



HATHI TRUST  
RESEARCH CENTER

# SEASR Analytics for HTRC

---

Loretta Auvil  
University of Illinois

# What is SEASR?

---

This project focus on

- developing,
- integrating,
- deploying, and
- sustaining

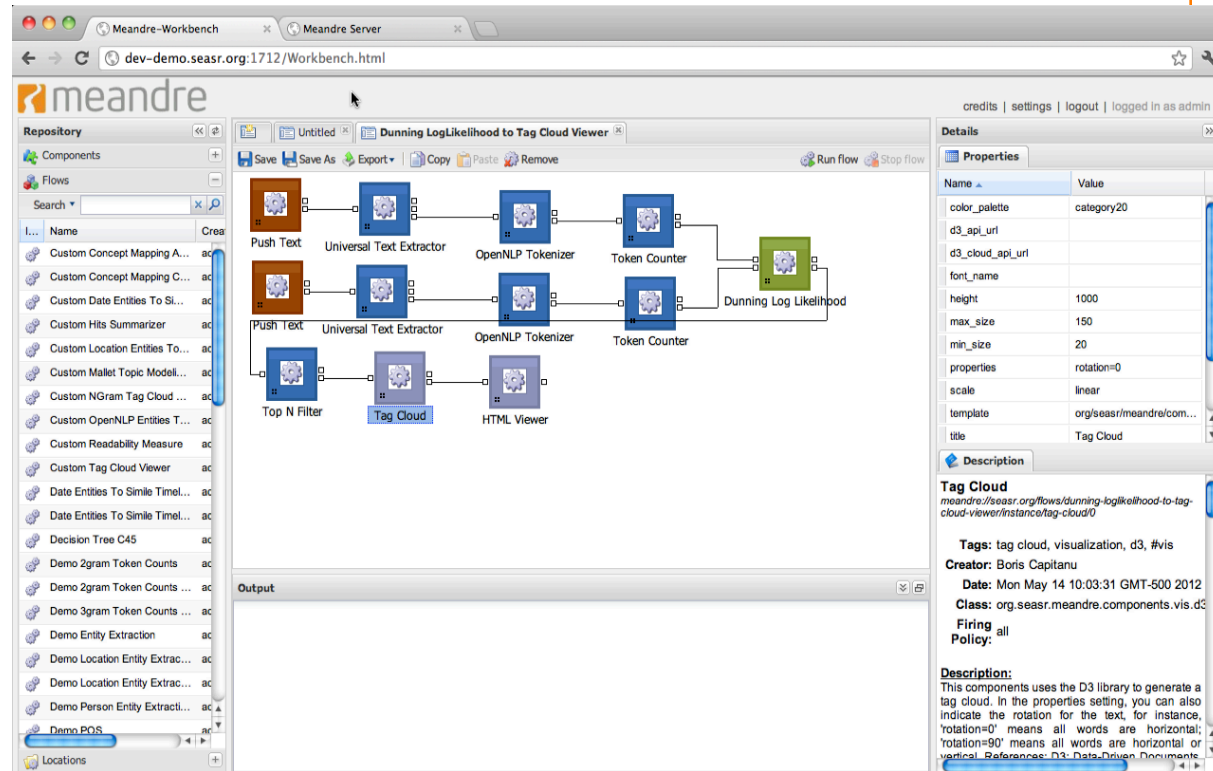
a set of reusable and expandable software components and a supporting framework,

to benefit a broad set of data mining applications for scholars in humanities.



# Meandre: Workbench Existing Flow

- Web-based UI
- Components and flows are retrieved from server
- Additional locations of components and flows can be added to server
- Create flow using a graphical drag and drop interface
- Change property values
- Execute the flow



The screenshot displays the Meandre Workbench interface. The main workspace shows a flow diagram titled "Dunning LogLikelihood to Tag Cloud Viewer". The flow consists of several components: "Push Text", "Universal Text Extractor", "OpenNLP Tokenizer", "Token Counter", "Dunning Log Likelihood", "Top N Filter", "Tag Cloud", and "HTML Viewer". The "Tag Cloud" component is highlighted, and its properties are shown in the right-hand panel.

**Properties Panel:**

Name	Value
color_palette	category20
d3_api_url	
d3_cloud_api_url	
font_name	
height	1000
max_size	150
min_size	20
properties	rotation=0
scale	linear
template	org/seasr/meandre/com...
title	Tag Cloud

**Description Panel:**

**Tag Cloud**  
meandre/seasr.org/flows/dunning-loglikelihood-to-tag-cloud-viewer/instance/tag-cloud/0

**Tags:** tag cloud, visualization, d3, #vis  
**Creator:** Boris Capitano  
**Date:** Mon May 14 10:03:31 GMT-0500 2012  
**Class:** org.seasr.meandre.components.vis.d3...  
**Firing Policy:** all

**Description:**  
This component uses the D3 library to generate a tag cloud. In the properties setting, you can also indicate the rotation for the text, for instance, "rotation=0" means all words are horizontal; "rotation=90" means all words are horizontal or vertical. [Reference: D3.js Data-Driven Documents](#)



# Outline

---

- Dunning Loglikelihood Comparison
- Entity Extraction
- Topic Modeling
- Spell Checking



