

Lecture: Visualizing Text

DATA ANALYSIS & VISUALIZATION
FALL 2021

*Dr. Muhammad Faisal Cheema
FASTNU*

Text visualization: What-Why-How

What is text data?

Documents

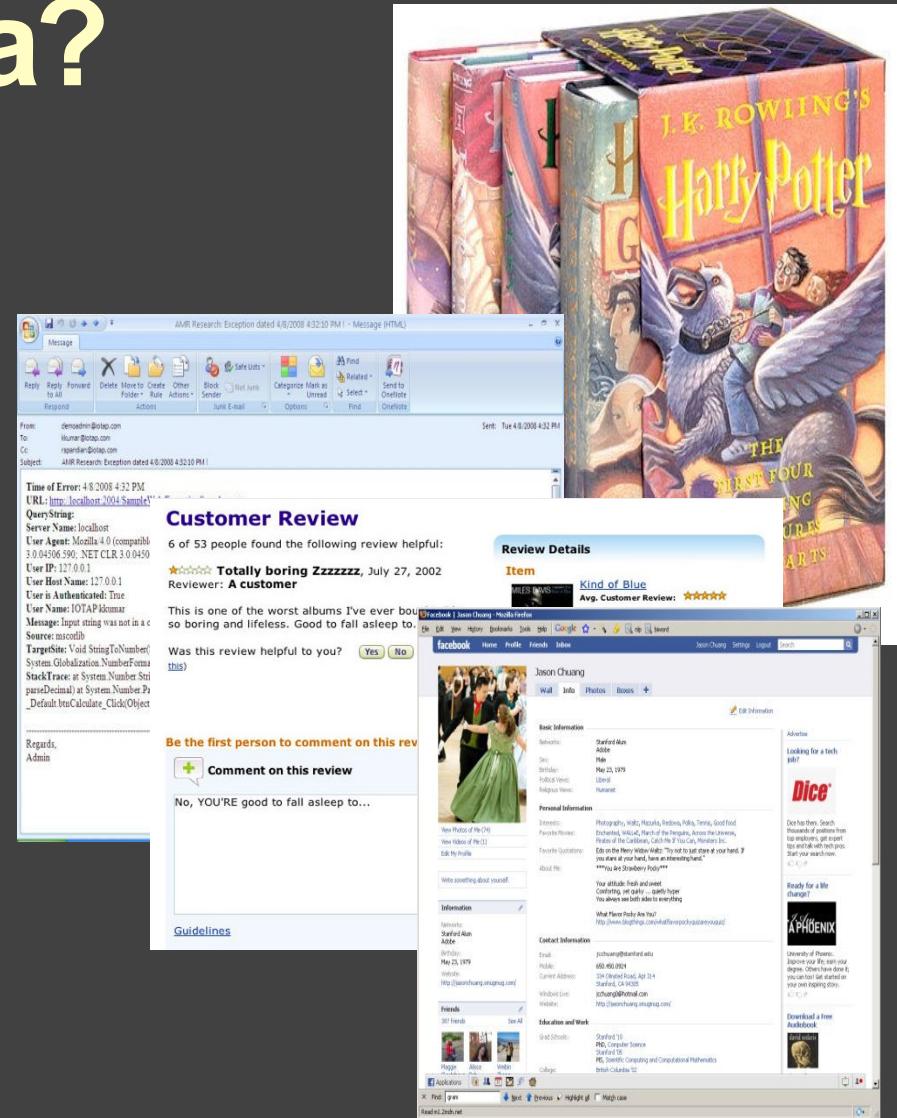
Articles, books and novels
E-mails, web pages, blogs

Text Snippets

Tweets, SMS messages
Tags, comments, profiles

And More...

Computer programs, logs
This slide!
Collections of documents



Why Text is Tough

Text is **not pre-attentive**

Text consists of abstract concepts

which are difficult to visualize

Text represents similar concepts in many different ways

space ship, flying saucer, UFO, figment of imagination

Text has very high dimensionality

Tens or hundreds of thousands of features

Many subsets can be combined together

Text Meaning is NOT pre-attentive

SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC

Why Text is Tough

Abstract concepts are difficult to visualize

Combinations of abstract concepts are even more difficult to visualize

time

shades of meaning

social and psychological concepts

causal relationships

Why Text is Tough

The Dog.



Why Text is Tough

The Dog.



The dog cavorts.

The dog cavorted.

Why Text is Tough

The man.



The man walks.

Why Text is Tough



The man walks the cavorting dog.

So far, we can sort of show this in pictures.

Why Text is Tough



As the man walks the cavorting dog, thoughts arrive unbidden of the previous spring, so unlike this one, in which walking was marching and dogs were baleful sentinels outside unjust halls.

How do we visualize this?

Why Text is Tough

Language only hints at meaning

Most meaning of text lies within our minds and common understanding

“How much is that kitty in the window?”

“how much” social system of barter and trade (not the size of the cat)

“kitty” implies a kitten here

“in the window” implies behind a store window, not really inside a window, requires notion of window shopping

Why Text is Tough

**General categories have no standard ordering
(nominal data)**

Categorization of documents by single topics misses important distinctions

Consider an article about

NAFTA

The effects of NAFTA on truck manufacture

The effects of NAFTA on productivity of truck manufacture
in the neighboring cities of El Paso and Juarez

Why Text is Tough

Other issues about language

ambiguous (many different meanings for the same words and phrases)

different combinations imply different meanings

Why Text is Tough

I saw Pathfinder on Mars with a telescope.

Pathfinder photographed Mars.

The Pathfinder photograph mars our perception of a lifeless planet.

The Pathfinder photograph from Ford has arrived.

The Pathfinder forded the river without marring its paint job.

Why Text is Easy

Text is easier when you have a lot of it

Highly redundant

Because people are good at finding associations, just about *any* simple algorithm can get “good” results for coarse tasks

- Pull out “important” phrases

- Find “meaningfully” related words

- Create “summary” from document

Major problem: Evaluation

People usually **search on relatively coarse meanings**

Why Text is Easy

Pretty much any simple technique can pull out phrases that seem to characterize a document
E.g. Most frequent words from an example lecture:

109	slide	69	to	37	view	37	version	37	graphic	37	first
37	back	36	previous	36	next	32	of	31	the		
30	recall	28	relevant	27	precision	25	retrieved	25	documents		
21	and	18	evaluate	15	a	13	what	13	vs	13	how
12	trec	12	is	12	high	12	for	10	relevance		
10	queries	10	on	9	information	8	x	8	why		
8	as	8	answer	7	search	7	maron	7	document		
7	blair	6	top	6	results	6	measure				
6	length	6	in	6	evaluation	6	curves				

Why Text is Easy

Same text, removing most frequent words in language and most frequent in this text:

30	recall	28	relevant	27	precision	25	retrieved	25	documents
18	evaluate	13	vs	12	trec	12	high	10	relevance
10	queries	9	information	8	x	8	answer	7	search
7	maron	7	document	7	blair	6	top	6	results
6	measure	6	length	6	evaluation	6	curves		

These words can act as a simple summary of the document

people are good at inferring the relations
redundancy in the word meanings

Why visualize text?

Understanding – read a document

Summaries – get the “gist” of a document

Clustering – group together similar contents

Quantify – convert to numerical measures

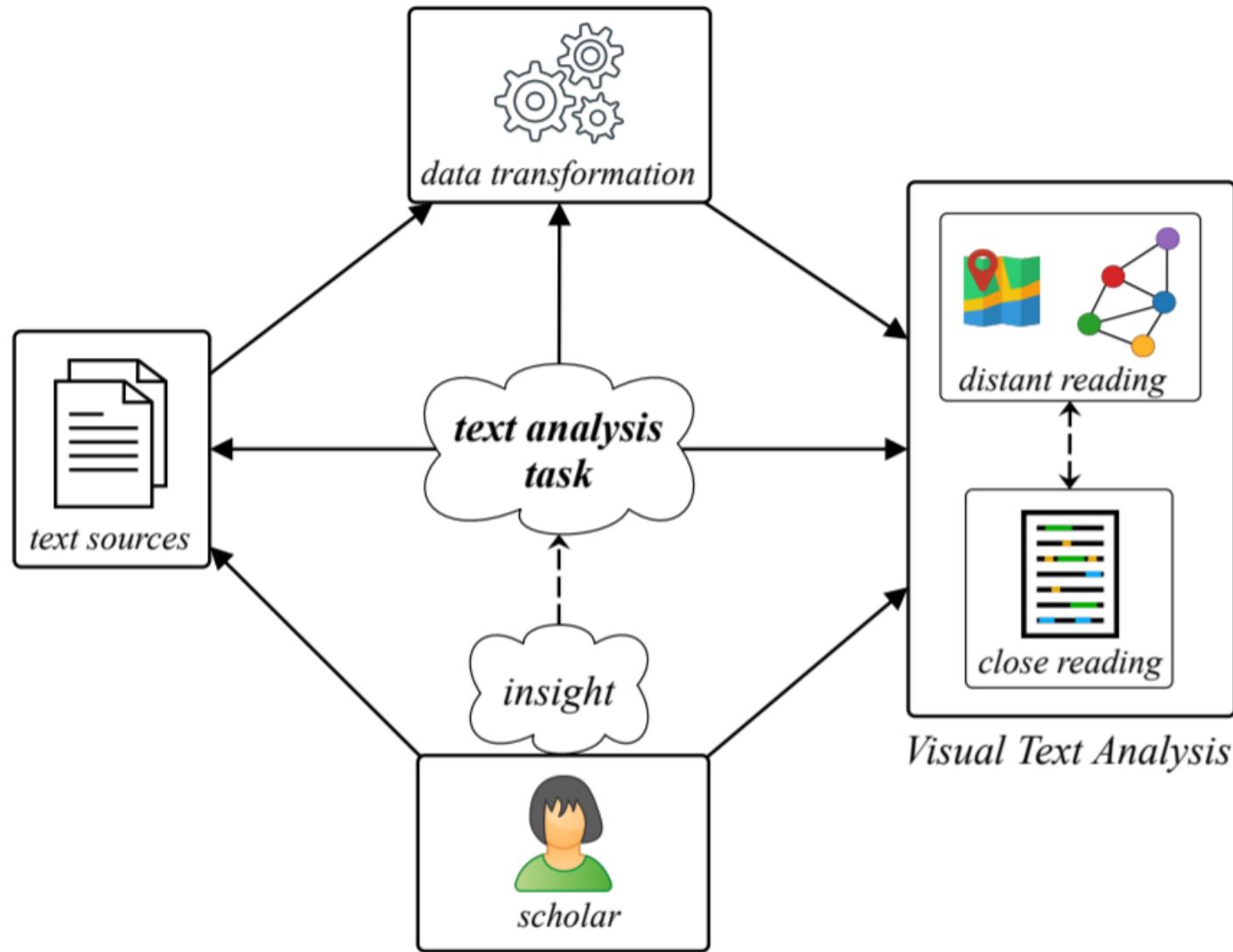
Correlate – compare patterns in text to those in other data, e.g., correlate with social network

Why visualize text?

