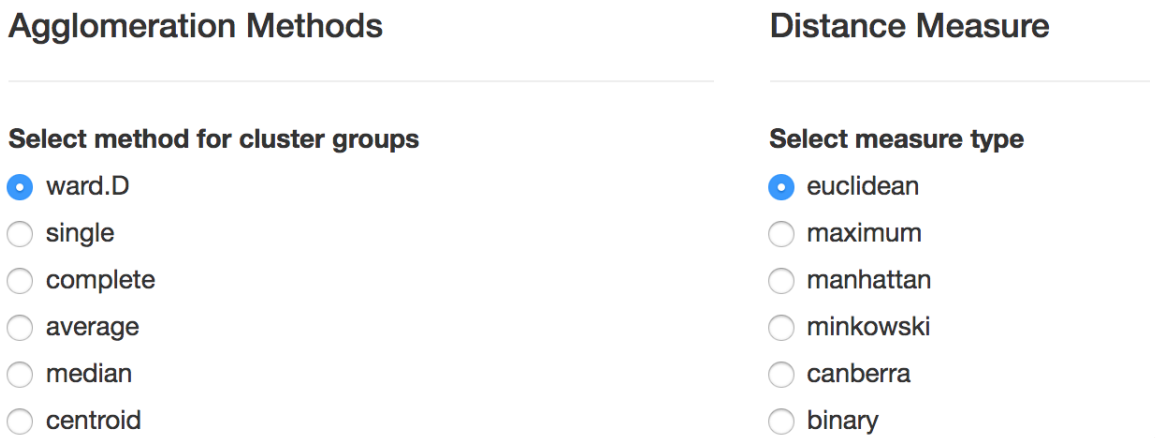


Cluster Analysis

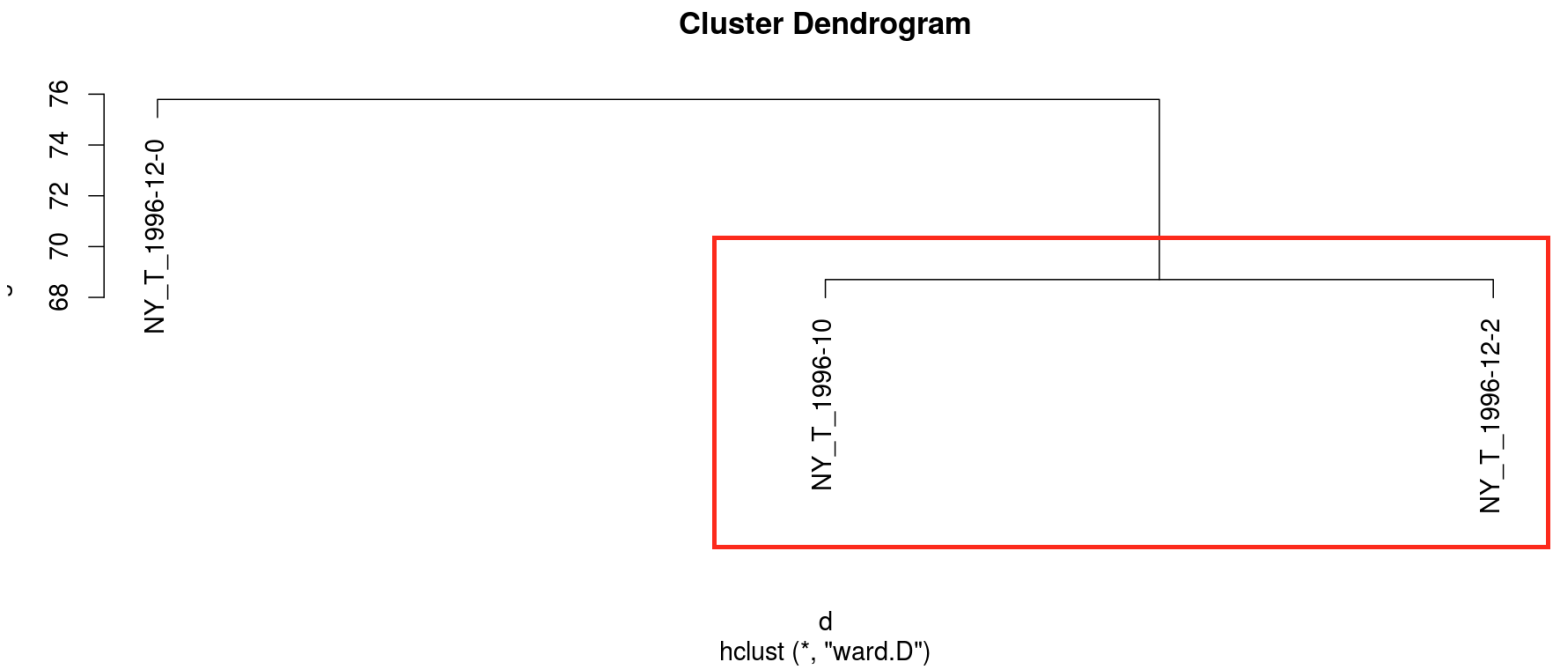


You need to have at least **three** documents

Documents will be grouped based on their term similarity measures



Cluster Analysis



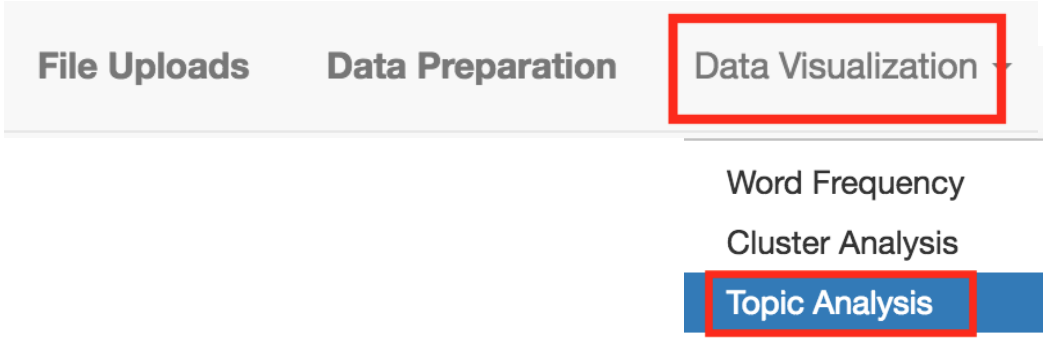
Topic Modeling

- **LDA** (Latent Dirichlet allocation)
- **STM** (Structural Topic model)
- Chronological topic visualization (lda): requires metadata

Topic Modeling Tuning

- Selection of topics (how many different themes)
- Selection of words per theme (how many words per topic)
- Selection of iteration

Topic Model Selection



LDA Topic Model

Model Creation

LDA Visualization

STM Visualization

Metadata Topic Visualization

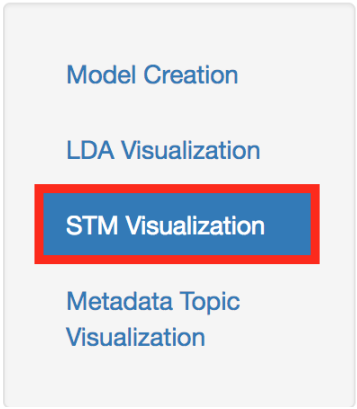
Run LDA Analysis

- ○ none
- ◉ run

Selected Topics LDA (lda.collapsed.gibbs package)

V1	V2	V3
policy	children	public
evidence	care	zimmer
president	vouchers	schools

STM Topic Model



Structural Topics STM

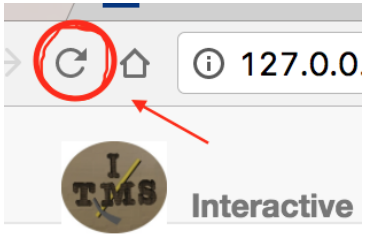
Topic 1:
compelling, definition, physical, choice, decisionmakers, influenced, notion, relationships, shifts, sweeping, trendthe, worknow, concepts, explained, made, behaviour, grows, practice, readings, collection

Topic 2:
changing, covers, early, flexibility, key, magic, organizational, paradigmatic, technology, theories, applies, conceptualize, consumer, internal, investigation, models, perspective, rainbow, topics, vital

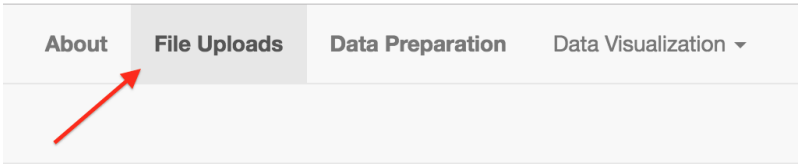
Topic 3:
investigating, adaptation, analyze, barriers, institution, positions, yale, hundred, lens, procedure, capture, central, concern, fixed, notions, shaping, sketches, audience, balance, dispute

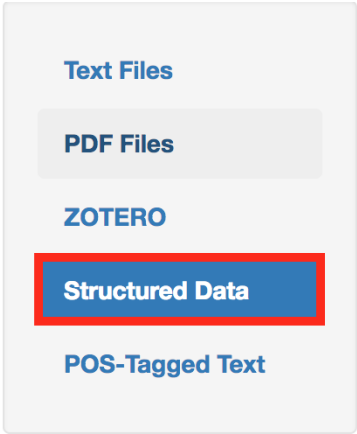
Other Formats - Google Books

Before switching to other data formats, refresh your local browser.



Start with **File Uploads** and select **Structured Data**





Other Formats - Google Books