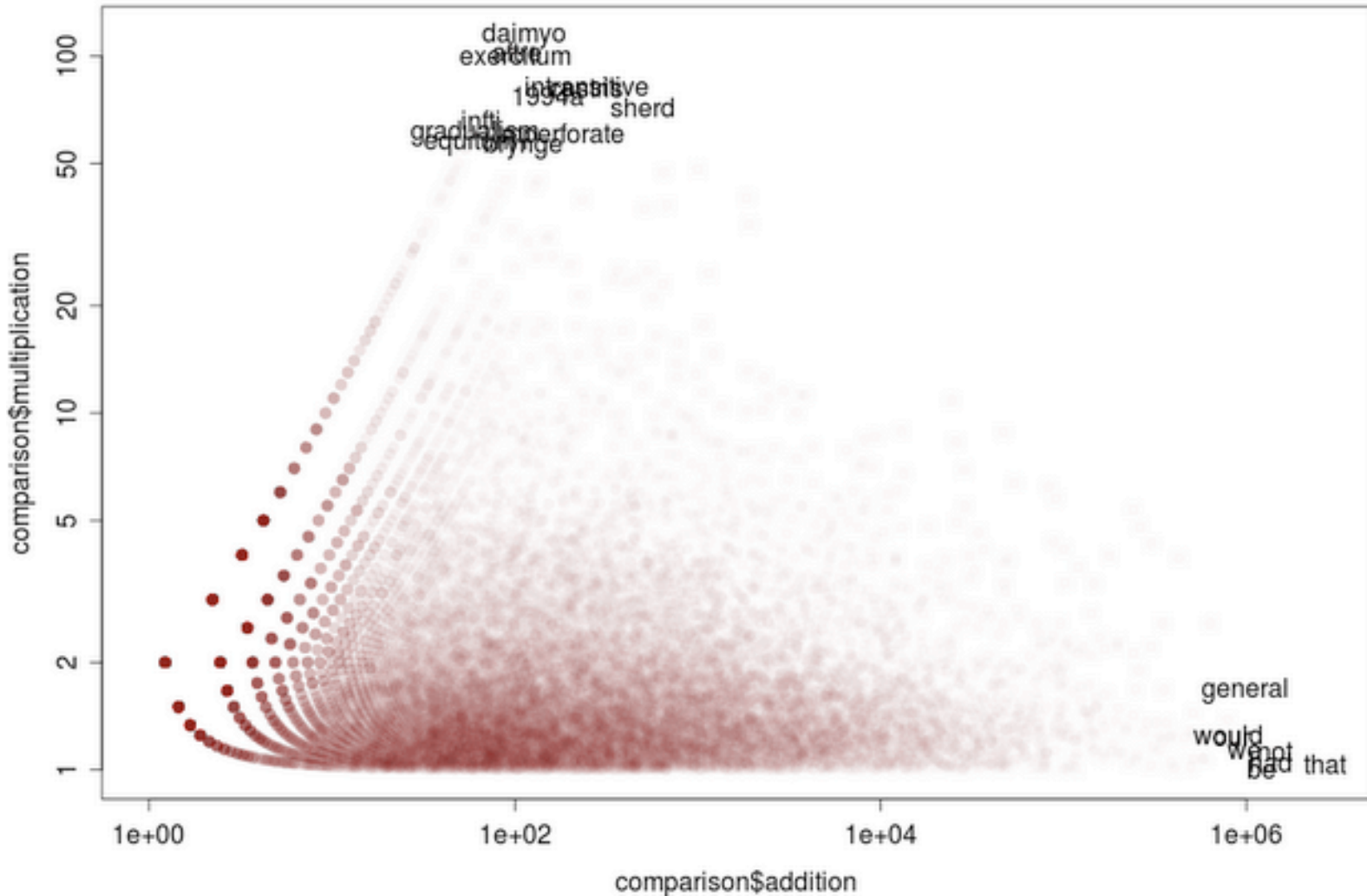
---

# Dunning Loglikelihood Comparison



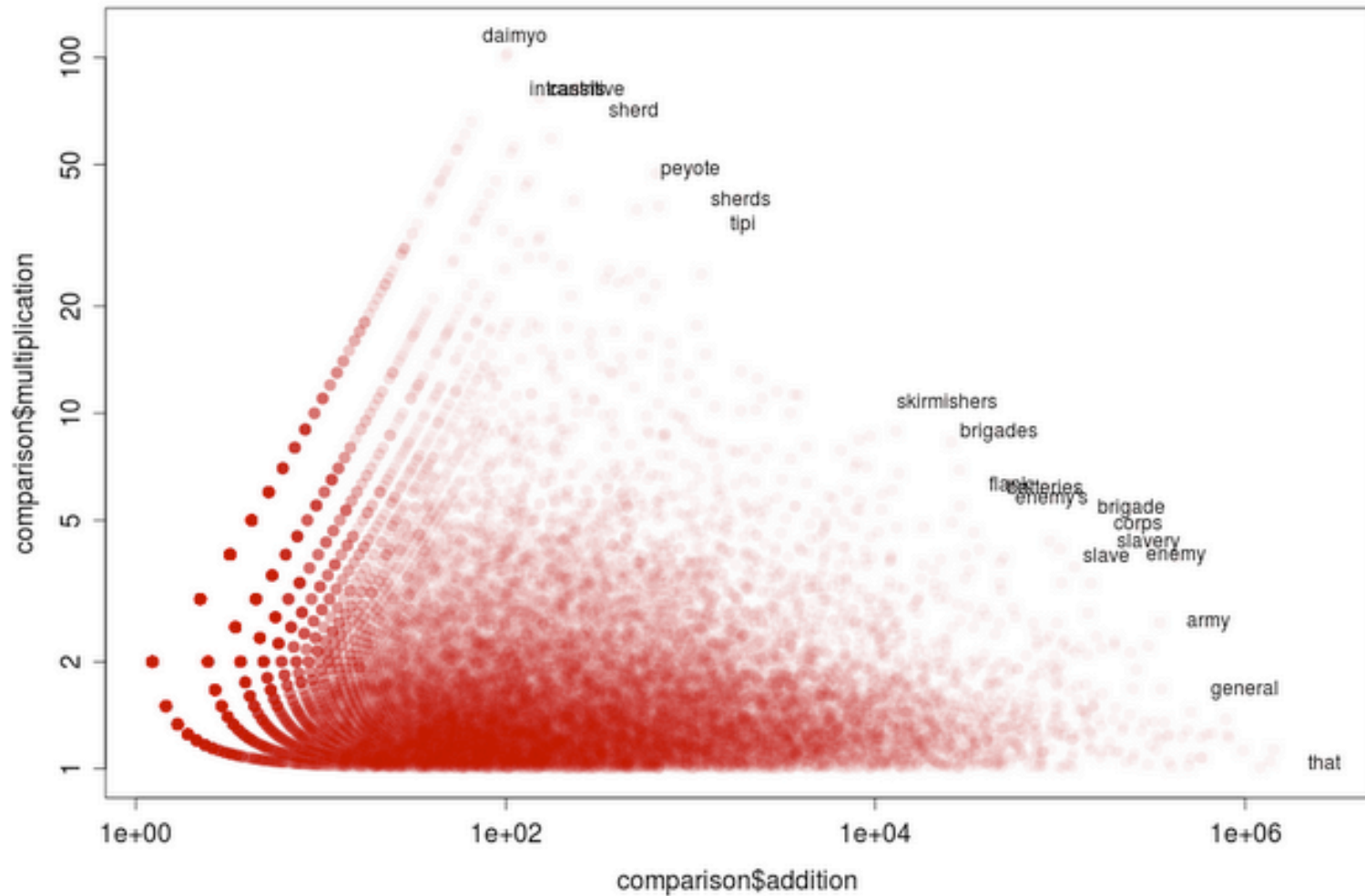
# Comparison of Documents

Over-representation of words in E  
Over F, by multiplication and addition



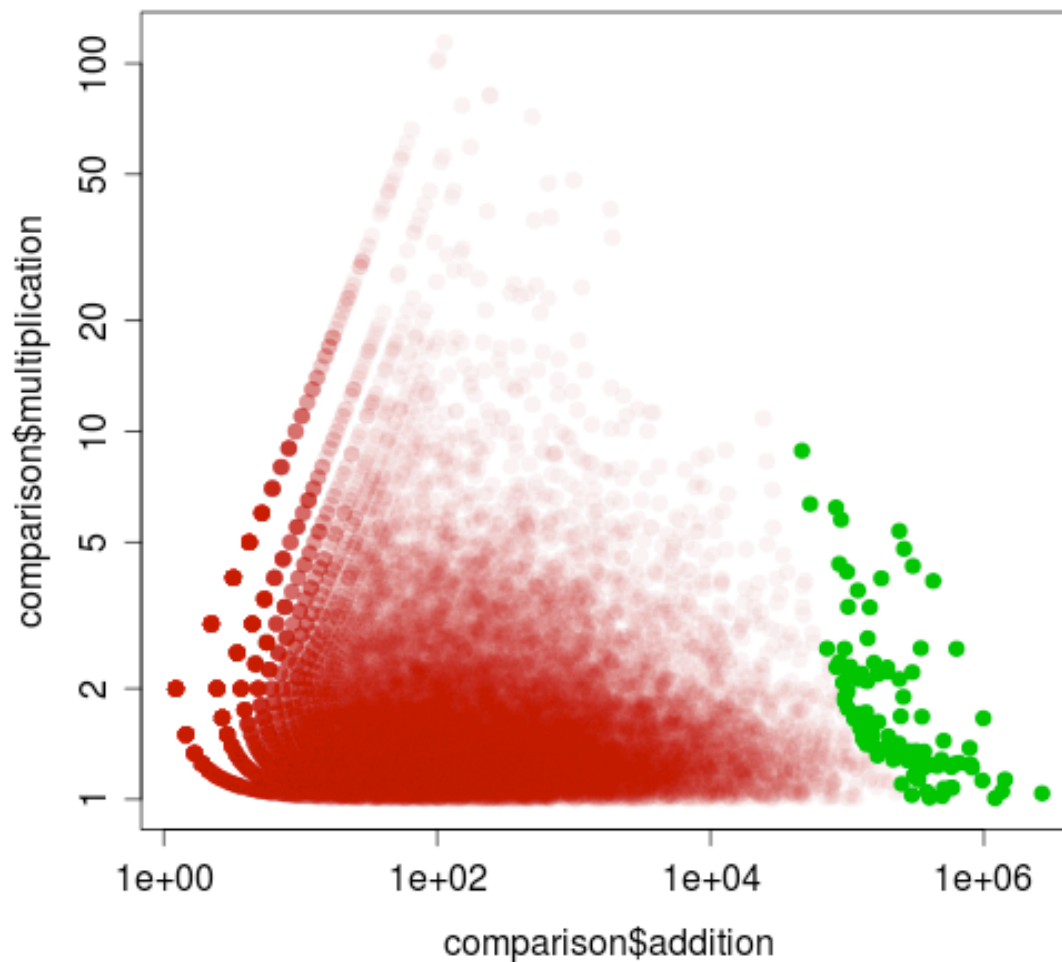
# Comparison of Documents

Over-representation of words in E  
Over F, by multiplication and addition



# Comparison with Dunning Loglikelihood

Same plot, Dunning's Points in Green

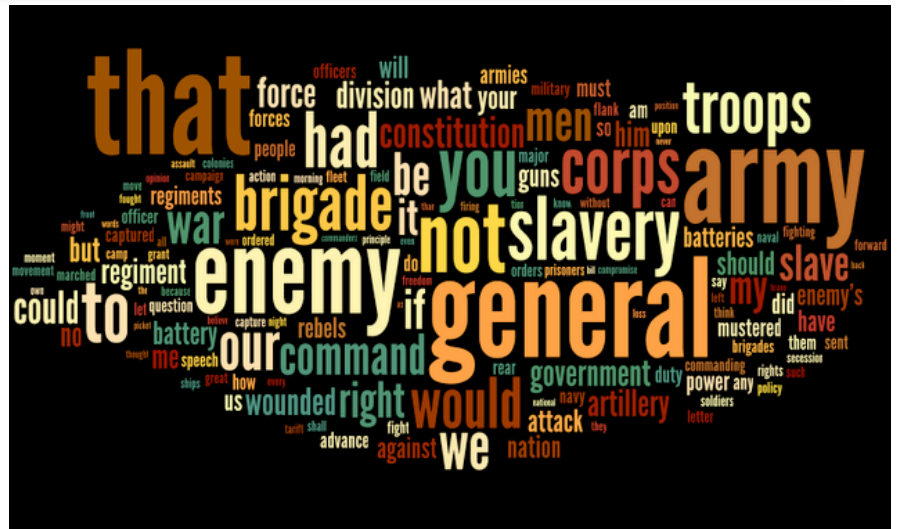




# Dunning Loglikelihood Tag Clouds

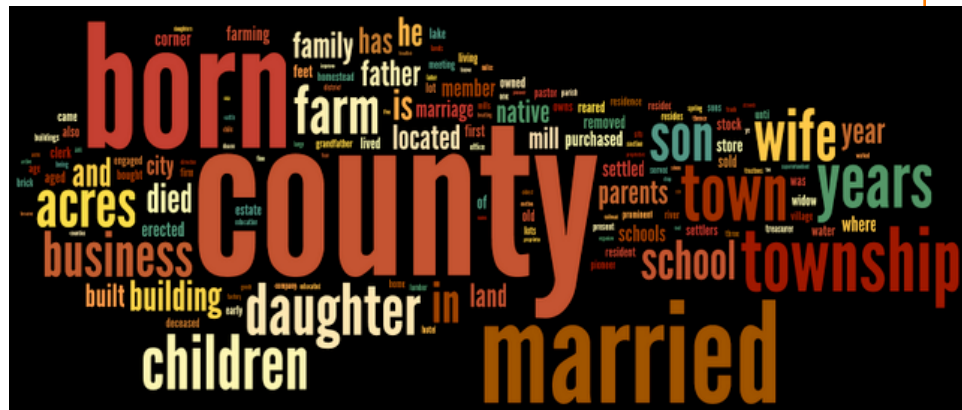
## Significantly overrepresented in E, in order:

- "that" "general" "army" "enemy"
- "not" "slavery" "to" "you"
- "corps" "brigade" "had" "troops"
- "would" "our" "we" "men"
- "war" "be" "command" "if"
- "slave" "right" "it" "my"
- "could" "constitution" "force" "what"
- "wounded" "artillery" "division" "government"



## Significantly overrepresented in F, in order:

- "county" "born" "married" "township"
- "town" "years" "children" "wife"
- "daughter" "son" "acres" "farm"
- "business" "in" "school" "is"
- "and" "building" "he" "died"
- "year" "has" "family" "father"
- "located" "parents" "land" "native"
- "built" "mill" "city" "member"







# Dunning Loglikelihood Comparison

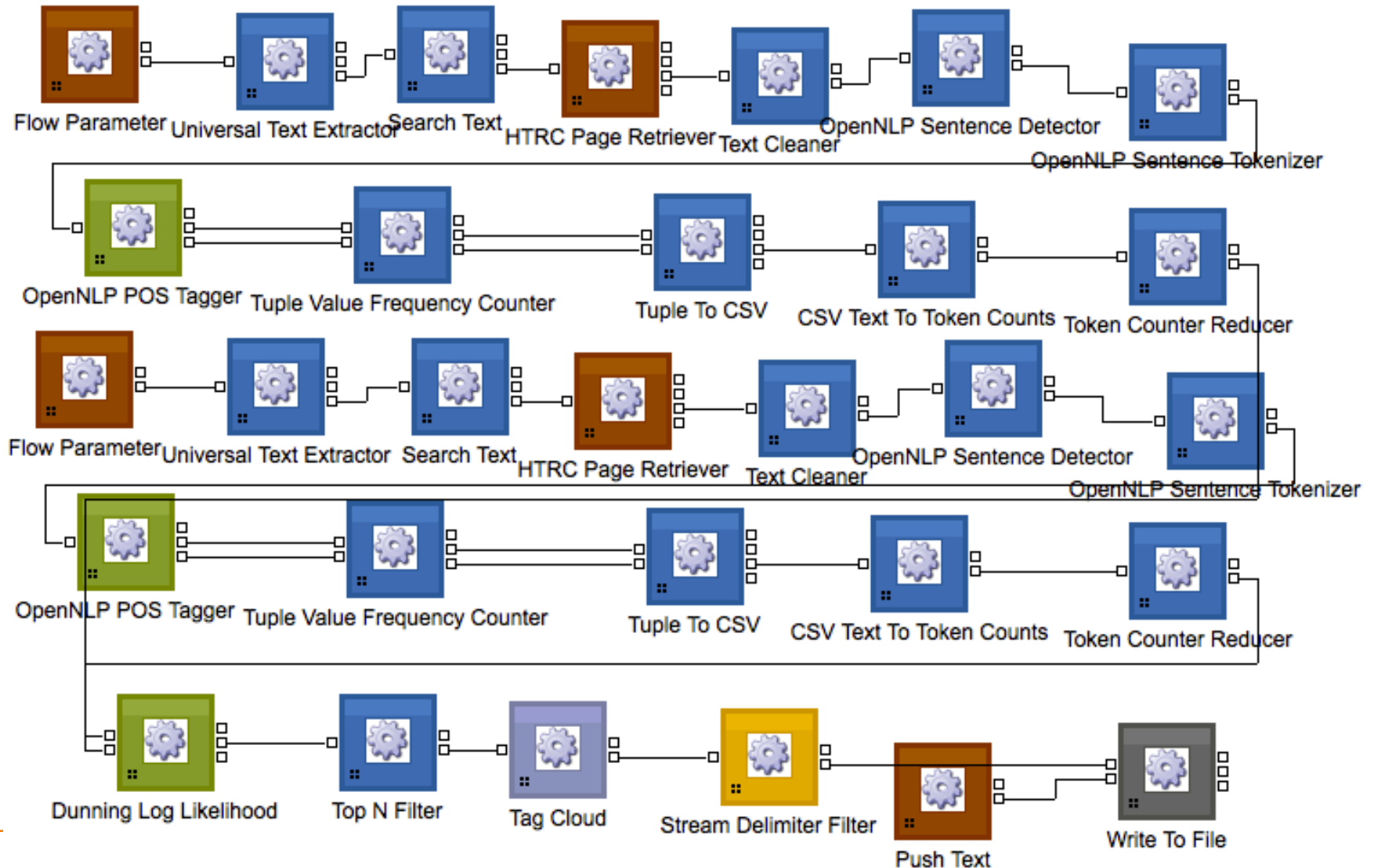
**Wordhoard:** This comparison highlights the lemmata that are disproportionately common or rare in Othello with respect to Shakespeare's tragedies as a whole.

Lemma	Relative use	Log likelihood ▼	Analysis parts per 10,000	Reference parts per 10,000	Analysis count	Reference count
she	+	141.0 ****	104.90	43.42	276	1,002
moor	+	98.4 ****	22.04	3.25	58	75
i	+	64.3 ****	445.04	343.39	1,171	7,925
we	-	61.7 ****	37.63	77.73	99	1,794
lieutenant	+	55.8 ****	11.02	1.34	29	31
handkerchief	+	54.9 ****	10.64	1.26	28	29
our	-	40.9 ****	19.76	44.11	52	1,018
willow	+	34.9 ****	6.84	0.82	18	19
it	+	32.6 ****	194.21	146.72	511	3,386
honest	+	32.0 ****	15.96	5.20	42	120
do	+	28.7 ****	148.22	109.54	390	2,528
her	+	27.7 ****	43.71	24.61	115	568
oh	+	25.3 ***	58.15	36.44	153	841
think	+	24.9 ***	32.68	17.29	86	399
wife	+	24.5 ***	15.96	6.15	42	142





# Meandre Flow



# Dunning Loglikelihood Calculations

	Analysis	Reference	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

$$E1 = c*(a+b)/(c+d)$$

$$E2 = d*(a+b)/(c+d)$$

$$G^2 = 2*((a*\ln(a/E1)) + (b*\ln(b/E2)))$$

For example, a log-likelihood value of 6.63 should occur by chance only about one in a hundred times. This means the significance of a  $G^2$  value of 6.63 is 0.01 .

$G^2$	Significance
15.13	$p < 0.0001$
10.83	$p < 0.001$
6.63	$p < 0.01$
3.84	$p < 0.05$



# Dunning Loglikelihood References

---

- <http://gautam.lis.illinois.edu/monkmiddleware/public/analytics/dunnings.html>
- <http://wordhoard.northwestern.edu/userman/analysis-comparewords.html>
- <http://sappingattention.blogspot.com/2011/10/comparing-corpus-by-word-use.html>
- <http://bookworm.culturomics.org/>
- <http://acl.ldc.upenn.edu/J/J93/J93-1003.pdf>
- <http://tdunning.blogspot.com/2008/03/surprise-and-coincidence.html>