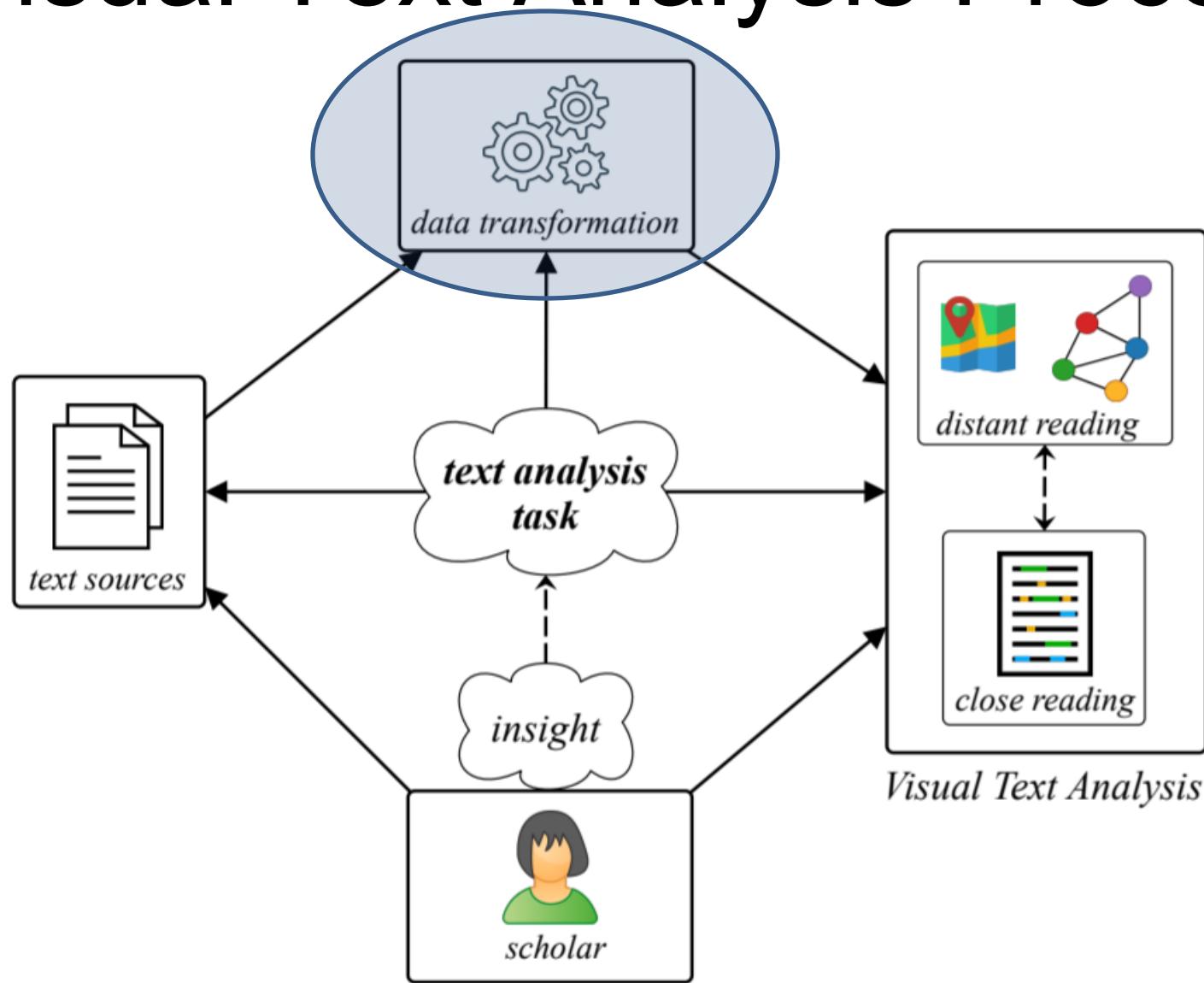
=

Visual Text Analysis Task

Visual Text Analysis Process



Visual Text Analysis Process



Data/Text Transformation

In other words ..

Text pre-processing

Example of a text pre-processing process

- **Tokenization of text**
- **Stemming of words**
- **Ordered lists of words**
- **Bag of word model**
- **Document vector space model**

Text pre-processing steps

Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#gocard, @stanfordfbball, Beat Cal!!!!!!!*

Entities? *San Francisco, O'Connor, U.S.A.*

Stemming

Group together different forms of a word.

Porter stemmer? *visualization(s), visualize(s), visually* → *visual*

Lemmatization? *goes, went, gone* → *go*

Ordered list of terms

Bag of Words Model

Ignore ordering relationships within the text

A document \approx vector of term weights

- Each dimension corresponds to a term (10,000+)
- Each value represents the relevance For example, simple term counts

