Week 6 – Visualizing Text Documents

Yasmin AlNoamany

L&S 88-2

University of California, Berkeley

Project pitches

Projects

1. 6°
2. Medical cannabis dispensary density and the opioid overdose death rate. What is the relationship?
3. Boston Marathon Run Times
4. Marriage rates from various parts of the country
5. Weather's Impact on Crime Rates
6. Brexit on Twitter: A Tale of Emoji
7. Social Media followers on Coca Cola Stock (KO)
8. Relationship between Local Finance, Education, and Poverty

Announcements

- Assignment 2 grades posted
  - Solutions will be sent out soon
- Assignment 3 discussion
  - If you have questions, post them here
  - Start your assignment early!
- Oct. 25: I'll be on travel
  - Make up lecture by Alberto Cairo will be giving from 4-5:30pm in the I School in South Hall (either room 210 or 202)
- Visiting lecture on Oct. 31th
  - From Google and other tech companies

Project

- 35% of your total grade
- Check out the project page carefully and go through the example projects.

- Remember that this is a teamwork project, not an assignment!

- Grading:
  - Is the visualization an effective representation of the data?
  - Does the visualization support different analytical questions about the data?
  - Is the visualization creative and does it illustrate some new ideas?
    - While it is not necessary to invent some new visualization technique for the project, designs that illustrate creativity and new thinking will generally be viewed positively. Of course, innovation cannot be a total substitute for utility.
  - Was your demonstration an effective presentation and illustration of your project and work?

- I'll send feedback on project ideas. You also can come and talk to me about your project

Feedback discussion

- Here is the evaluation form
  - Fill it if you haven't done so

- Assignment 1 ave time is 2 hours
- Assignment 2 ave time is 3-4 hours

- Links of resources/readings/etc.
  - Syllabus will be your only reference

- "Hands-on R and Tableau"
  - We covered intro to R in the last class. I hope it was slow and helpful.
  - I'll provide few minutes for Tableau today

- "We want more coding experience":
  - Concepts are important!
  - I designed the class to cover concepts and hands-on experience on most of the topics
  - We can't cover everything in the class, so I generated notebooks for you in the form of tutorials
  - I assigned those as readings for Week 2, 3 so you have time to learn and practice getting data from the web

- "One tool versus multiple tools":
  - We use Python as our main tool for doing most of the tasks − Note, I don't teach python, I teach how to do tasks in Python
  - Other tools are important to learn about so you can choose later which tool suite your needs!
  - This is normal in vis. Classes to feel overwhelmed with different tools

Previous lecture

- Types of data models
- Charts types
- Analyzing and Visualizing data using R and Rstudio
- Introduction to visualization using Tableau

Today's lecture

- Application of Web content mining
- Text representation
- Pre-processing
- Single Document Visualization
  - Work/Tag Clouds
  - Wordle
- Document Collection Visualizations
  - Jigsaw

Resources

- Ch 9 of Interactive Data visualization book - Text and documents
- Working With Text Data – Python tutorial
- Text preprocessing notebooks in python from Stanford
- Optional:
  - Practical Data Mining with Python
  - Chapter 10 of Search User Interfaces
  - Chapter 11 of Search User Interfaces

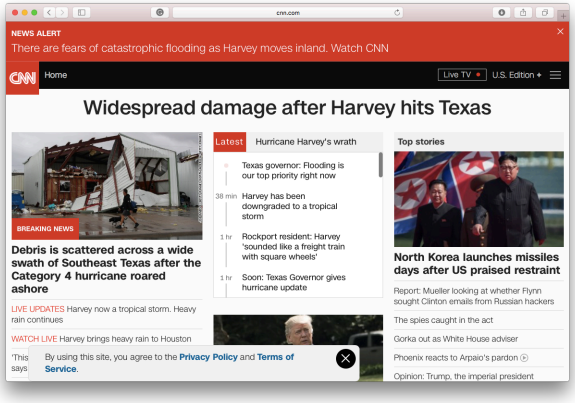
Objectives

- Define corpus and metadata
- Distinguish different levels of text representations
- Describe text preprocessing process
- Name some of the common tasks a user might want to perform on text data
- Name advantages and disadvantages of using word clouds for document visualization
- Name different visualization methods for single document visualization and for document collection visualization