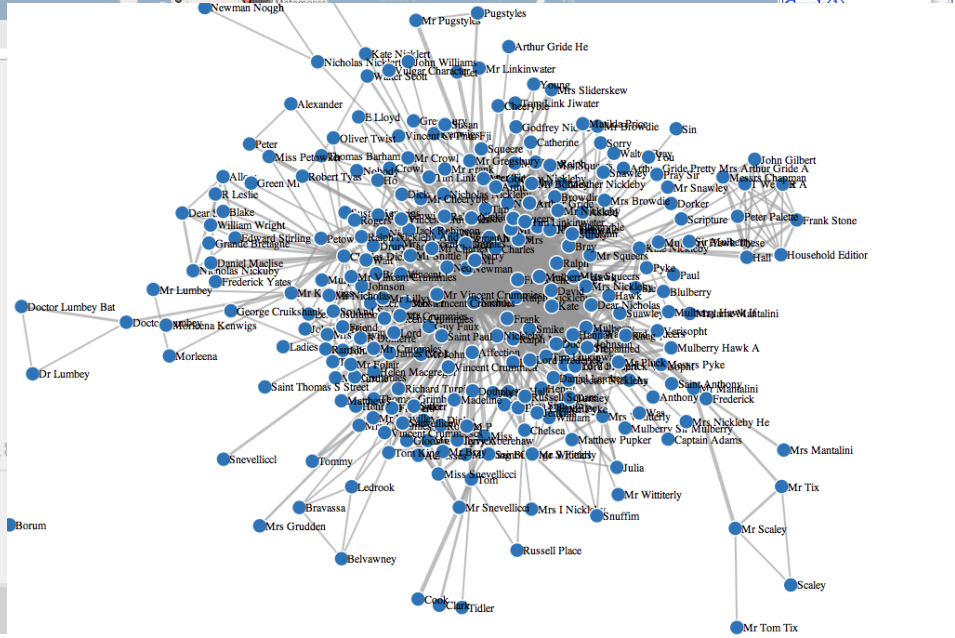
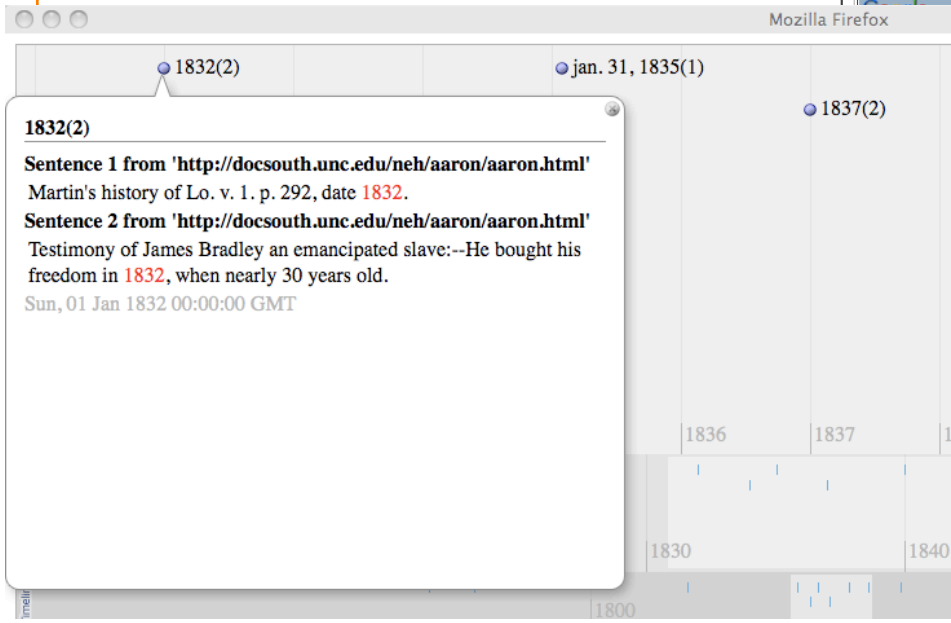
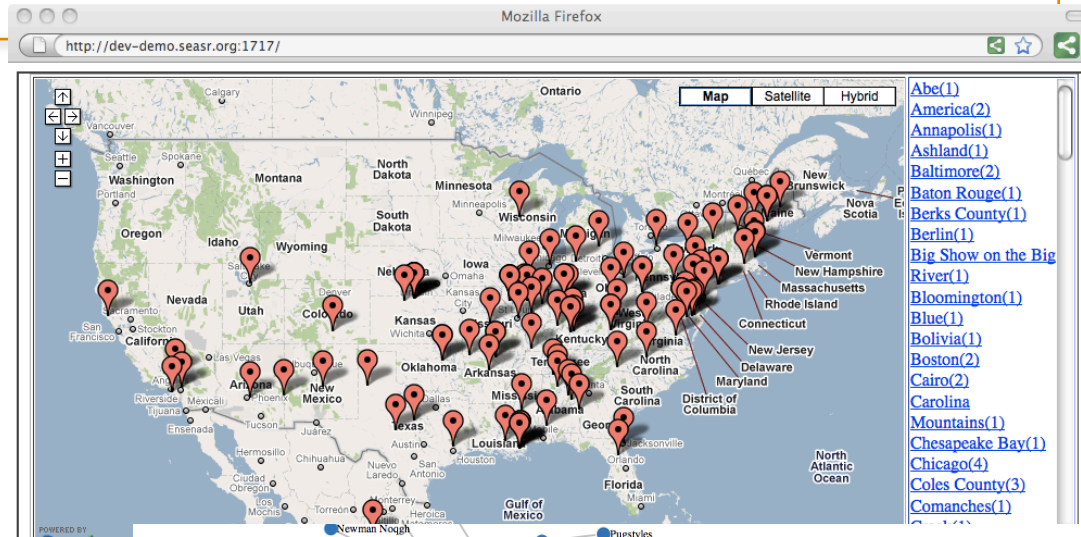
---

# Entity Extraction



# SEASR @ Work – Entity Mash-up

- Entity Extraction
  - Locations viewed on Google Map
  - Dates viewed on Simile Timeline
  - Entities in social network



# Text Preprocessing

---

- Syntactic analysis
  - Tokenization
  - Lemmitization
  - Ngrams
  - Part Of Speech (POS) tagging
  - Stop Word Removal
  - Shallow parsing
  - Custom literary tagging
- Semantic analysis
  - Information Extraction
    - Named Entity tagging
    - Unnamed Entity tagging
  - Co-reference resolution
  - Ontological association (WordNet, VerbNet)
  - Semantic Role analysis
  - Concept-Relation extraction



# Semantic Analysis

---

- Deep Parsing
  - more sophisticated syntactic, semantic and contextual processing must be performed to extract or construct the answer
- Information extraction is the identification of specific semantic elements within a text (e.g., entities, properties, relations)
- Extract the relevant information and ignore non-relevant information (important!)
- Link related information and output in a predetermined format



# Information Extraction Approaches

---

- Terminology (name) lists
  - This works very well if the list of names and name expressions is stable and available
- Tokenization and morphology
  - This works well for things like formulas or dates, which are readily recognized by their internal format (e.g., DD/MM/YY or chemical formulas)
- Use of characteristic patterns
  - This works fairly well for novel entities
  - Rules can be created by hand or learned via machine learning or statistical algorithms
  - Rules capture local patterns that characterize entities from instances of annotated training data



# Semantic Analytics

---

## – Named Entity (NE) Tagging

NE:Person                      NE:Time  
Mayor **Rex Luthor** announced **today** the establishment  
of a new research facility in **Alderwood**. It will be  
NE:Location  
known as **Boynton Laboratory**.  
NE:Organization



# Semantic Analysis

---

- Semantic Category (unnamed entity, UNE)  
Tagging

Mayor Rex Luthor announced today the establishment

UNE:Organization

of a **new research facility** in Alderwood. It will be

known as Boynton Laboratory.



# Semantic Analysis

---

- Co-reference Resolution for entities and unnamed entities

Mayor Rex Luthor announced today the establishment

UNE:Organization

of a **new research facility** in Alderwood. **It** will be

known as **Boynton Laboratory**.



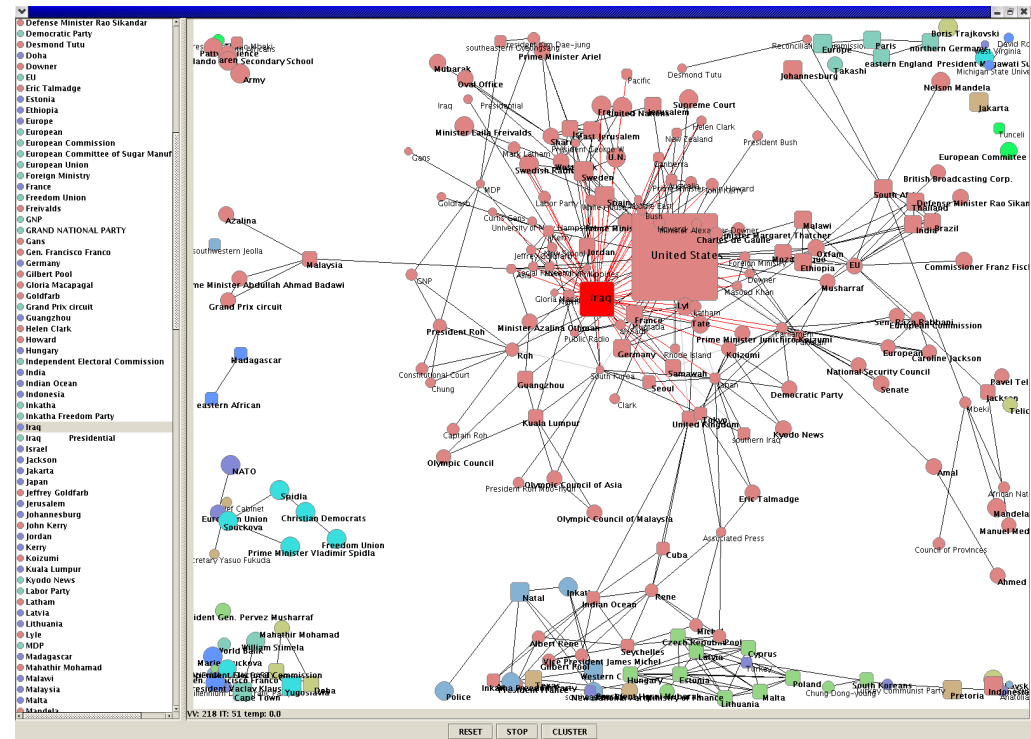


# Streaming Text: Knowledge Extraction

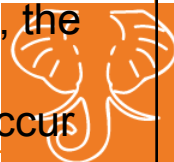
- Leveraging some earlier work on information extraction from text streams

## Information extraction

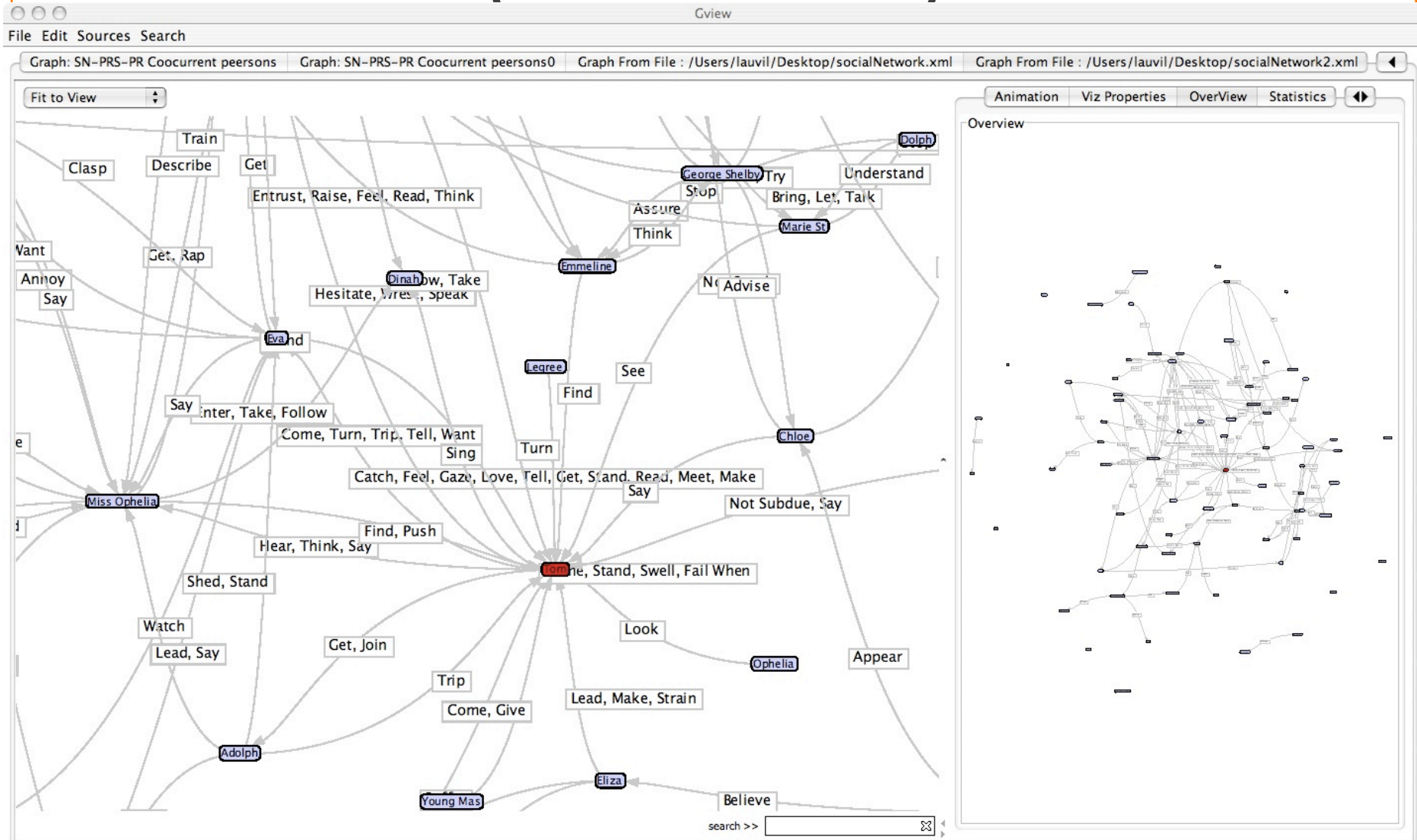
- process of using advanced automated machine learning approaches
- to identify entities in text documents
- extract this information along with the relationships these entities may have in the text documents



The visualization above demonstrates information extraction of names, places and organizations from real-time news feeds. As news articles arrive, the information is extracted and displayed. Relationships are defined when entities co-occur within a specific window of words.