Aggregate into a document-term matrix

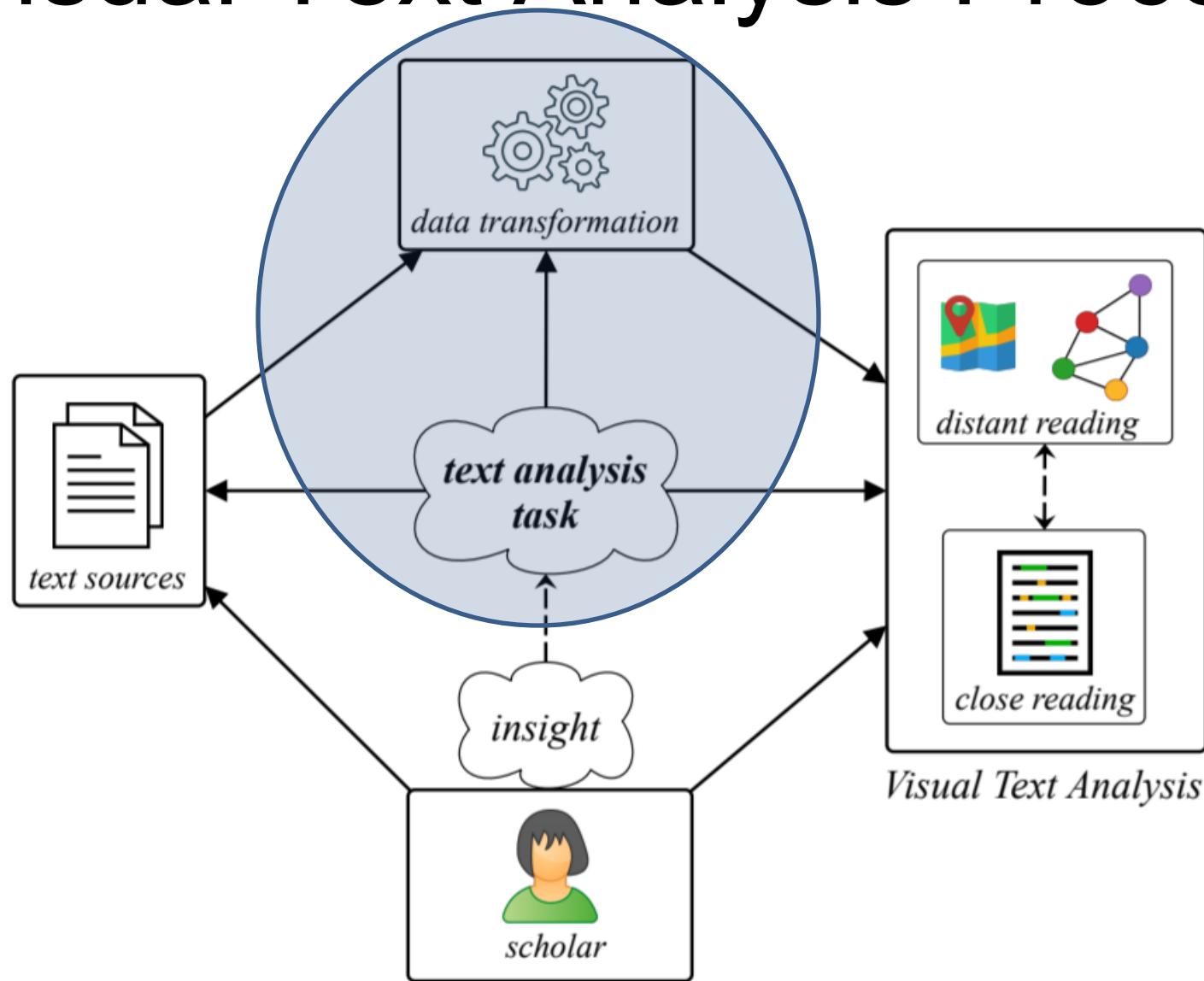
- Document vector space model

Document-Term Matrix

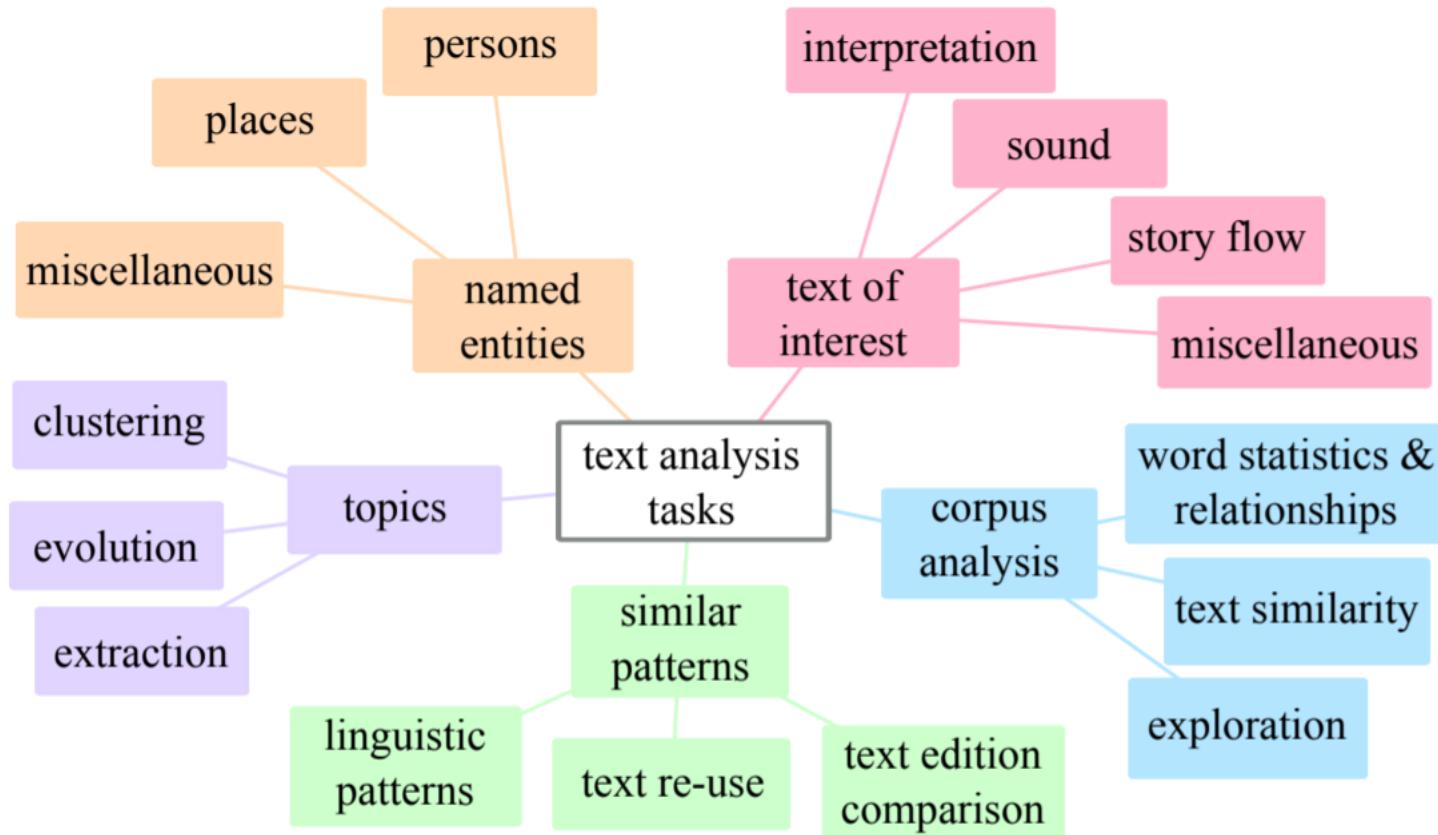
Each document is a vector of term weights
Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Visual Text Analysis Process



Taxonomy of Text analysis tasks



Text analysis tasks

Named Entities

The analysis of named entities is a common text analysis task to extract places (fictional or reported geographies), persons / characters in a story or miscellaneous entities from a single text or a whole collection of text.

- Persons
- Places
- Miscellaneous

Topics

The analysis of topics consists of tasks like topic extractions, so that major topics in the source texts can be tracked and topic-related text passages can be discovered. The presence of temporal data allows for the analysis of topical evolution, and on the basis of the found topics, a topical clustering of a corpus is possible.

- Clustering
- Evolution
- Extraction

Text analysis tasks

Corpus Analysis

An important text analysis task is explore text corpora containing a high number of texts. Tasks such as the analysis of word statistics & relationships among text, text similarity between the texts etc. play important role in corpus analysis tasks

- Word statistics & similarity
- Text similarity
- Exploration

Text of interest

Some tasks focus an individual literary work, which we call text of interest. The underlying research tasks vary from visualizing text interpretations to the analysis of sound of literary works (mostly poems), Interpretation and to the story flow analysis of a given source text.

- Sound
- Story flow
- Miscellaneous

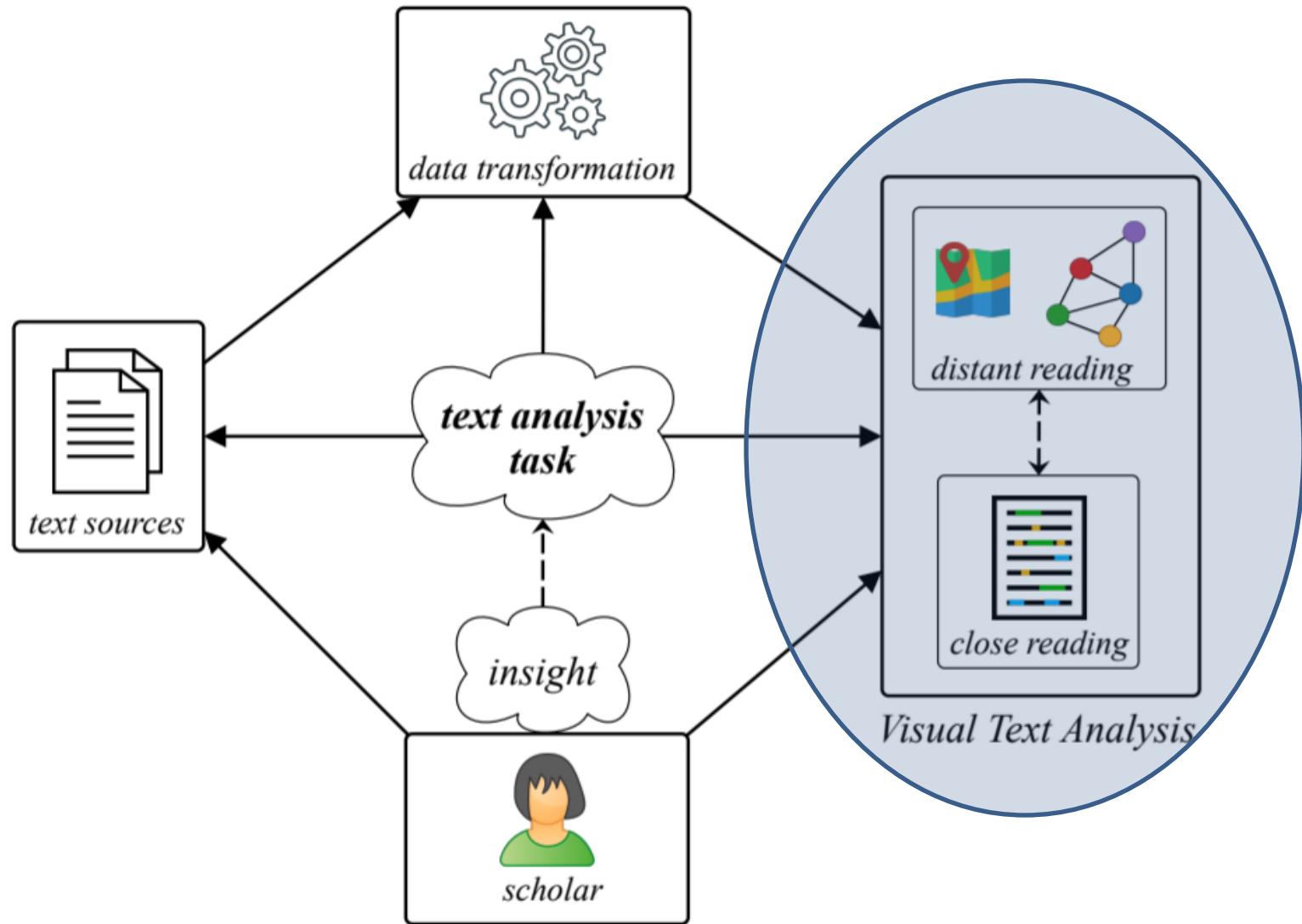
Text analysis tasks

Similar Patterns

An analysis of similar patterns that includes the discovery, the alignment and the visualization of similar text segments among the texts of a given collection is a typical text analysis task. Dependent on the length of patterns, the tasks are divided into three sets. While the analysis of linguistic patterns concerns short phrases, text re-use analysis focuses on determining deliberately re-used text segments (e.g., quotes or plagiarized passages)., whereas text editions focus on comparison of whole text documents.

- Linguistic Patterns
- Text re-use
- Text edition comparison

Visual Text Analysis Process



Visual Text analysis

Close reading

Close reading is the thorough interpretation of a text passage (usually single document) by determination of central themes and the analysis of their development. E.g.