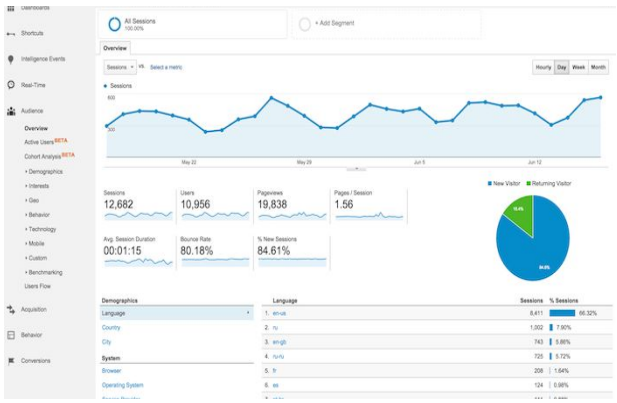
Types of Web Mining

Web Mining

Web Content Mining



Web Usage Mining

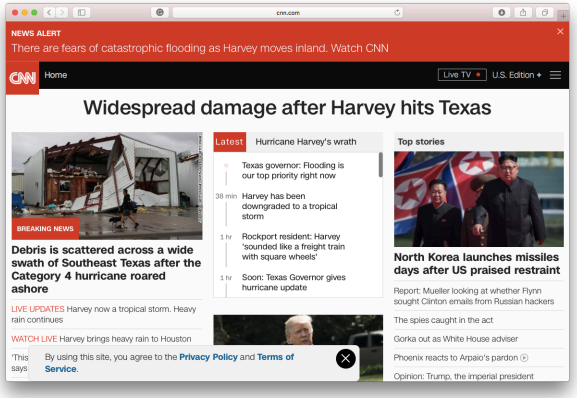


Web Structure Mining

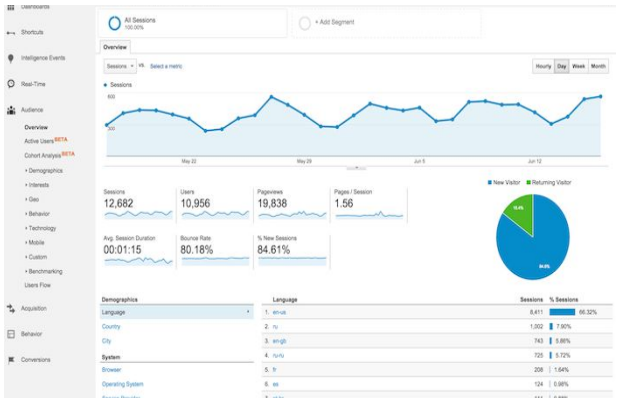
Types of Web Mining

Web Mining

Web Content Mining



Web Usage Mining



Web Structure Mining

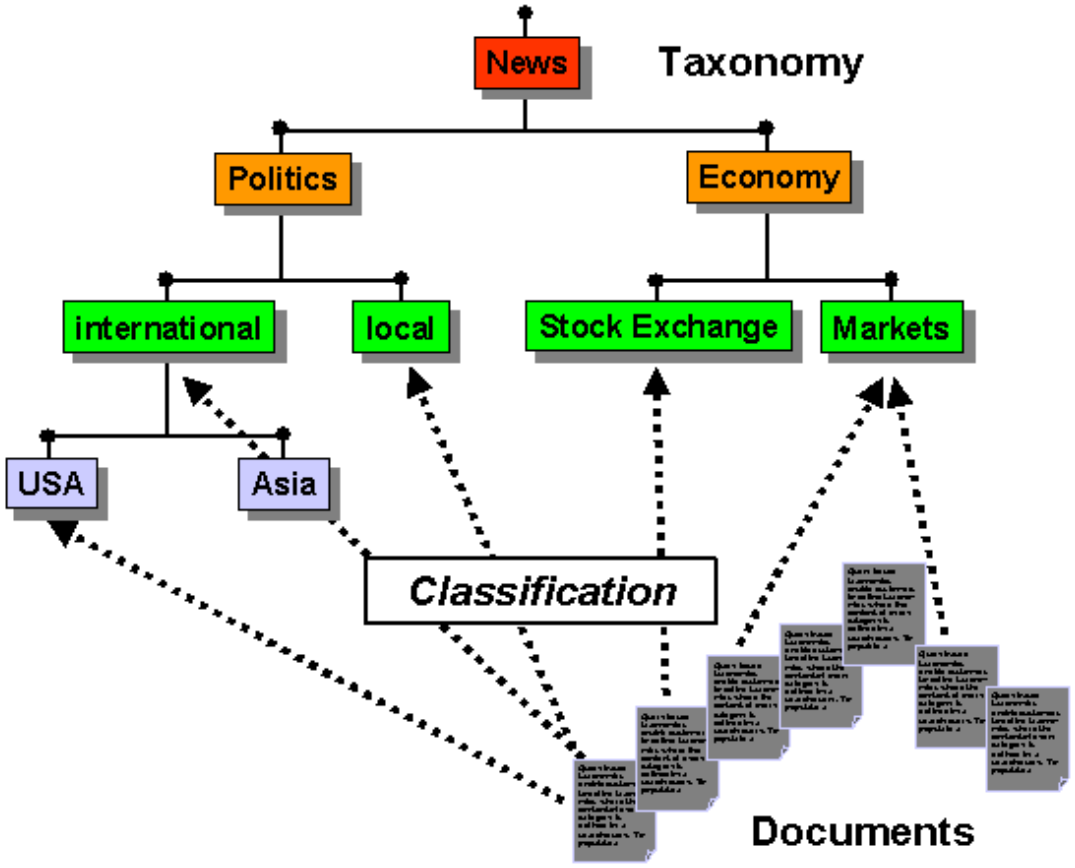
Applications of Web content mining