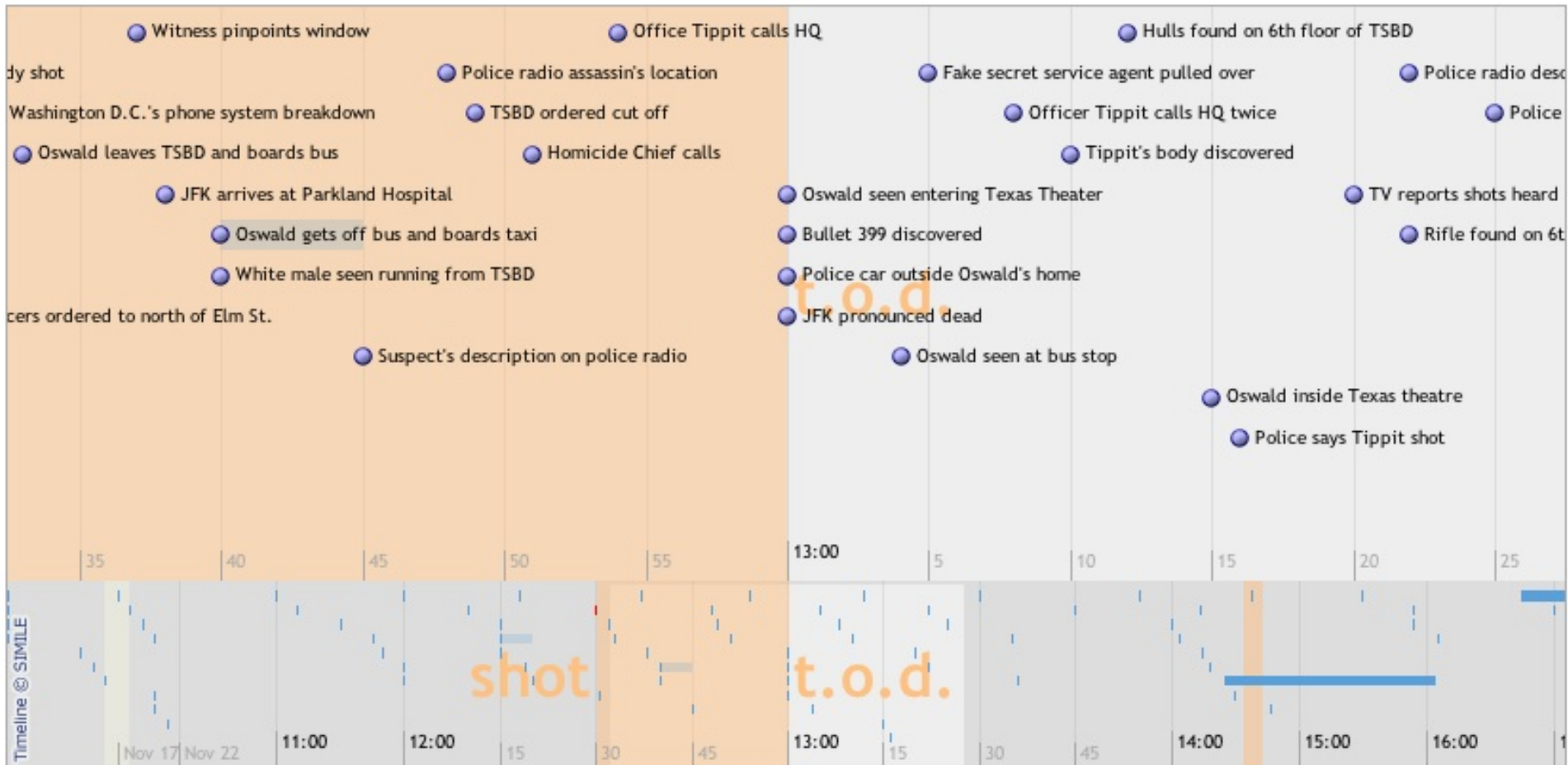


# Results: Social Network (Tom in Red)



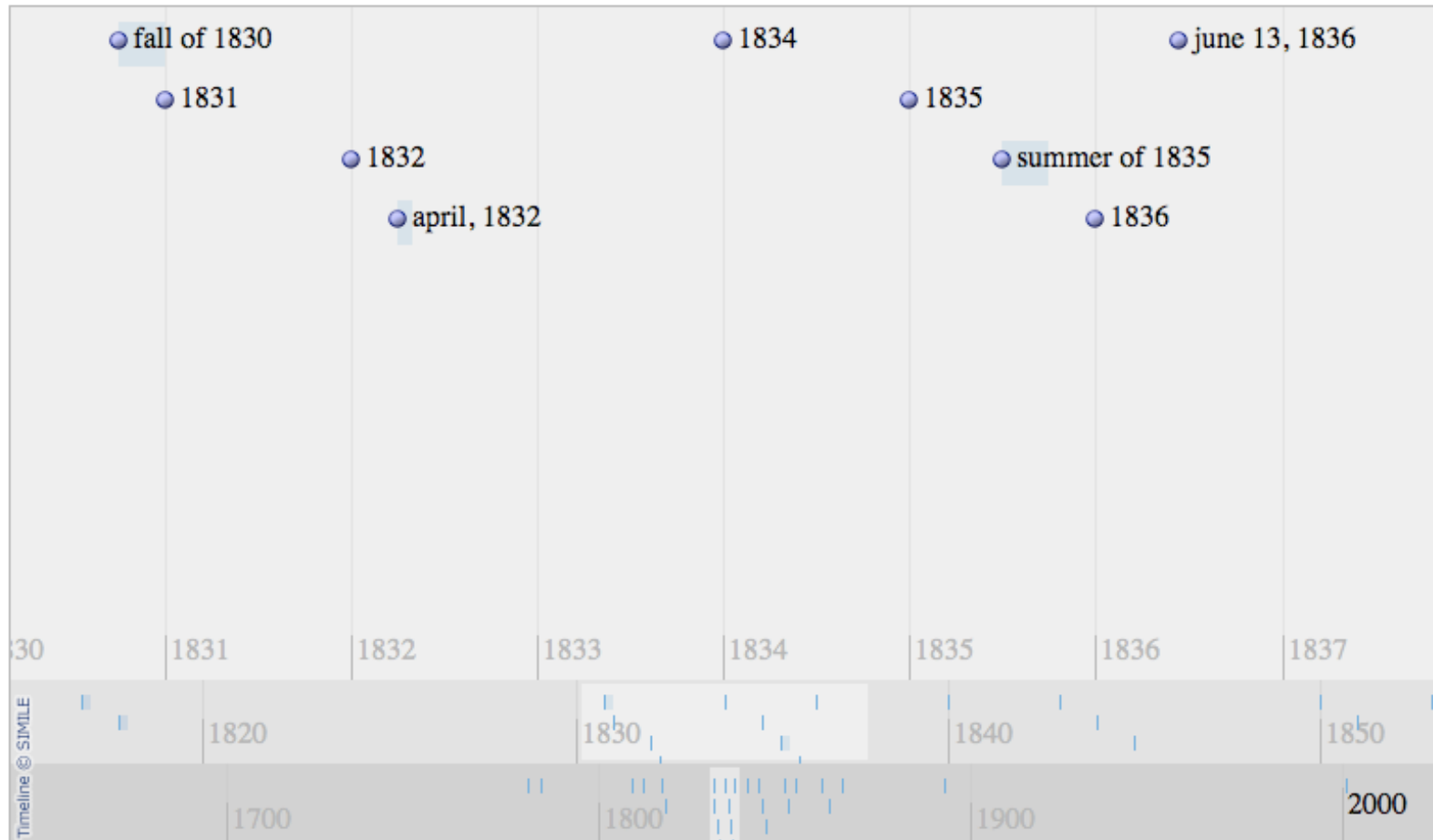
# Simile Timeline

- Constructed by Hand

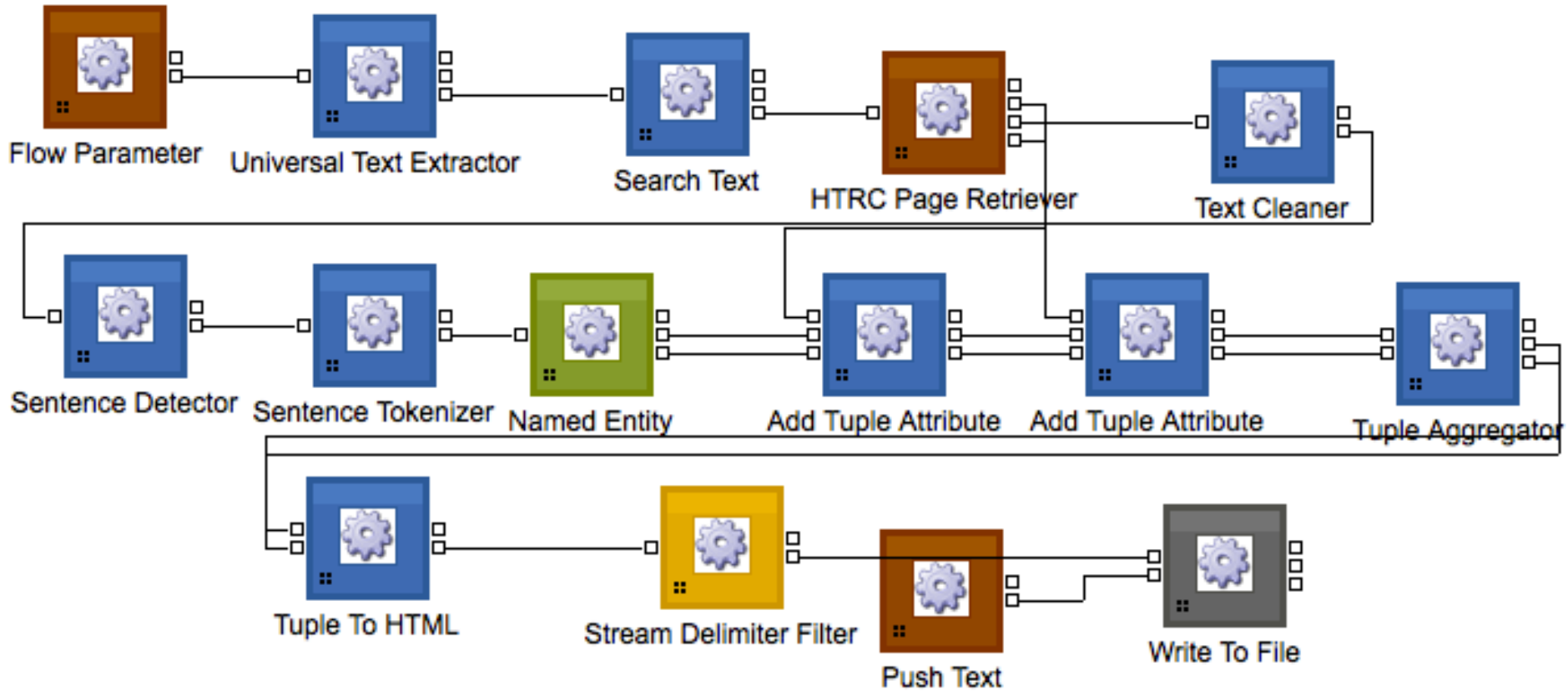


# Simile Timeline in SEASR

- Dates are automatically extracted with their sentences



# Flow for Dates to Simile



---

# Topic Modeling



# Text Analytics: Topic Modeling

- Given: Set of documents
- Find: To reveal the semantic content in large collection of documents
- Usage: Mallet Topic Modeling tools
- Output:
  - Shows the percentage of relevance for each document in each cluster
  - Shows the key words and their counts for each topic



# Topic Modeling: LDA Model

## Topics

## Documents

## Topic proportions and assignments

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a matter of numbers since, particularly as more and more genomes are completely mapped and sequenced, "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

- LDA assumes that there are K topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.



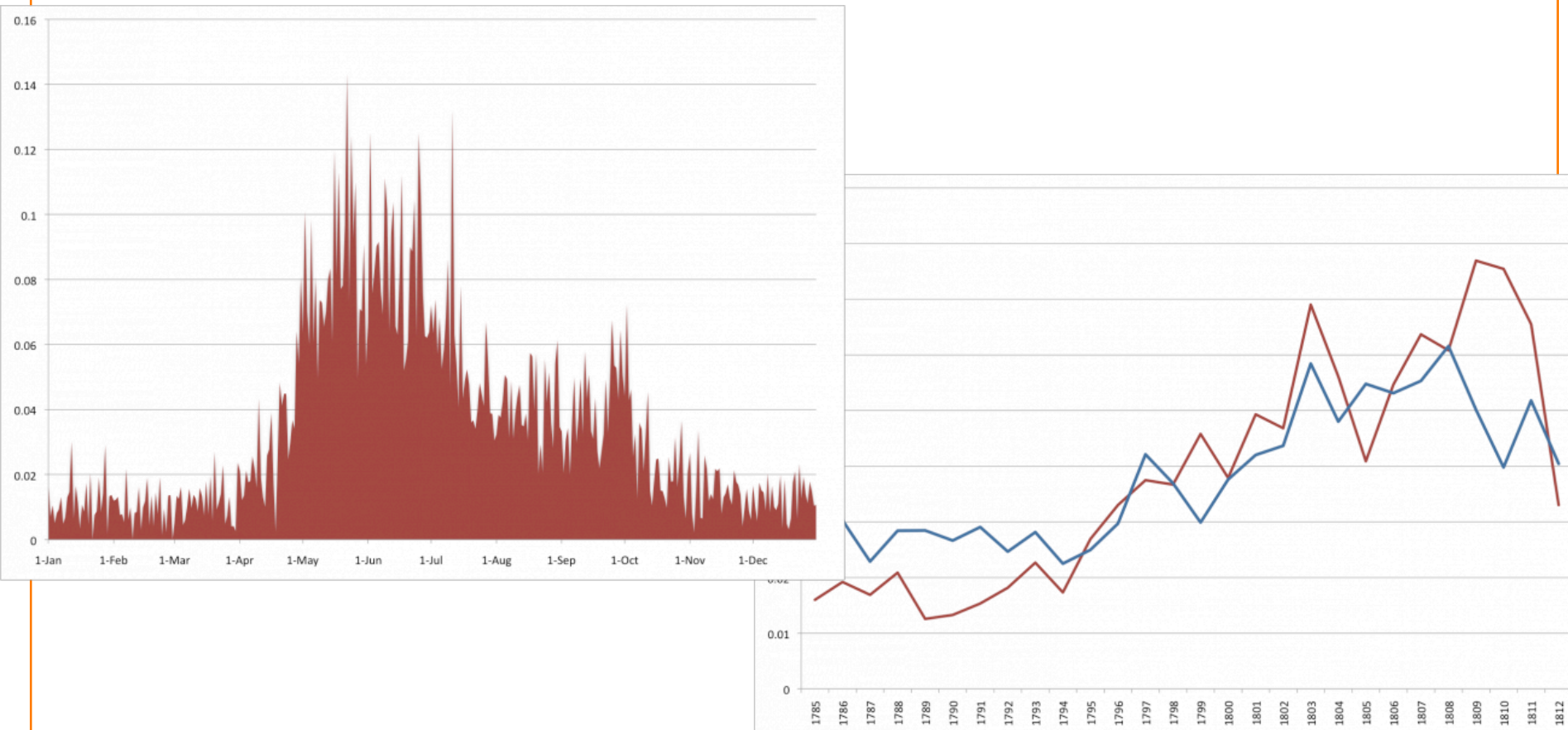
# Topic Modeling: Martha Ballard's Diary