- individuals, events, and ideas, their development and interaction
- used words and phrases
- vocabulary (word frequency, distribution, structure)
- semantic structure
- content

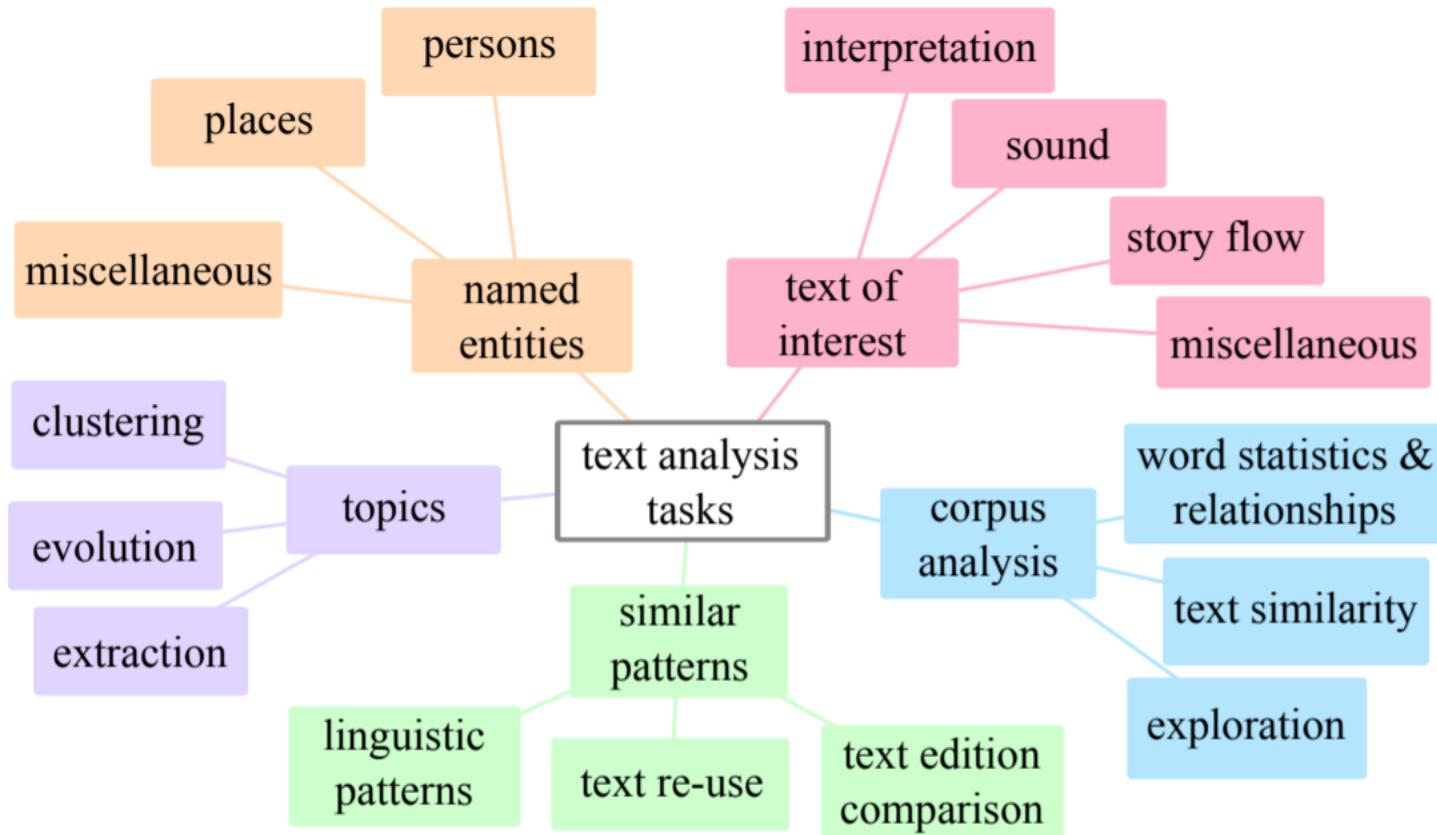
Visual Text analysis

Distant reading

While close reading retains the ability to read the source text without dissolving its structure, distant reading does the exact opposite. It aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s). E.g.

- document themes
- changes over time
- document relationships
- document similarity

Lets map text analysis tasks to Close and Distant reading visualizations



Close reading visualization techniques

IMPORTANT: Focus is on highlighting actual TEXT

- Colors
- Font sizes
- Glyphs
- Connections

Close reading techniques

Color

This elegant shell occurs very rarely on the coasts of this country; we have observed it sparingly distributed on the sands near Tenby, in Pembrokeshire. Da Costa says, he was informed that it is found near Bangor, among the rocks from Bangor Ferry to Anglesea, in Wales, by which he could only mean that the species is an inhabitant of the Menai, the arm of Beaumaris bay, communicating with the St. George's channel which divides Caernarvonshire from the island of Anglesea. The same writer notes it likewise from Cornwall. Dr. Pultney describes it as a scarce shell, which he had found at Weymouth. Having Da Costa's specimens of this shell, and also that of his *Pectunculus Vetula* before us, we should not refrain from observing, that the opinion of Dr. Pultney respecting these shells is incorrect; they are not merely transitions in growth, or varieties of the same kind, the difference between the two is obvious, and fully authorize us to consider them as distinct species. It should be understood in advancing this remark, that the shell which Da Costa figures and describes, for *Pectunculus Vetula* is clearly the Linnaean *Venus Paphia*, a shell well known as a native of the West Indies, and never found to our knowledge in any of the European seas. Da Costa was aware, after his work had been published, that he had erroneously confounded the variety of *Fasciatus*, Fig. 1. 1. in our Plate, with the West Indian shell; he had conceived the latter to be the same shell in a more perfect condition, and caused it to be engraved accordingly.

This elegant shell occurs very rarely on the coasts of this country; we have observed it sparingly distributed on the sands near Tenby, in Pembrokeshire. Da Costa says, he was informed that it is found near Bangor, among the rocks from Bangor Ferry to Anglesea, in Wales, by which he could only mean that the species is an inhabitant of the Menai, the arm of Beaumaris bay, communicating with the St. George's channel which divides Caernarvonshire from the island of Anglesea. The same writer notes it likewise from Cornwall. Dr. Pultney describes it as a scarce shell, which he had found at Weymouth. Having Da Costa's specimens of this shell, and also that of his *Pectunculus Vetula* before us, we should not refrain from observing, that the opinion of Dr. Pultney respecting these shells is incorrect; they are not merely transitions in growth, or varieties of the same kind, the difference between the two is obvious, and fully authorize us to consider them as distinct species. It should be understood in advancing this remark, that the shell which Da Costa figures and describes, for *Pectunculus Vetula* is clearly the Linnaean *Venus Paphia*, a shell well known as a native of the West Indies, and never found to our knowledge in any of the European seas. Da Costa was aware, after his work had been published, that he had erroneously confounded the variety of *Fasciatus*, Fig. 1. 1. in our Plate, with the West Indian shell; he had conceived the latter to be the same shell in a more perfect condition, and caused it to be engraved accordingly.

Close reading techniques

Font size

Once upon a midnight dreary, while I pondered weak and weary,
Over many a quaint and curious volume of forgotten lore,
While I nodded, nearly napping, suddenly there came a tapping,
As of some one gently rapping, rapping at my chamber door.
"Tis some visitor," I muttered, "tapping at my chamber door -
Only this, and nothing more."

Ah, distinctly I remember it was in the bleak December,
And each separate dying ember wrought its ghost upon the floor.
Eagerly I wished the morrow; - vainly I had sought to borrow
From my books surcease of sorrow - sorrow for the lost Lenore -
For the rare and radiant maiden whom the angels named Lenore -
Nameless here for evermore.

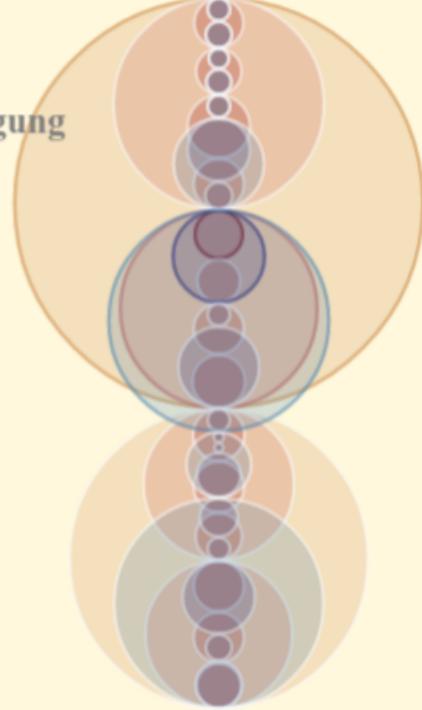
Once upon a midnight dreary, while I pondered weak
and weary,
Over many a quaint and curious volume of forgotten lore,
While I nodded, nearly napping, suddenly there came a
tapping,

As of some one gently **rap**ping, rapping****
at my chamber door.

"Tis some visitor," I muttered, "**tapping** at my
chamber door -
Only this, and nothing more."