Select your search terms and submit

Choose file format

- ○ XML
- ○ JSON
- ◉ Google Books Search

Enter your search terms for Google Books, separated by spaces

social science

Submit

Current limitation is 40 books

Show 25 entries Search:

titles	authors	dates	corpus
Readings in the Philosophy of Social Science	Michael Martin, Lee C. McIntyre	1994	Readings in the Philosophy of Social Science the first comprehensive anthology in the

Future Options

Shiny Web Application is highly customizable

1. Part-of-speech tagging (tm package)
2. Network analysis (igraph package)
3. Name Entity Recognition (NLP package)
4. Twitter Streaming (twitterR package) - will requires user's twitter set-up for streaming but information will be provided how to set it up

Open for other suggestions and collaboration - contact `obscrivn@indiana.edu`

Acknowledgements

I would like to thank WIM for providing this opportunity.

Contributors: Jefferson Davis, Irina Trapido, Jay Lee

References I

[1] Many open source R packages: tm, shiny, NLP, stringi, stringr, topicmodels, lda and many more

[2] Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press

[3] Gries, Stefan Th. 2015. *Quantitative designs and statistical techniques*. In Douglas Biber Randi Reppen (eds.), The Cambridge Handbook of English Corpus Linguistics. Cambridge: Cambridge University Press

[4] Jockers, Matthew. 2014. Text Analysis with R for Students of Literature. Quantitative Methods in the Humanities and Social Sciences. Springer International Publishing, Cham

[5] Moretti, Franco. 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. Verso

[6] Oelke, Daniella, Dimitrios Kokkinakis, and Mats Malm. 2012. Advanced visual analytics methods for literature analysis. Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social 561Sciences, and Humanities, pages 35–44
image credits: https://media.giphy.com/media/10zsjaH4g0GgmY/giphy.gif