Label	Words
MIDWIFERY	birth deld safe morn receivd calld left cleverly pm labour fine reward arivd infant expected recd shee born patient
CHURCH	meeting attended afternoon reverend worship foren mr famely performd vers attend public supper st service lecture discoarst administred supt
DEATH	day yesterday informd morn years death ye hear expired expird weak dead las past heard days drowned departed evinn
GARDENING	gardin sett worked clear beens corn warm planted matters cucumbers gatherd potatoes plants ou sowd door squash wed seeds
SHOPPING	lb made brot bot tea butter sugar carried oz chees pork candles wheat store pr beef spirit churnd flower
ILLNESS	unwell mr sick gave dr rainy easier care head neighbor feet relief made throat poorly takeing medisin ts stomach



# Topic Modeling: Martha Ballard's Diary

- Coordinating topics with additional data



# Topic Modeling: Pennsylvania Gazette

Label	Words
<i>RUNAWAY</i>	away reward servant old whoever named year feet jacket high paid hair pair secure coat run inches
<i>GOVT –U.S.</i>	state government constitution law united power citizen people public congress right legislature
<i>REAL ESTATE</i>	good house acre sold land meadow well mile premise plantation stone containing mill dwelling orchard
<i>GOVT –REVOLT</i>	country america war great liberty nation people american men let cause peace enemy present state she
<i>CLOTH</i>	silk cotton ditto white black linen cloth women blue worsted men fine thread plain coloured



# Topic Modeling: Historical Newspapers

Topics	Explanation
black* price* worth* white* goods* yard* silk* made* lot* week ladies wool* inch* ladles* sale* prices* pair* suits* fine*	Reflects discussion of the market and sales of goods, with some words that relate to cotton and others that reflect other goods being sold alongside cotton (such as wool).
state* people* states* bill* law* made united* party* men* country* government* county* public* presi- dent* money* committee* general* great question*	Political language associated with the political debates that dominated much of newspaper content during this era. The association of the topic “money” is particularly telling, as economic and fiscal policy were particularly important discussion during the era.
market* cotton* york* good* steady* closed* prices* corn* texas* wheat* fair* stock* choice* year* lower* receipts* ton* crop* higher*	All these topics reflect market-driven language related to the buying and selling cotton and, to a much smaller extent, other crops such as corn.



# Topic Modeling: Mining the Dispatch

---

- **Topic words**

- negro, years, reward, boy, man, named, jail, delivery, give, left, black, paid, pay, ran, color, richmond, subscriber, high, apprehension, age, ranaway, free, feet, delivered

- **Advertisement**

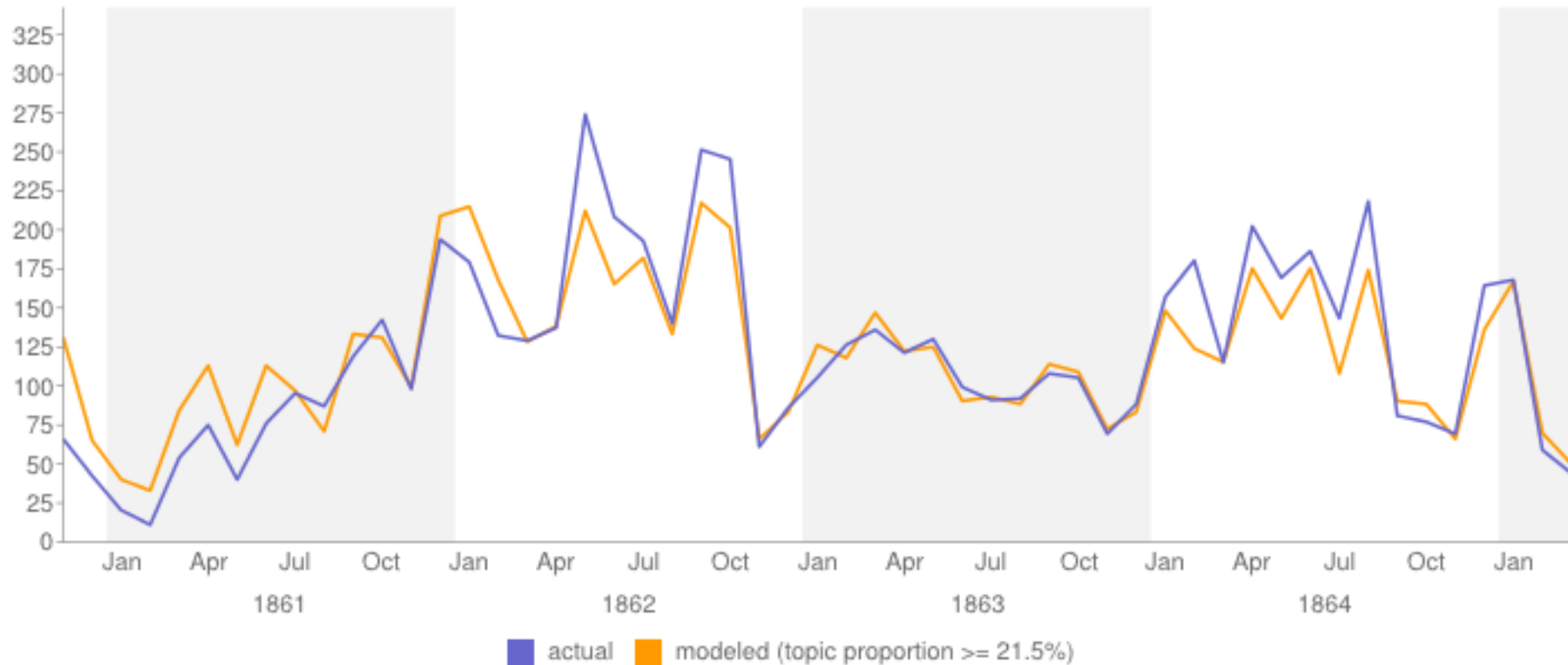
- Ranaway.—\$10 reward.**

- Ranaway from the subscriber, on the 3d inst., my slave woman Parthena. Had on a dark brown and white calico dress. She is of a ginger-bread color; medium size; the right fore-finger shortened and crooked, from a whitlow. I think she is harbored somewhere in or near Duvall's addition. For her delivery to me I will pay \$10.

- de 6—ts G. W. H. Tyler.



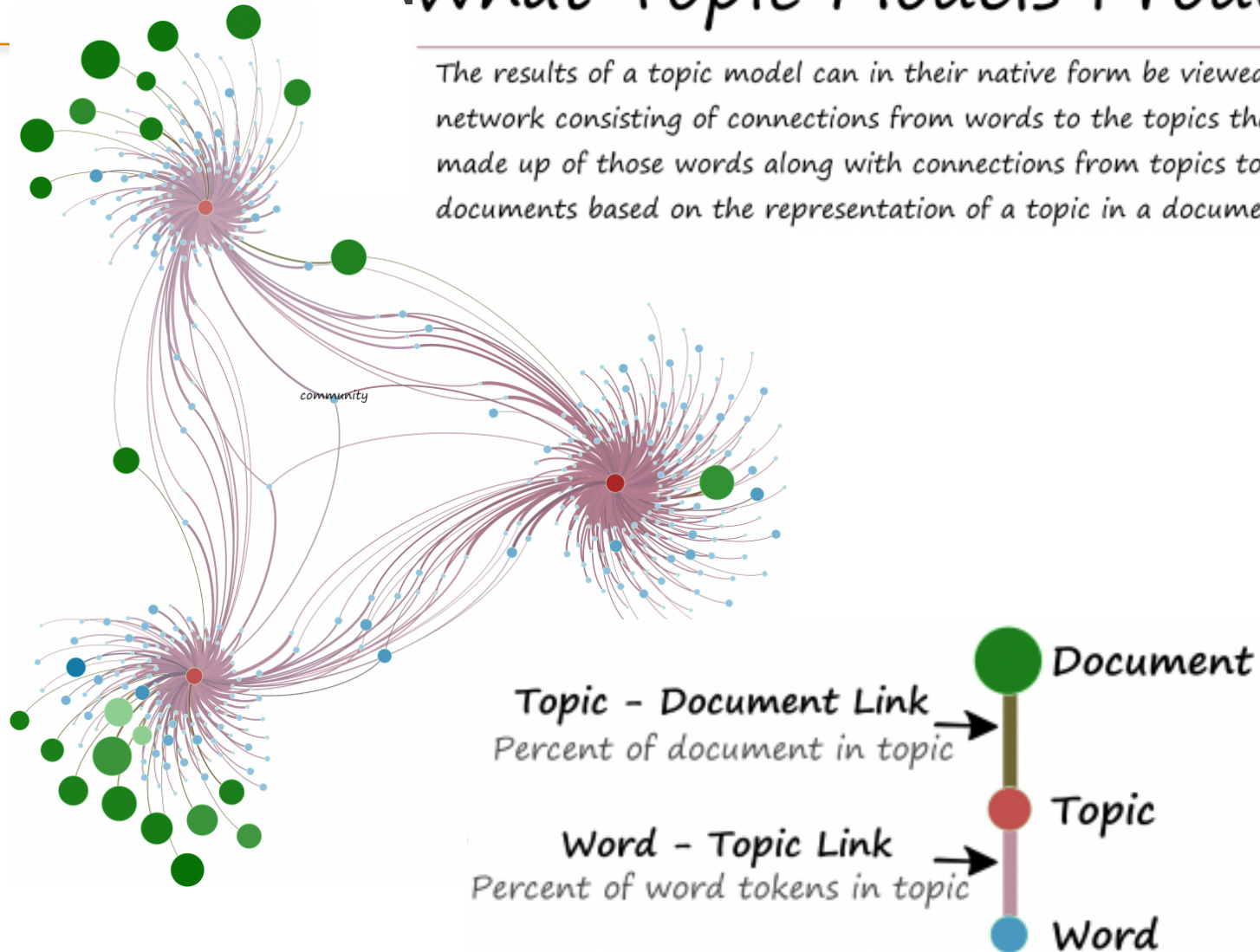
# Topic Modeling: Mining the Dispatch



# Topic Modeling: Link-Node

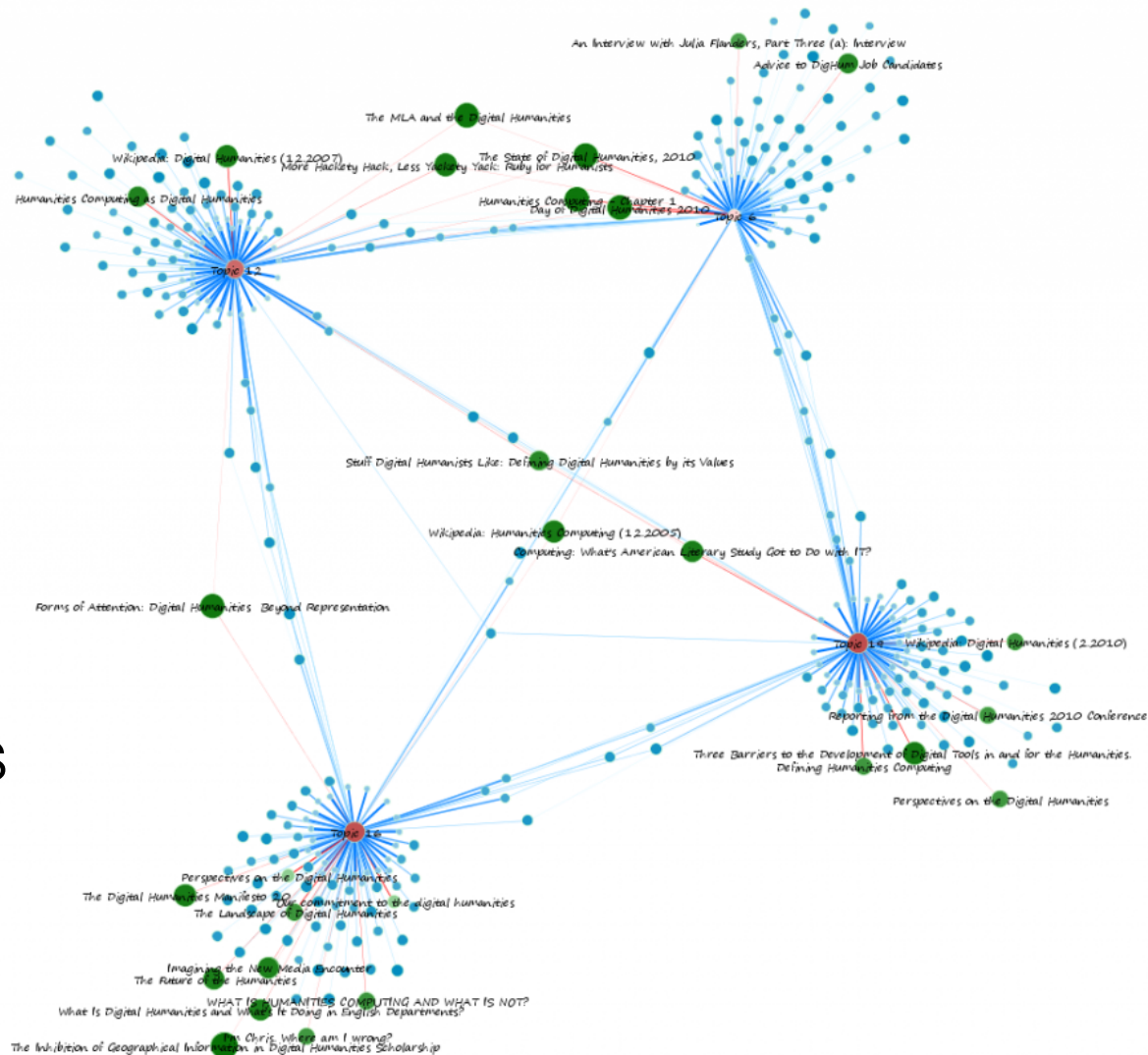
## What Topic Models Produce

The results of a topic model can in their native form be viewed as a network consisting of connections from words to the topics that are made up of those words along with connections from topics to documents based on the representation of a topic in a document.