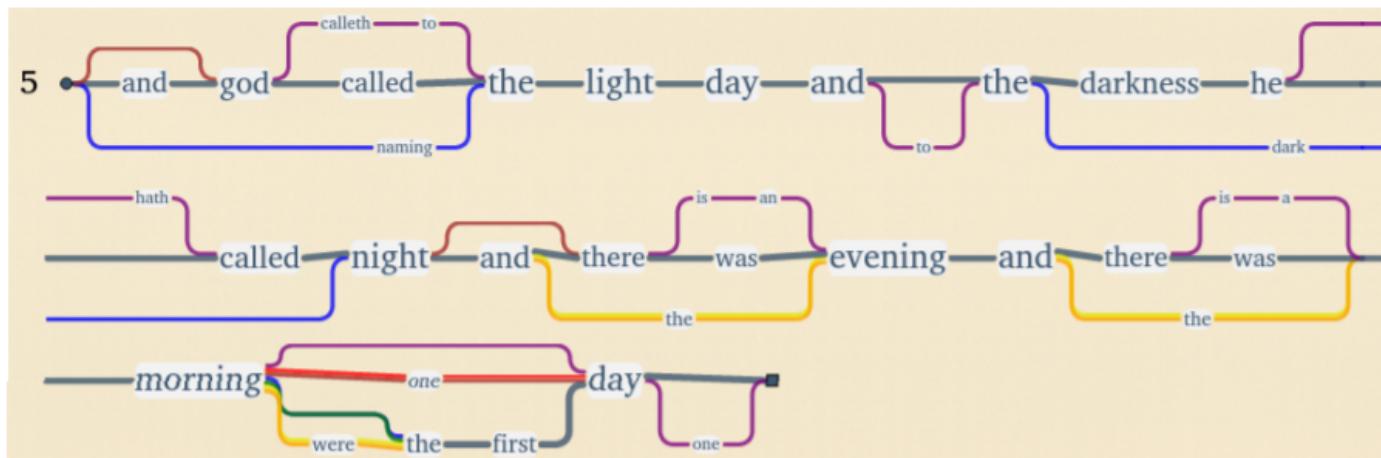
Close reading techniques Glyphs

Rainer Maria Rilke *octave sestet* *line*
Das I. Sonett *quatrain tercet couplet* *phr* *s*



Da stieg ein Baum. O reine Übersteigung!
O Orpheus singt! O hoher Baum in Ohr!
Und alles schwieg. Doch selbst in der Verschweigung
ging neuer Anfang, Wink und Wandlung vor.
Tiere aus Stille drangen aus dem klaren
gelösten Wald von Lager und Genist;
und da ergab sich, daß sie nicht aus List
und nicht aus Angst in sich so leise waren,
sondern aus Hören. Brüllen, Schrei, Geröhr
schien klein in ihren Herzen. Und wo eben
kaum ein Hütte war, dies zu empfangen,
ein Unterschlupf aus dunkelstem Verlangen
mit einem Zugang, dessen Pfosten beben,—
da schufst du ihnen Tempel im Gehör.

Close reading techniques Connections



human events it becomes necessary for a people to advance from that they have hitherto remained, & to assume among the powers of the earth the station to which the laws of nature & of nature's god entitle them, a decent respect of mankind requires that they should declare the causes which impel them to

be sacred & undeniable, that all men are created equal & independant, that they derive rights inherent & inalienable, among which are the preservation of their life, the pursuit of happiness, and to secure these ends, governments are instituted among men, deriving their just powers from the consent of the governed, that whenever any form of government becomes destructive of these ends, it is the right of the people to alter or to abolish it, and to provide a new government, laying its foundation on such principles & organizing its powers in such a manner as will seem most likely to effect their safety and happiness. Prudence, indeed, will suggest that mankind are more disposed to suffer while evils are suffered than to abolish the forms to which they are accustomed, but when

the unanimous Declaration of the thirteen united States of America

When in the Course of human events, it becomes necessary for one people to dissolve political bands which have connected them with another, and to assume, among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation.

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty, and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, that whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to provide a new Government, laying its foundation on such principles and organizing its powers in such a manner as will seem most likely to effect their Safety and Happiness. Prudence, indeed, will suggest that Governments long established should not be changed for light and transient causes; accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are suffered than to abolish the forms to which they are accustomed, but when

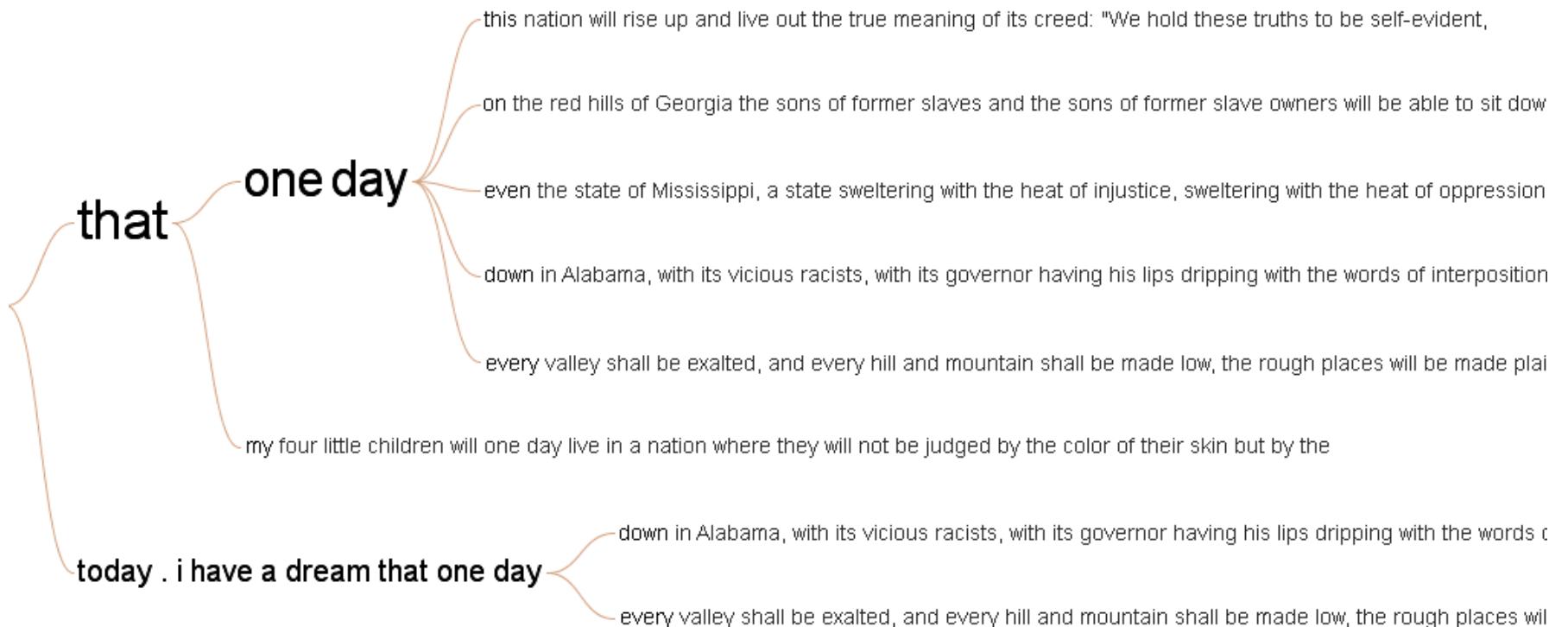
TextArc (One Document)



TextArc

- Three rules:
 - Show the entire text in an ellipse around the page: line-by-line and word-by-word
 - Like tag clouds, use larger font-size and brighter text for frequent words
 - Central words move to the middle (links to its mentions)

Word Tree (One Document)

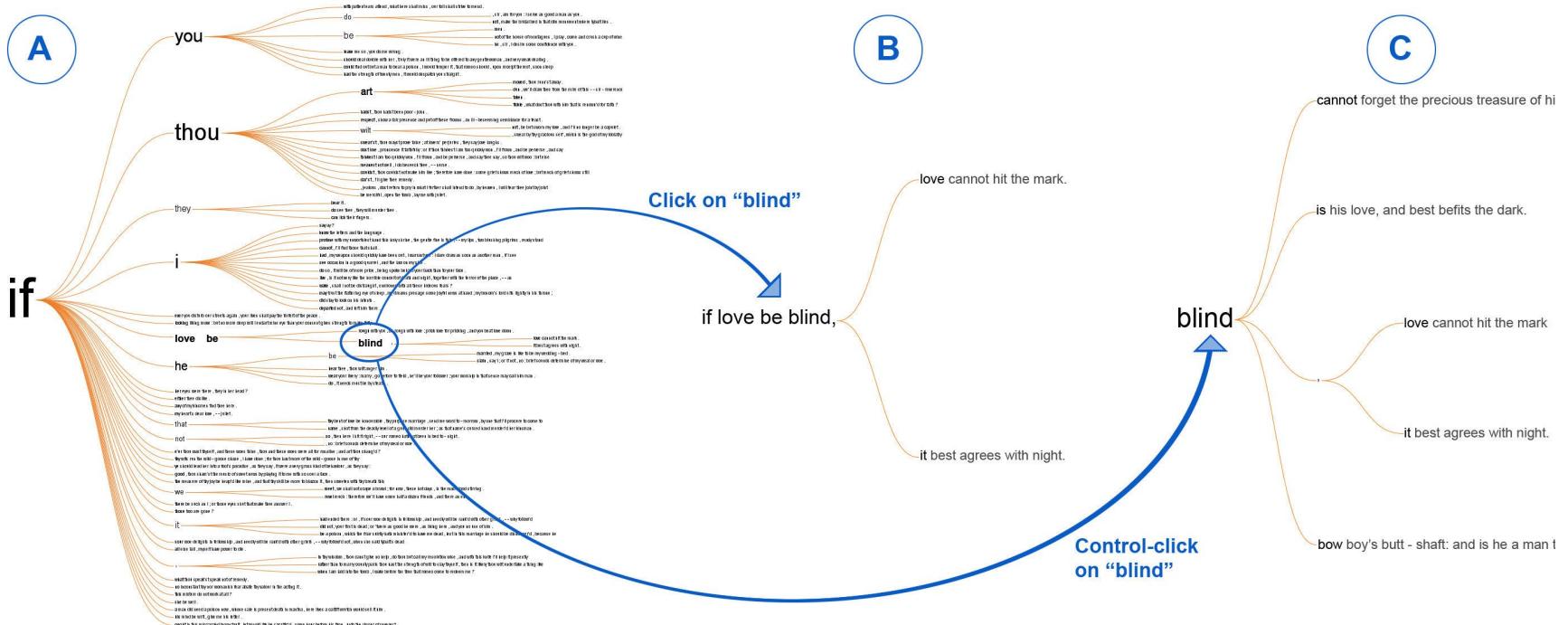


[Wattenberg & Viegas, 2007]