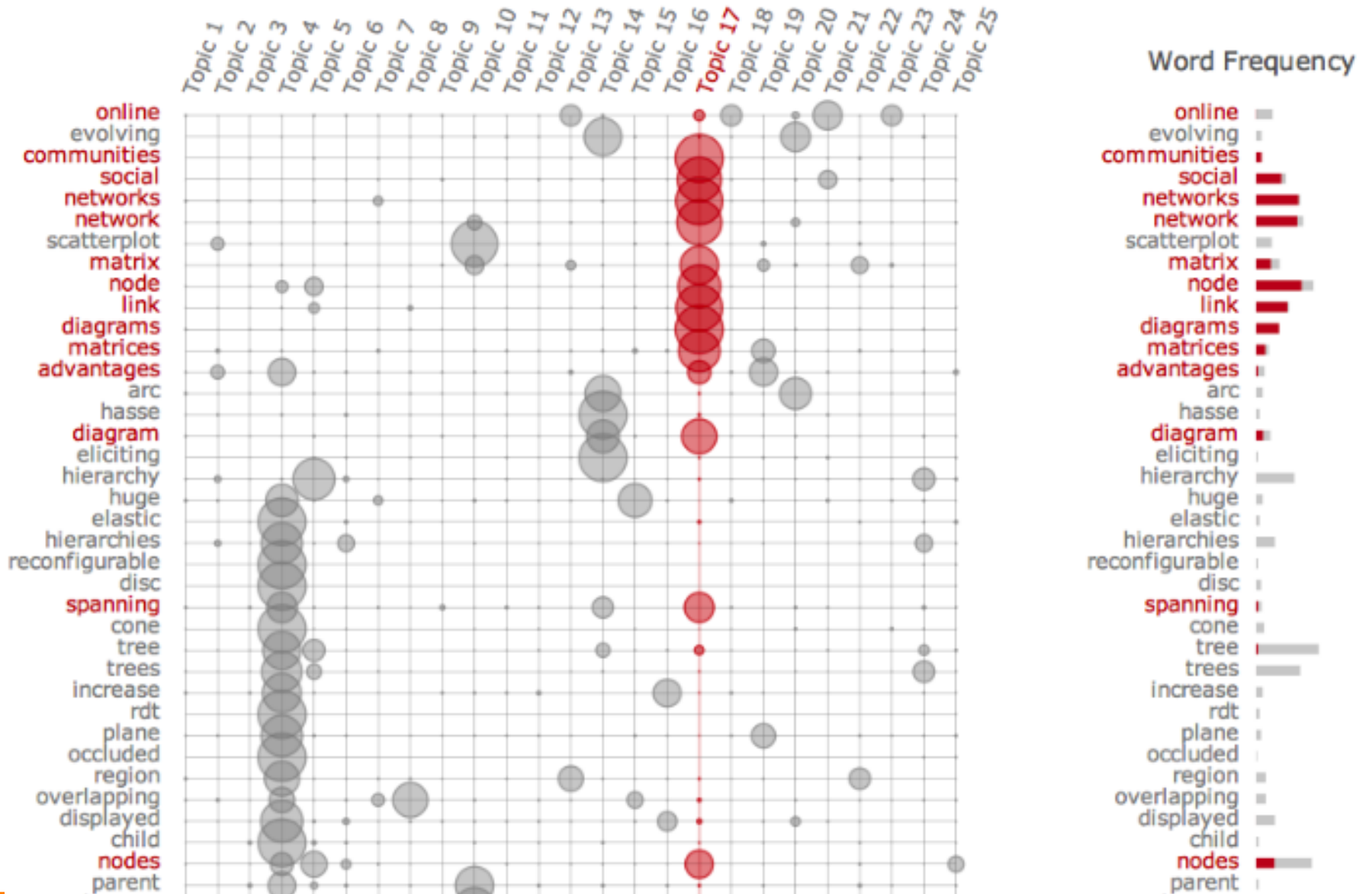
# Topic Modeling: Link-Node Visualization

Extract the tokens, word counts, and their connections from the Mallet topic model files into a graph file that generates edges and nodes, allowing us to view the ten topics as a network model in Gephi





# Topic Modeling: Matrix





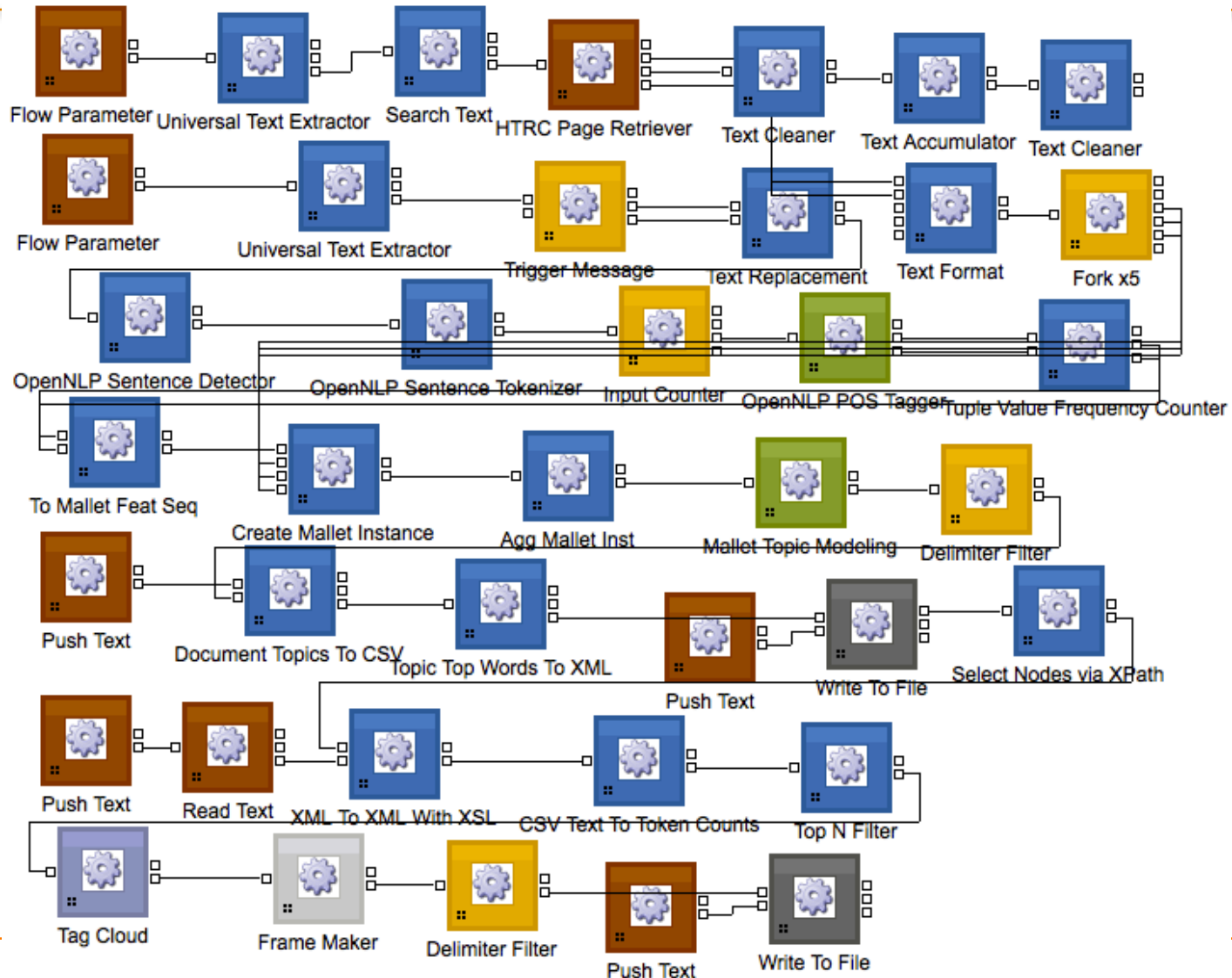
# Topic Modeling Process

---

- Load the documents
- Segment the documents
- Extract nouns (POS analysis)
- Create the Mallet data structures for each segment
- Mallet for topic modeling
- Save results
- Parse keyword results
- Create tagclouds of keywords



# Topic Modeling Flow



# HTRC Topics

- Search for “dickens” in collections
- 1148 documents
- 100 topics, showing 2 below



# Topic Model Explorer

Token : london

Submit

Documents List
















File	Title	Year	First Name	Last Name	Nation	Ge	Count
2393	The mysteries of the court of Lor	1849	George	Reynolds	British	M	1008
3080	Black Bess; or, the knight of the	1866	Edward	Viles	British	M	716
3081	Blueskin: a romance of the last c	1866	Edward	Viles	British	M	516
2392	Mary Price ; or the memoirs of a	1853	George	Reynolds	British	M	376
3132	Ten thousand a-year	1840	Samuel	Warren	British	M	362

Topics List

Topic Id	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5	Actions
2	london	birmingham	england	messrs .	portman square	 
3	india	germany	united states	london	lowell	 
4	england	sheffield	london	wilmington	france	 
5	london	fairfield	england	lima	miss malcolm	 
12	london	england	hampshire	bar	grand	 



Topic Location Correlation for topic 2

File	Title	Year	LastNa	Nation	Ge	Segment	NumTypes	NumTokens	Actions
1410	Almack's : a novel	1826	Hudsor	British	F	329	5	5	
1101	The two aristocracies : a no	1857	Gore	British	F	107	4	6	
1158	The bosom friend : a novel	1845	Grey	British	F	78	4	5	
1158	The bosom friend : a novel	1845	Grey	British	F	109	4	7	
1410	File 98.xml, segment 101						4	6	
1410	imprudent as a marriage between mr . wickham and our poor lydia would be , we are now anxious to be assured it has taken place , for there is but too much reason to fear they are not gone to scotland . colonel forster came yesterday , having left brighton the day before , not many hours after the express . though lydia's short letter to mrs . f . gave them to understand that they were going to gretna green , something was dropped by denny expressing his belief that w . never intended to go there , or to marry lydia at all , which was repeated to colonel f . who instantly taking the alarm , set off from b . intending to trace their route . he did trace them easily to clapham , but no farther ; for on entering that place they removed into a hackney-coach and dismissed the chaise that brought them from epsom . all that is known after this is , that they were seen to continue the <b>london</b> road . i know not what to think . after making every possible inquiry on that side <b>london</b> , colonel f . came on into hertfordshire , anxiously renewing them at all the turnpikes , and at the inns in barnet and hatfield , but without any success , no such people had been seen to pass through . with the kindest concern he came on to longbourn , and broke his apprehensions to us in a manner most creditable to his heart . i am sincerely grieved for him and mrs . f . but no one can throw any blame on them . our distress , my dear lizzy , is very great . my father and mother believe the worst , but i cannot think so ill of him . many circumstances might make it more eligible for them to be married privately in town than to pursue their first plan ; and even if he could form such a design against a young woman of lydia's connexions , which is not likely , can i suppose her so lost to every thing?— impossible . i grieve to find , however , that colonel f . is not disposed to depend upon their marriage ; he shook his head when i expressed my hopes , and said he feared w . was not a man to be trusted . my poor mother is really ill and keeps her room . could she exert herself it would be better , but this is not to be expected ; and as to my father , i never in my life saw him so affected . poor kitty has anger for having concealed their attachment ; but as it was a matter of confidence one cannot wonder . i am truly glad . dearest lizzy . that you					4	6		
1410							4	6	
1410							4	5	
1410							4	6	
1410							4	5	
1410							4	4	
1515							4	6	
2030							4	4	
2266							4	4	
2268							4	6	

# Additional Topic Modeling Variations

---

- Topics over time
- Connections between topics
- Hierarchy of topics



# Topic Modeling References

---

- <http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/>
- <http://dsl.richmond.edu/dispatch/pages/intro>
- <http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/>
- [http://www.ics.uci.edu/~newman/pubs/JASIST\\_Newman.pdf](http://www.ics.uci.edu/~newman/pubs/JASIST_Newman.pdf)
- <https://dhs.stanford.edu/visualization/topic-networks/>
- Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 96–104, Portland, OR, USA, 24 June 2011. © 2011 Association for Computational Linguistics
- Matthew Jockers, *Macroanalysis: Digital Methods and Literary History*, UIUC Press, 2013
- [Termite: Visualization Techniques for Assessing Textual Topic Models](#), [Jason Chuang](#), Christopher D. Manning, [Jeffrey Heer](#), *Advanced Visual Interfaces*, 2012





---

# Spell Checking



# Correlation-Ngram Viewer

## Pearson Correlation Algorithm

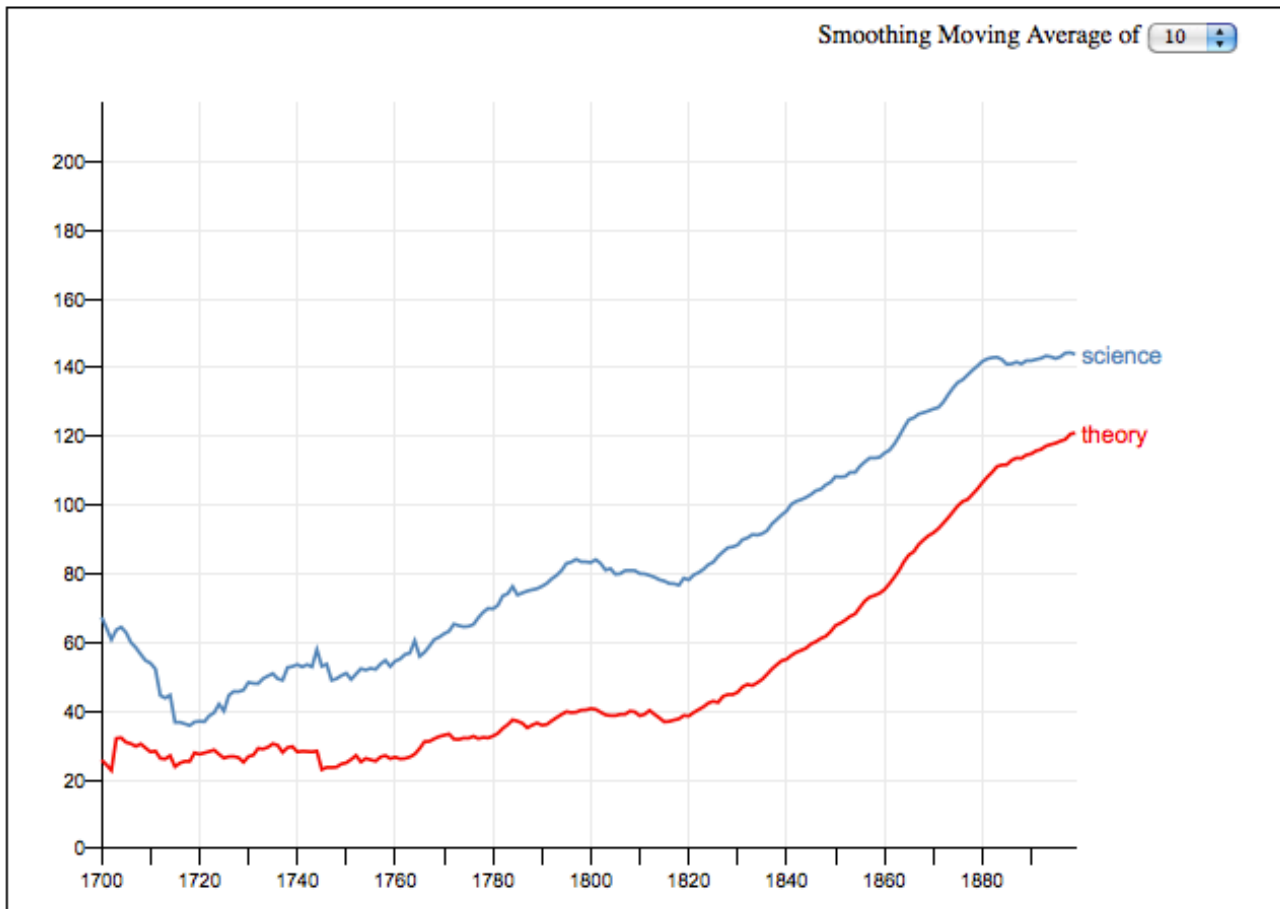
Ngram Viewer

science,theory|

Search

Note: When entering an ngram, use a comma to separate ngrams that you want to compare.

Smoothing Moving Average of 10



Ngrams Included

ngram	variations	sum_match
science	science	3535055
science	Science	790086
science	SCIENCE	130651
science	fcience	48279
science	scienc	1637
science	Science	1203
science	fcinc	461
science	fcinc	456
science	Scienc	239
science	science	182
science	8cience	162
science	SCIENCE	138
science	Scienc	80
science	8SCIENCE	76
science	scienc	60
science	SCIENCE	60
science	Scienc	28
science	Fcienc	19
science	Scienc	19
science	Scienc	17
science	Scienc	14
science	sCIENCE	12
science	Science	7
science	scienc	6
science	scienc	4
science	SCIENCE	4

# OCR Correction

- HTRC Example of one of the worst pages of text based on number of corrections per word rate = 0.1994

Testing/src/Page\_Corrected.txt

A D 1817 57<sup>th</sup> GEO III C xxix 653 the faicl Commiffioners or #trustees# or other #persons# as #aforesaid# to #issue# their Preceptor Precepts to the Sheriff or Sheriffs or Bailiff or other proper Officer of the City Borough or County wherein #such# parochial or other #district# hall be #situate# to deliver #possession# of the #said# #premises# to #such# #person# or #persons# as #shall# in #such# Precept or Precepts be nominated to receive the fame and the #said# Sheriff or Sheriffs or Bailiff and every other proper Officer is hereby #authorised# and required to deliver #such# #possession# accordingly of the #said# #premises# and to levy #such# Coils as mail accrue from the #issuing# #costs# and Execution of #such# Precept or Precepts on the #person# or #persons# fo #refusing# to give #possession# as #aforesaid# by #distress# and Sale of his her or their Goods XCI And be it further enacted That all and every #person# and Mortgagees not m ooot on #persons# who mail have any Mortgage or Mortgages on #such# #houses# Buildings Lands Tenements and Hereditaments not being in Pof en er \*f rm T rrp i n • r r

i r n r in cipa and #interest# teilion thereof by virtue or iuch Mortgage or Mortgages mall on and Six Months the Tender of the Principal Money and #interest# due thereon #interest# to contogether with the Amount of Six Calendar Months #interest# on the vey to Commiffaid Principal by the #said# #commissions# or #trustees# or other #persons# fioners» c j having the Control of the Pavements in the Streets or public Places in #such# parochial or other #district# within the Jurifdi ion of this A 61 wherein the #said# #houses# Buildings Lands Tenements and Hereditaments #shall# lie or be as #aforesaid# or by #such# #person# or #persons# as they mail appoint immediately convey affign and transfer #such# Mortgage or Mortgages to the #said# #commissions# or #trustees# or other #persons# as #aforesaid# or to #such# #person# or #persons# as they #shall# appoint or in cafe #such# Mortgagee or Mortgagees #shall# have or on Six Notice in Writing from the #said# #commissions# or #trustees# or other Months Notice #persons# as #aforesaid# or from #such# #person# or #persons# as they #shall# appoint that they will pay off and #discharge# the Principal Money and #interest# which #shall# be due on the #said# Mortgage or Mortgages at the End or Expiration of Six Calendar Months to be computed from the Day of giving #such# Notice that then at the End of the #said# Six Calendar Months on Payment of the Principal and #interest# fo due #such# Mortgagee or Mortgagees #shall# convey #assign# and transfer his her or their #interest# in the #premises# to the #said# #commissions# or #trustees# or other #persons# as #aforesaid# or to #such# #person# or #persons# as #shall# be appointed in #trust# for them and in cafe the on #refusal# TnMortgagee or Mortgagees #shall# #refuse# to convey and #assign# as afore #terest# on Mortfaid on #such# Tender or Payment that then all #interest# on every #such# #gauge# to ceale Mortgage #shall# from thenceforth #cease# and determine XCII Provided always and be it further #enacted# That in cafe Mortgagees not the Sum due upon any #such# Mortgage or Mortgages with all #interest# to be paid more due thereon #shall# amount to more than the real Value of the than the real #premises# to be #ascertained# as directed by this Aft then the #said# #commissions# or #trustees# or other #persons# as #aforesaid# #shall# not be liable to pay to the Mortgagee or Mortgagees more than #such# real Value of #such# #premises# fo #ascertained# as areofaid XCIII And be it further enacted That the Conveyance of any Bargains and #such# #estate# or #interest# of any Feme Covert to the #said# #commissions# Sales to have the or #trustees# or other #persons# as #aforesaid# for the time being or any Force of Fines Five or more of them or any #person# or #persons# in #trust# for them and Recoveries by Indenture or Indentures of Bargain and Sale #sealed# and delivered by #such# Feme Covert in the #presence# of and #attested# by Two credible

Testing/src/Page.txt

A D 1817 57<sup>th</sup> GEO III C xxix 653 the faicl Commiffioners or Trustees or other Perfons as aforefaid to iffue their Preceptor Precepts to the Sheriff or Sheriffs or Bailiff or other proper Officer of the City Borough or County wherein fuch parochial or other Diftrict hall be fituate to deliver Poffeffion of the faid Premifes to fuch Perfon or Perfons as fhall in fuch Precept or Precepts be nominated to receive the fame and the faid Sheriff or Sheriffs or Bailiff and every other proper Officer is hereby authorized and required to deliver fuch Poffeffion accordingly of the faid Premifes and to levy fuch Coils as mail accrue from the iffuing Cofts and Execution of fuch Precept or Precepts on the Perfon or Perfons fo refusing to give Poffeffion as aforefaid by Distrefs and Sale of his her or their Goods XCI And be it further enacted That all and every Perfon and Mortgagees not m ooot on Perfons who mail have any Mortgage or Mortgages on fuch Houfes Buildings Lands Tenements and Hereditaments not being in Pof en er \*f rm T rrp i n • r r

i r n r in cipa and Intereft teilion thereof by virtue or iuch Mortgage or Mortgages mall on and Six Months the Tender of the Principal Money and Intereft due thereon Intereft to contogether with the Amount of Six Calendar Months Intereft on the vey to Commiffaid Principal by the faid Commiffioners or Trustees or other Perfons fioners» c j having the Control of the Pavements in the Streets or public Places in fuch parochial or other Diftrict within the Jurifdi ion of this A 61 wherein the faid Houfes Buildings Lands Tenements and Hereditaments fhall lie or be as aforefaid or by fuch Perfon or Perfons as they mail appoint immediately convey affign and transfer fuch Mortgage or Mortgages to the faid Commiffioners or Trustees or other Perfons as aforefaid or to fuch Perfon or Perfons as they fhall appoint or in cafe fuch Mortgagee or Mortgagees fhall have or on Six Notice in Writing from the faid Commiffioners or Trustees or other Months Notice Perfons as aforefaid or from fuch Perfon or Perfons as they fhall appoint that they will pay off and difcharge the Principal Money and Intereft which fhall be due on the faid Mortgage or Mortgages at the End or Expiration of Six Calendar Months to be computed from the Day of giving fuch Notice that then at the End of the faid Six Calendar Months on Payment of the Principal and Intereft fo due fuch Mortgagee or Mortgagees fhall convey affign and transfer his her or their Intereft in the Premifes to the faid Commiffioners or Trustees or other Perfons as aforefaid or to fuch Perfon or Perfons as fhall be appointed in Truft for them and in cafe the on Refusal TnMortgagee or Mortgagees fhall refuse to convey and affign as afore tereft on Mortfaid on fuch Tender or Payment that then all Intereft on every fuch gage to ceale Mortgage fhall from thenceforth ceafe and determine XCII Provided always and be it further #enacted# That in cafe Mortgagees not the Sum due upon any fuch Mortgage or Mortgages with all Intereft to be paid more due thereon fhall amount to more than the real Value of the than the real Premifes to be #ascertained# as directed by this Aft then the faid Commiffioners or Trustees or other Perfons as aforefaid fhall not be liable to pay to the Mortgagee or Mortgagees more than fuch real Value of fuch Premifes fo #ascertained# as areofaid XCIII And be it further enacted That the Conveyance of any Bargains and fuch Eftate or Intereft of any Feme Covert to the faid Commiffioners Sales to have the or Trustees or other Perfons as aforefaid for the time being or any Force of Fines Five or more of them or any Perfon or Perfons in Truft for them and Recoveries by Indenture or Indentures of Bargain and Sale fealed and delivered by fuch Feme Covert in the Prefence of and attested by Two credible

# Worst Page

A D 1817 57°GEO III C xxix 653 the faicl Commiffioners or Truftees or other Perfons as aforefaid to iffue their Preceptor Precepts to the Sheriff or Sheriffs or Bailiff or other proper Officer of the City Borough or County wherein fuch parochial or other Diftrift hall be fituate to deliver Poffeffion of the faid Premifes to fuch Perfon or Perfons as fhall in fuch Precept or Precepts be nominated to receive the fame and the faid Sheriff or Sheriffs or Bailiff and every other proper Officer is hereby authorized and required to deliver fuch Poffeffion accordingly of the faid Premifes and to levy fuch Coils as mail accrue from the iffuing Cofts and Execution of fuch Precept or Precepts on the Perfon or Perfons fo refufing to give Poffeffion as aforefaid by Diftreffs and Sale of his her or their Goods XCI And be it further enacted That all and every Perfon and Mortgagees not m oeci on Perfons who mail have any Mortgage or Mortgages on fuch Houfes Buildings Lands Tenements and Hereditaments not being in Pof en er °f rm T rrp i n • r r

i r n r in cipa and Intereft ieiiiion thereof by virtue or iuch Mortgage or Mortgages mall on and Six Months the Tender of the Principal Money and Intereft due thereon Intereft to contogether with the Amount of Six Calendar Months Intereft on the vey to Commiffaid Principal by the faid Commiffioners or Truftees or other Perfons fioners» c j having the Control of the Pavements in the Streets or public Places in fuch parochial or other Diftrift within the Jurifdi ion of this A 61 wherein the faid Houfes Buildings Lands Tenements and Hereditaments fhall lie or be as aforefaid or by fuch Perfon or Perfons as they mail appoint immediately convey affign and transfer fuch Mortgage or Mortgages to the faid Commiffioners or Truftees or other Perfons as aforefaid or to fuch Perfon or Perfons as they fhall appoint or in cafe fuch Mortgagee or Mortgagees fhall have or on Six Notice in Writing from the faid Commiffioners or Truftees or other Months Notice Perfons as aforefaid or from fuch Perfon or Perfons as they fhall appoint that they will pay off and difcharge the Principal Money and Intereft which fhall be due on the faid Mortgage or Mortgages at the End or Expiration of Six Calendar Months to be computed from the Day of giving fuch Notice that then at the End of the faid Six Calendar Months on Payment of the Principal and Intereft fo due fuch Mortgagee or Mortgagees fhall convey affign and transfer his her or their Intereft in the Premifes to the faid Commiffioners or Truftees or other Perfons as aforefaid or to fuch Perfon or Perfons as fhall be appointed in Truft for them and in cafe the on Refufal TnMortgagee or Mortgagees fhall refufe to convey and affign as afore tereft on Mortfaid on fuch Tender or Payment that then all Intereft on every fuch gage to ceale Mortgage fhall from thenceforth ceafe and determine XCII Provided always and be it further enacted That in cafe Mortgagees not the Sum due upon any fuch Mortgage or Mortgages with all Intereft to be paid more due thereon fhall amount to more than the real Value of the than the real Premifes to be afcertained as directed by this Aft then the faid Commiffioners or Truftees or other Perfons as aforefaid fhall not be liable to pay to the Mortgagee or Mortgagees more than fuch real Value of fuch Premifes fo afcertained as areofaid XCIII And be it further enacted That the Conveyance of any Bargains and fuch Eftate or Intereft of any Feme Covert to the faid Commiffioners Sales to have the or Truftees or other Perfons as aforefaid for the time being or any Force of Fines Five or more of them or any Perfon or Perfons in Truft for them and Recoveries« by Indenture or Indentures of Bargain and Sale fealed and delivered by fuch Feme Covert in the Prefence of and attefted by Two credible