Word Tree

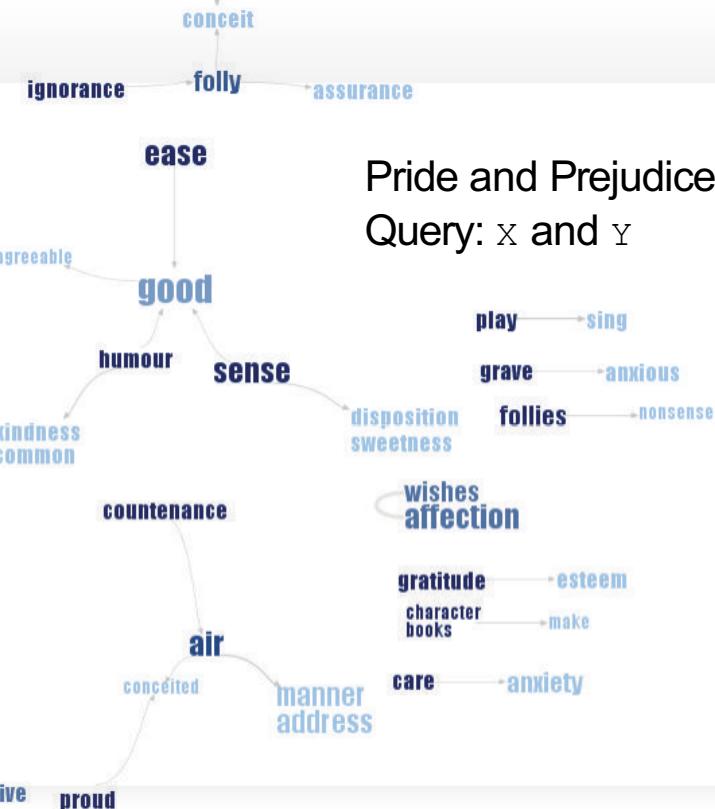
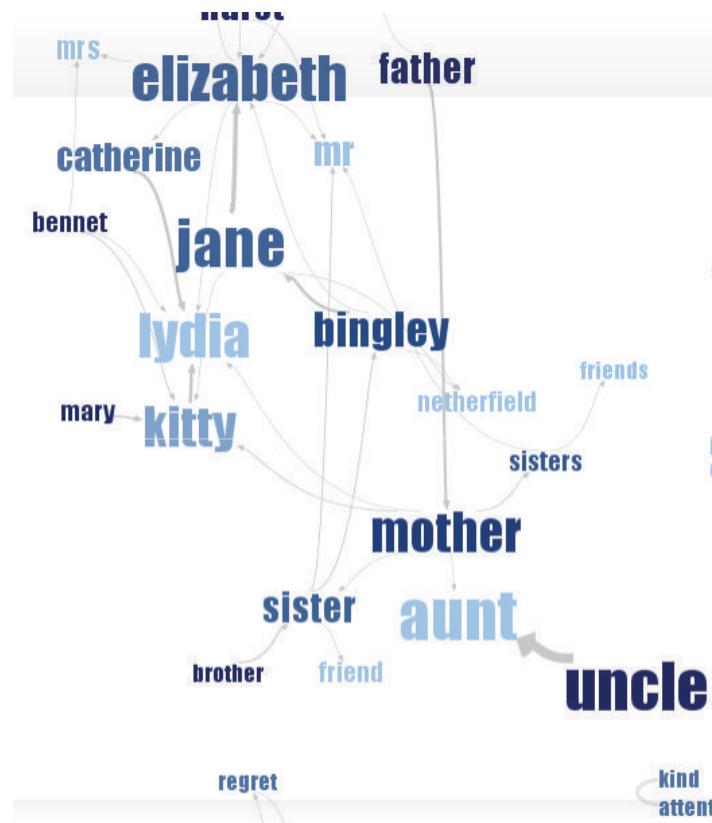
- A "Visual Concordance"
- Shows phrasing, relationships between words
- Starting point is a single word or snippet
- Branches to show common words/phrases that follow
- Goal is to show context: "keyword-in-context"

Interaction in Word Tree



[Wattenberg & Viegas, 2007]

Phrase Nets



Pride and Prejudice
Query: X and Y

[van Ham, 2009]

Words are more than just character sequences

Fed Drapes

Clark Coolidge

FELL FAR BUT THE BARN (came) up & smacked me
Who're you, bleeding? Fled.
Blat in back of a Vistrola Car
is so red is such that sun
fell in the rushes & pen bear appear

the white wrong numeral on the wall
can't take if off with the clock
down with the clock it ...
way
on the board - couch with brass, kindergarten clenched joints
backed violet rip into the gas valve
it hemmed & snowed

the wrong way
remnant face
rubber
the pucker

Rhymes

Phonetic Rhymes

Character Clusters

Levenshtein Distance

Identical Rhyme/Rhyme Riche ●

AAR AES
Perfect Rhyme ● ● ●
AEKT

Semirhyme ..

Syllabic Rhyme

Consonant Slant Rhyme ●

Vowel Slant Rhyme ● ● ● ● ● OW1

Pararhyme

Eye Rhyme

Alliteration ● ● ● ● ● KL1

Assonance ● ● ● ● ●

Poemage

- Support close reading—in-depth reading to generate as much productive meaning as possible
- Search for poetic devices: affect, imagery, pun, metaphor
- Sound and linguistic devices → Rhyming
 - Identical: pare/pair
 - Perfect: picky/tricky
 - Assonance & consonance: blue/estuaries, shell/chiffon
 - Eye rhyme: cough/bough
- Support exploration: scholars do not want computers to "solve" poems

Interface

Poemage v 0.1

Poem View

Machinations Calcite
Clark Coolidge

acetone imprinted
oblique swatch on the skin car barn oil wall
ocarina & mumps
much wet green
I'd leave sole key to this game to my friend, sheep water cat

acted impressed
weaving candle turn on computer cigarette paper wall
tarheels & balance
a lot of yellow stick neck
He'll have to hurry & carry away, to my blue friend hustling bringing
his moon & car

agate inked
merry melodies drool on chalk of wet lead star tool
crayon & sands
length of granite duck - drill
It's sucking up the strand, his crystal flag, & the eels tube for that,
their parade wizzed fun

arctic duck
splinter dry -ice spazzduke- ing ace supper at church
hard pinks & sponge breath
many forarms drift

Roller window going up on I repeat my offer food list in iron flakes

Path View

Modes: 1 2 3 shuffle nodes

Set View

SONIC RHYMES

Identical Rhyme/Rhyme Riche AET

Perfect Masculine

Perfect Feminine

Perfect Dactylic

Semirhyme

Syllabic Rhyme

Consonant Slant Rhyme K

Vowel Slant Rhyme

Pararhyme

Syllabic 2 Rhyme

Alliteration SW

Assonance

Consonance SH CH SK

clear beautiful mess

hover word show uncertainty custom set

show words show context fill intersecting paths

CONTEXT SLIDER

Poemage v0.1

Distant reading visualization techniques

- Heat maps
- Tag clouds
- Maps
- Time lines
- Graphs
- Miscellaneous

Distant reading techniques

Heat maps



Distant reading techniques

Tag clouds



Tag Cloud (One Document)

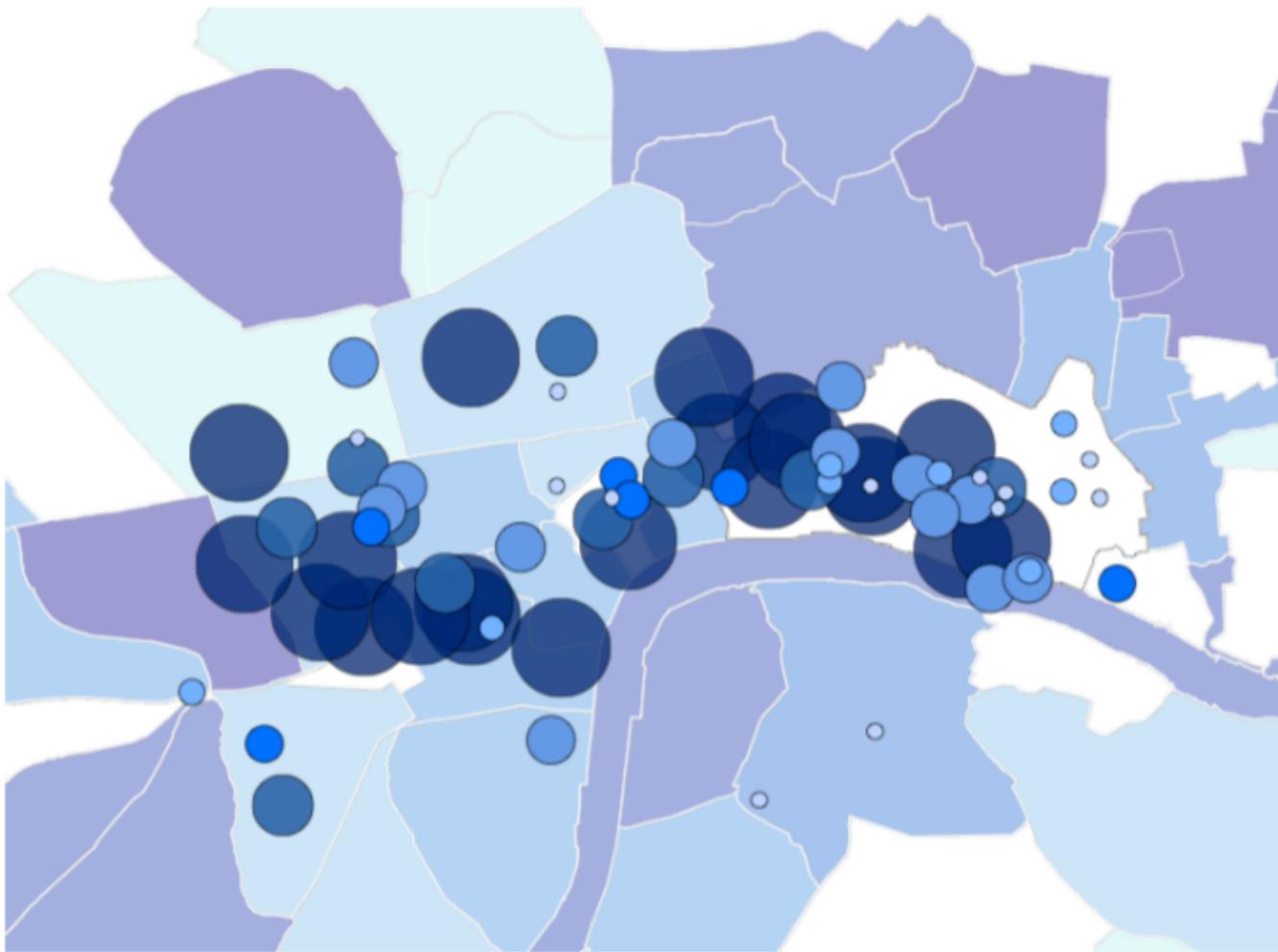
- Derived data: number of occurrences of words
 - Channel: Font size
 - Potential problem: Think about ink...