# Corrected Page

A D 1817 57°GEO III C xxix 653 the faicl Commiffioners or #trustees# or other #persons# as #aforesaid# to #issue# their Preceptor Precepts to the Sheriff or Sheriffs or Bailiff or other proper Officer of the City Borough or County wherein #such# parochial or other #district# hall be #situate# to deliver #possession# of the #said# #premises# to #such# #person# or #persons# as #shall# in #such# Precept or Precepts be nominated to receive the fame and the #said# Sheriff or Sheriffs or Bailiff and every other proper Officer is hereby #authorised# and required to deliver #such# #possession# accordingly of the #said# #premises# and to levy #such# Coils as mail accrue from the #issuing# #costs# and Execution of #such# Precept or Precepts on the #person# or #persons# fo #refusing# to give #possession# as #aforesaid# by #distress# and Sale of his her or their Goods XCI And be it further enacted That all and every #person# and Mortgagees not m oeci on #persons# who mail have any Mortgage or Mortgages on #such# #houses# Buildings Lands Tenements and Hereditaments not being in Pof en er °f rm T rrp i n • r r

i r n r in cipa and #interest# ieiion thereof by virtue or iuch Mortgage or Mortgages mall on and Six Months the Tender of the Principal Money and #interest# due thereon #interest# to contogether with the Amount of Six Calendar Months #interest# on the vey to Commiffaid Principal by the #said# #commissioners# or #trustees# or other #persons# fioners» c j having the Control of the Pavements in the Streets or public Places in #such# parochial or other #district# within the Jurifd i on of this A 61 wherein the #said# #houses# Buildings Lands Tenements and Hereditaments #shall# lie or be as #aforesaid# or by #such# #person# or #persons# as they mail appoint immediately convey afign and transfer #such# Mortgage or Mortgages to the #said# #commissioners# or #trustees# or other #persons# as #aforesaid# or to #such# #person# or #persons# as they #shall# appoint or in cafe #such# Mortgage or Mortgagees #shall# have or on Six Notice in Writing from the #said# #commissioners# or #trustees# or other Months Notice #persons# as #aforesaid# or from #such# #person# or #persons# as they #shall# appoint that they will pay off and #discharge# the Principal Money and #interest# which #shall# be due on the #said# Mortgage or Mortgages at the End or Expiration of Six Calendar Months to be computed from the Day of giving #such# Notice that then at the End of the #said# Six Calendar Months on Payment of the Principal and #interest# fo due #such# Mortgagee or Mortgagees #shall# convey #assign# and transfer his her or their #interest# in the #premises# to the #said# #commissioners# or #trustees# or other #persons# as #aforesaid# or to #such# #person# or #persons# as #shall# be appointed in #trust# for them and in cafe the on #refusal# TnMortgagee or Mortgagees #shall# #refuse# to convey and #assign# as afore #terest# on Mortfaid on #such# Tender or Payment that then all #interest# on every #such# #gauge# to ceale Mortgage #shall# from thenceforth #cease# and determine XCIL Provided always and be it further #enacted# That in cafe Mortgagees not the Sum due upon any #such# Mortgage or Mortgages with all #interest# to be paid more due thereon #shall# amount to more than the real Value of the than the real #premises# to be #ascertained# as directed by this Aft then the #said# #commissioners# or #trustees# or other #persons# as #aforesaid# #shall# not be liable to pay to the Mortgagee or Mortgagees more than #such# real Value of #such# #premises# fo #ascertained# as areofaid XCIII And be it further enacted That the Conveyance of any Bargains and #such# #estate# or #interest# of any Feme Covert to the #said# #commissioners# Sales to have the or #trustees# or other #persons# as #aforesaid# for the time being or any Force of Fines Five or more of them or any #person# or #persons# in #trust# for them and Recoveries« by Indenture or Indentures of Bargain and Sale #sealed# and delivered by #such# Feme Covert in the #presence# of and #attested# by Two credible



# Some Stats

	Google Ngram	HTRC 250K Books
Total number of ngrams:	359,511,583,097	20,173,974,251
Total number of ngrams (ignoring punctuation chars):	306,780,490,555	
Total number of ngrams (ignoring numbers only & repeating characters, other noise that I could easily identify):	293,760,570,946	19,282,108,416
Total number of corrections that we have made:	1,660,948,155	131,571,046
<b>Percent of Cleaning</b>	<b>0.57%</b>	<b>0.68%</b>
Unique ngrams before cleaning	7,380,256	
Unique ngrams after cleaning	4,977,548	
Number of generated rules:	154,227	
Number of valid rules:	99,455	99,455
Number of rules that are shorter than 5 chars and ignored	7,076	



# Spellchecking Analysis

---

- Not just OCR detection but OCR correction
- Can also be used for cleaning other messy data



# Spell Check Component

---

- Wrapped existing spellchecker from `com.swabunga.spell`
- Input
  - Dictionary to define the correct words
  - Transformations is a set of rules that should be tried on misspelled words before taking the spell checker's suggestions
  - Token counts is a set of counts that can be used to choose word when spell checker suggests multiple ones
- Output
  - Replacement Rules are the transformation rules for misspelled words
  - Replacements are suggestions for misspelled words
  - Corrected Text is the original text with corrections applied
  - Uncorrected Misspellings is the list of words for which a correction/ replacement could not be found





# Spell Check Properties

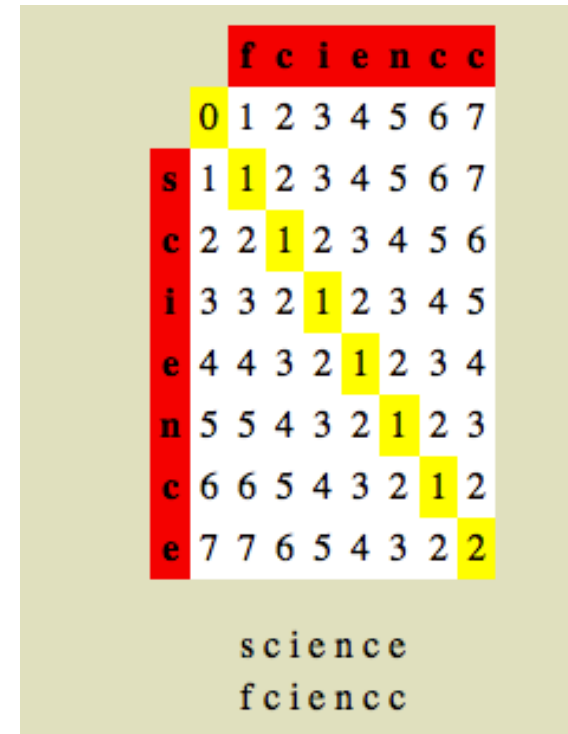
---

- `do_correction = false`
- `enable_transforms_only = true`
- `enable_levenshtein = true`
- `ignore_digitwords = true`
- `ignore_uppercase = false`
- `ignore_internetaddresses = true`
- `ignore_mixedcase = false`
- `levenshtein_distance = .25`
- `min_rule_support = 1`
- `output_misspellings_with_counts = false`
- `transform_threshold = 30`
- `_debug_level = info`
- `_ignore_errors = false`



# Adding Levenshtein

- Use the Levenshtein algorithm to filter the list of suggestions considered
- The Levenshtein distance is a metric for measuring the amount of difference between two sequences. The value of this property is expressed as a percentage that will depend on the length of the misspelled word



# Transformation Rules

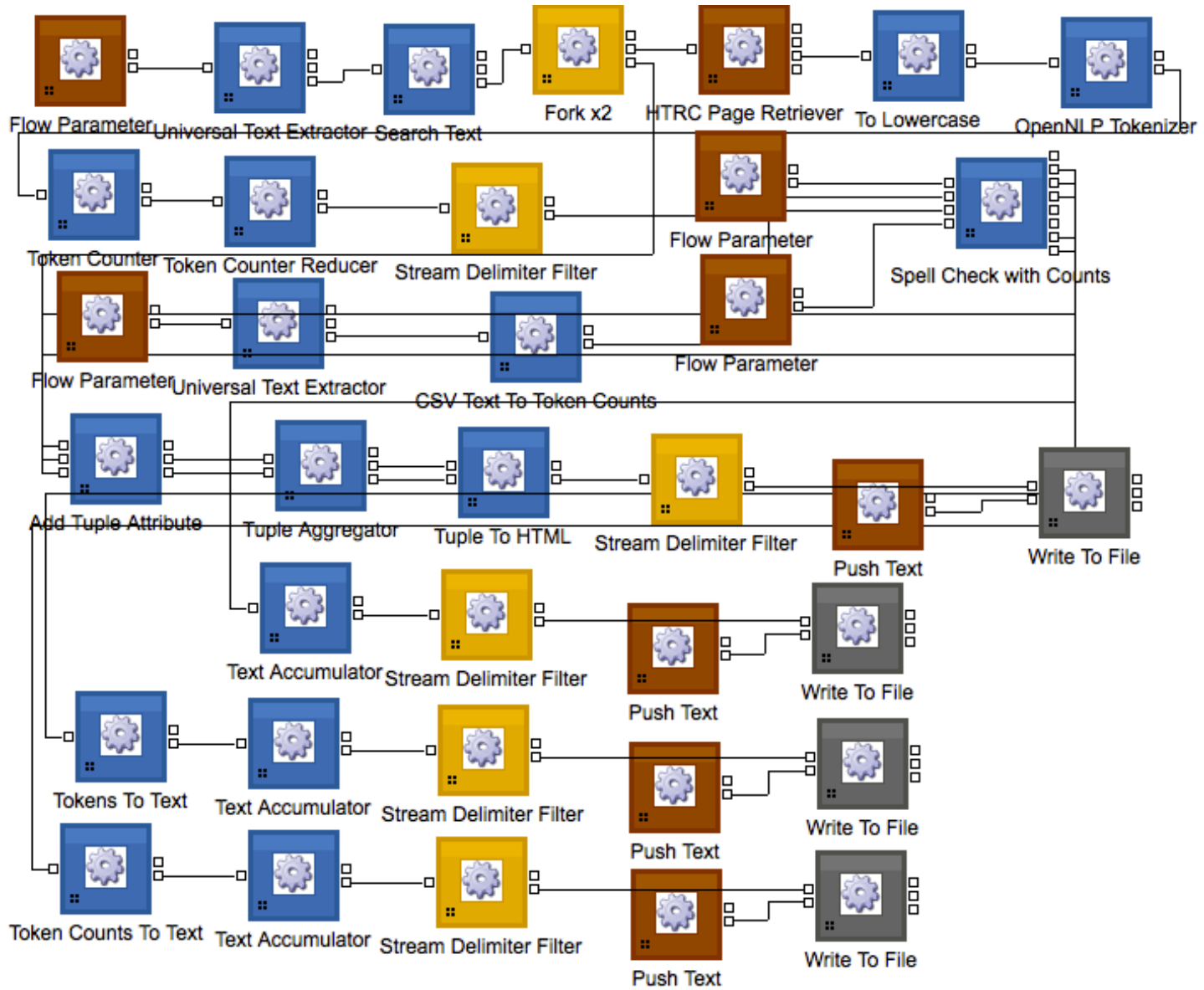
---

## Complete List

o=0; i=1; l=1; z=2; o=3; e=3; s=3; d=3; t=4; e=4;  
l=4; s=0; s=5; c=6; e=6; fi=6; o=6; l=7; z=7; y=7;  
j=8; g=8; s=8; a=9; c=9; g=9; o=9; ti=9; b={h,o};  
c={e,o,q}; cl={ct,d}; ct={cl,d,dl,dt,ft}; d={cl,ct};  
dl=ct; dt=ct; e=c; fl={ss,st}; ft=ct; h={li,b,ii,ll}; i=  
{l,r,t}; in=m; j=y; l={i,t}; li=h; m={rn,lll,iii,in,ni}; n=  
{ll,il,ii,h}; ni=m; oe=ce; r=ll; rn=m; s=f; sh=  
{fli,ih,jb,jh,m,sb}; ss=fl; st=fl; t=l; tb=th; th=tb;  
v=y; u={ll,n,ti,ii}; y={j,v};



# Spell Check Flow



# Demonstration

---

- Dunning Loglikelihood
- Entity Extraction
  - Entity Extraction List
  - Location Entity Extraction
  - Date Entity Extraction
  - Person Entity Extraction
- Topic Modeling
- Spell Checking