[Scray, CC-BY-SA-3.0]

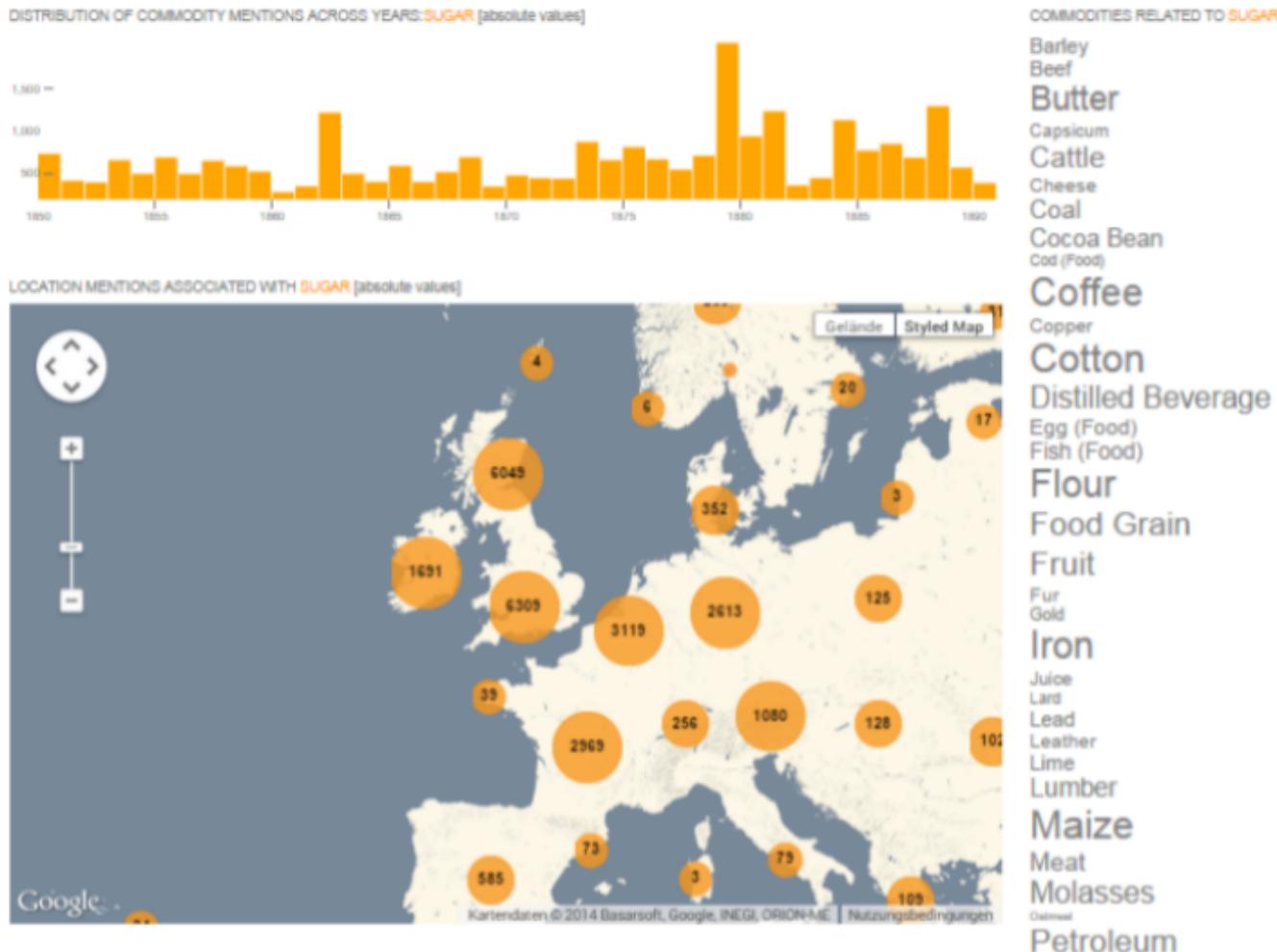
Distant reading techniques

Maps



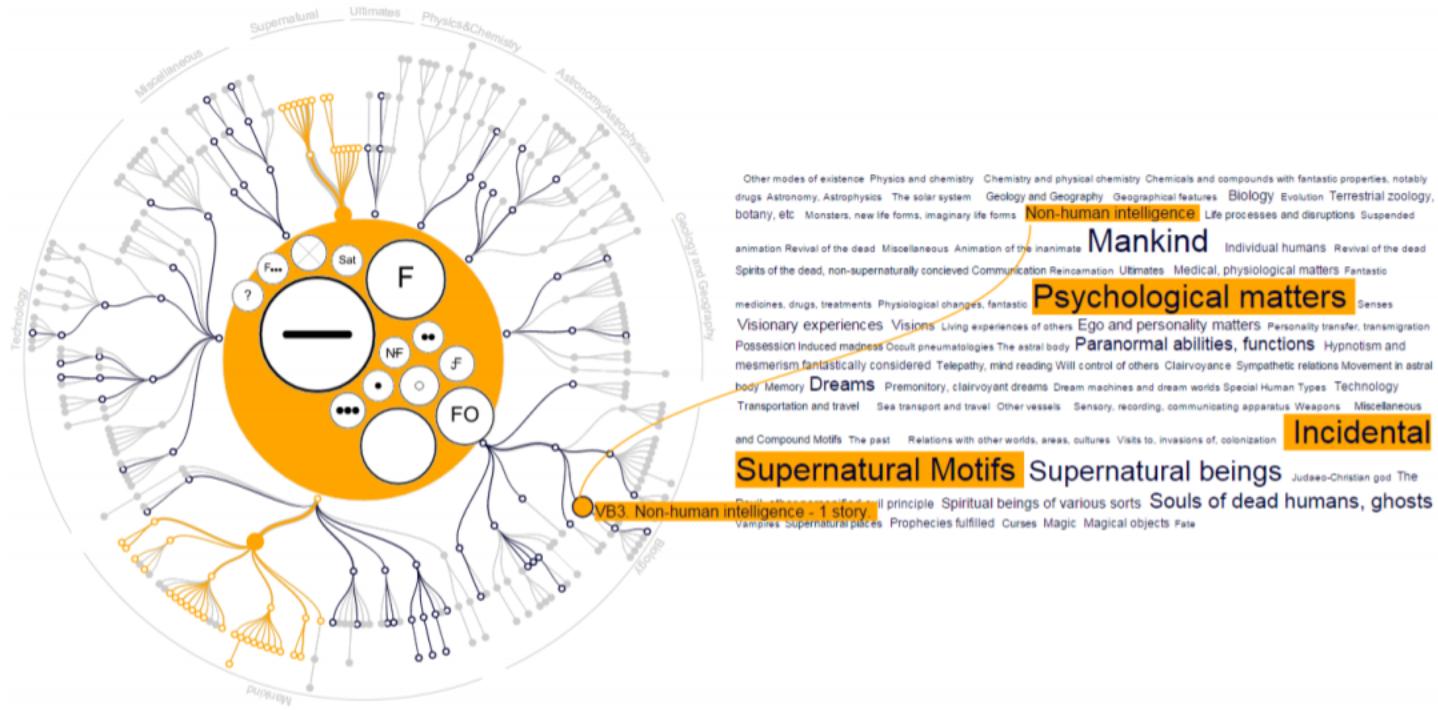
Distant reading techniques

Time lines



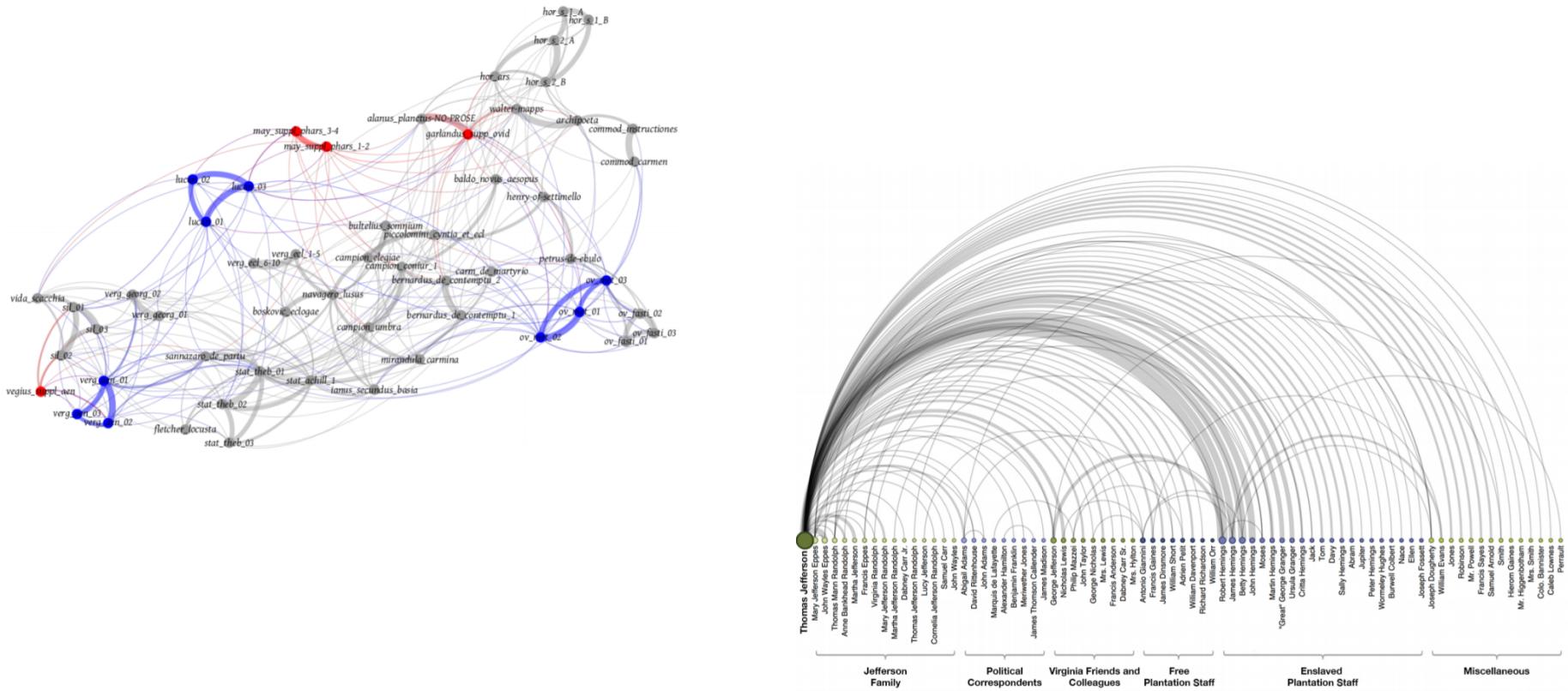
Distant reading techniques

Graphs



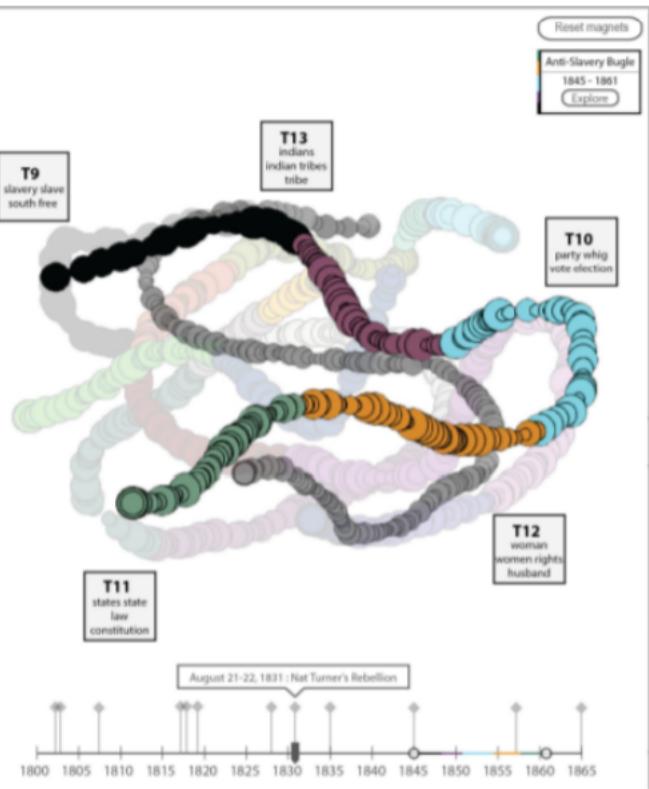
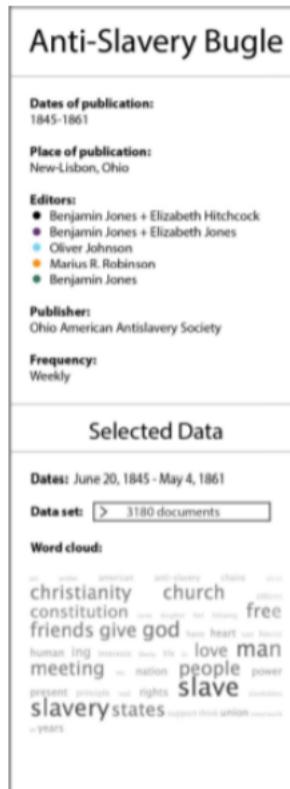
Distant reading techniques

Graphs ...



Distant reading techniques

Miscellaneous



Comparing Documents

- Word choice/usage
- Relationships
- Phrasing

Tag Cloud (Two Documents)

State of the Union Address, 2002 vs. 2011

act afghanistan allies
american attack best budget
camps children citizens coalition
congress continue corps **Country** create
danger depend destruction develop economy encourage
enemies evil extend fight free **freedom**
government health help history home homeland
hope increase islamic **jobs** join lives mass
military moment months **nation** opportunity
people police power protect rebuild
regimes resolve retirement **security**
spending states tax terror
terrorists thank thousands
together tonight training true united
war ways weapons women
work workers **world**

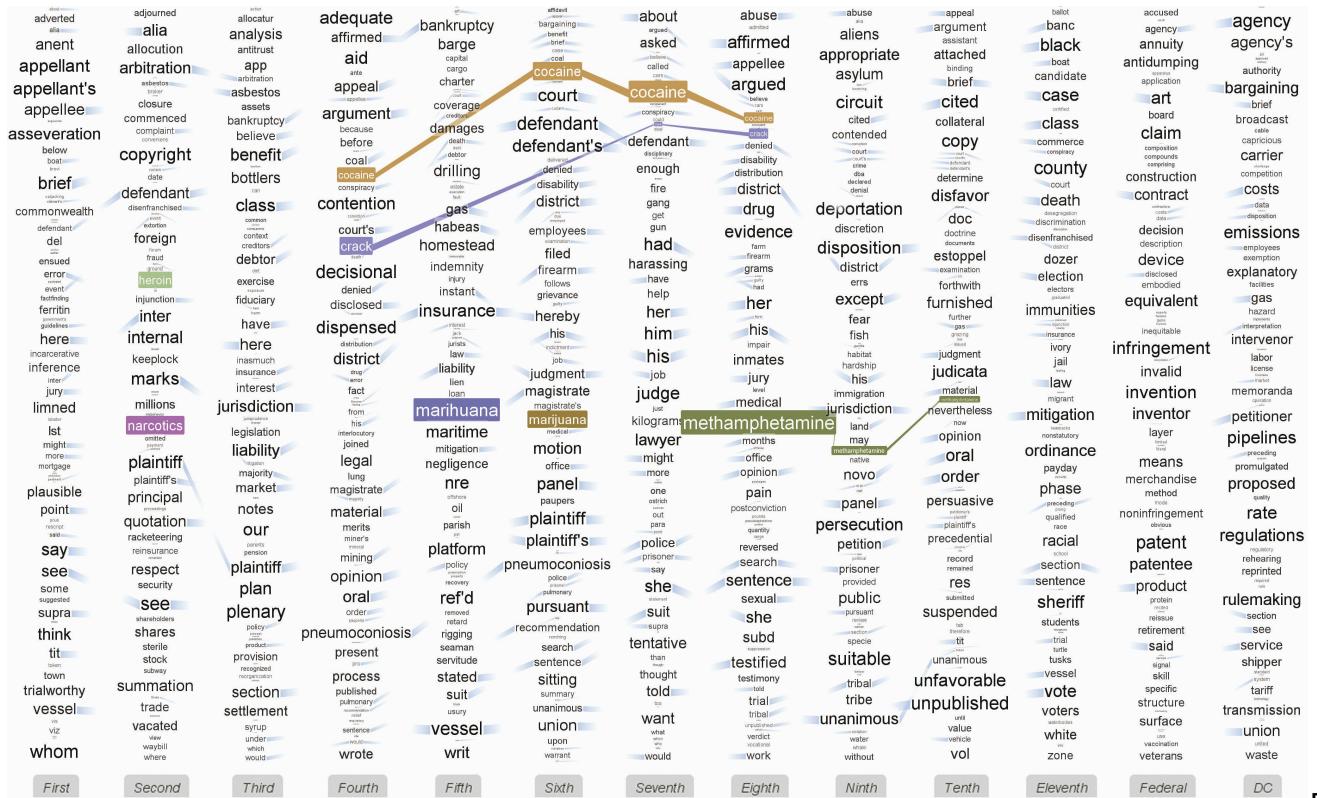
President Bush, January 29, 2002

afghan ago already **american** behind
believe best better building
care century challenge chance change child children clean
college company compete **congress country**
create cuts deficit democrats different don done
dream economy education energy family
future generation give goal
government health help home idea
innovation internet invest **jobs** law
life live money **nation** passed
people percent possible projects race reform
republicans research responsibility schools
spending states step students success
support sure **tax** teachers technology things together
tonight troops willing win **work** workers
world years

President Obama, January 25, 2011

[Pyrsmis, CC-BY-SA-3.0]

Parallel Tag Clouds (Multiple Documents)



[Collins et al., 2009]