- Extracting useful knowledge from the contents of Web documents or other semantic information about Web resources

- Applications:
  - document clustering or categorization
  - topic identification/tracking
  - concept discovery
  - focused crawling
  - content-based personalization
  - intelligent search tools

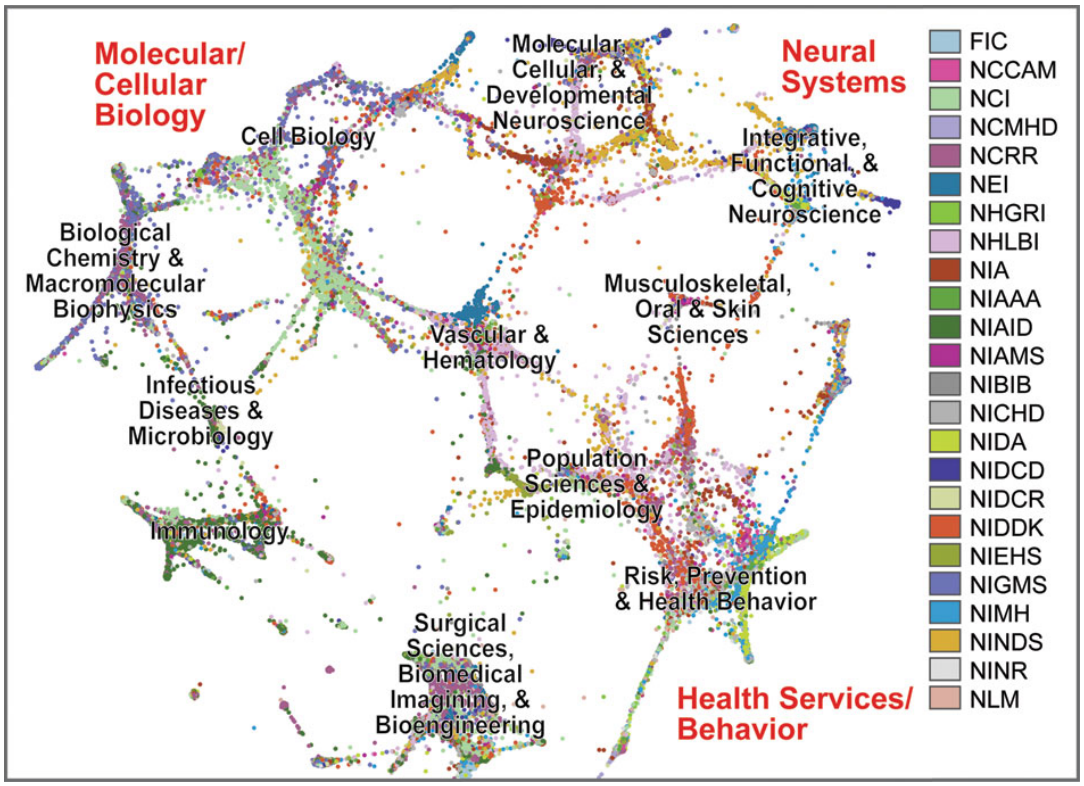
Document categorization

- Adding structures to the text corpus



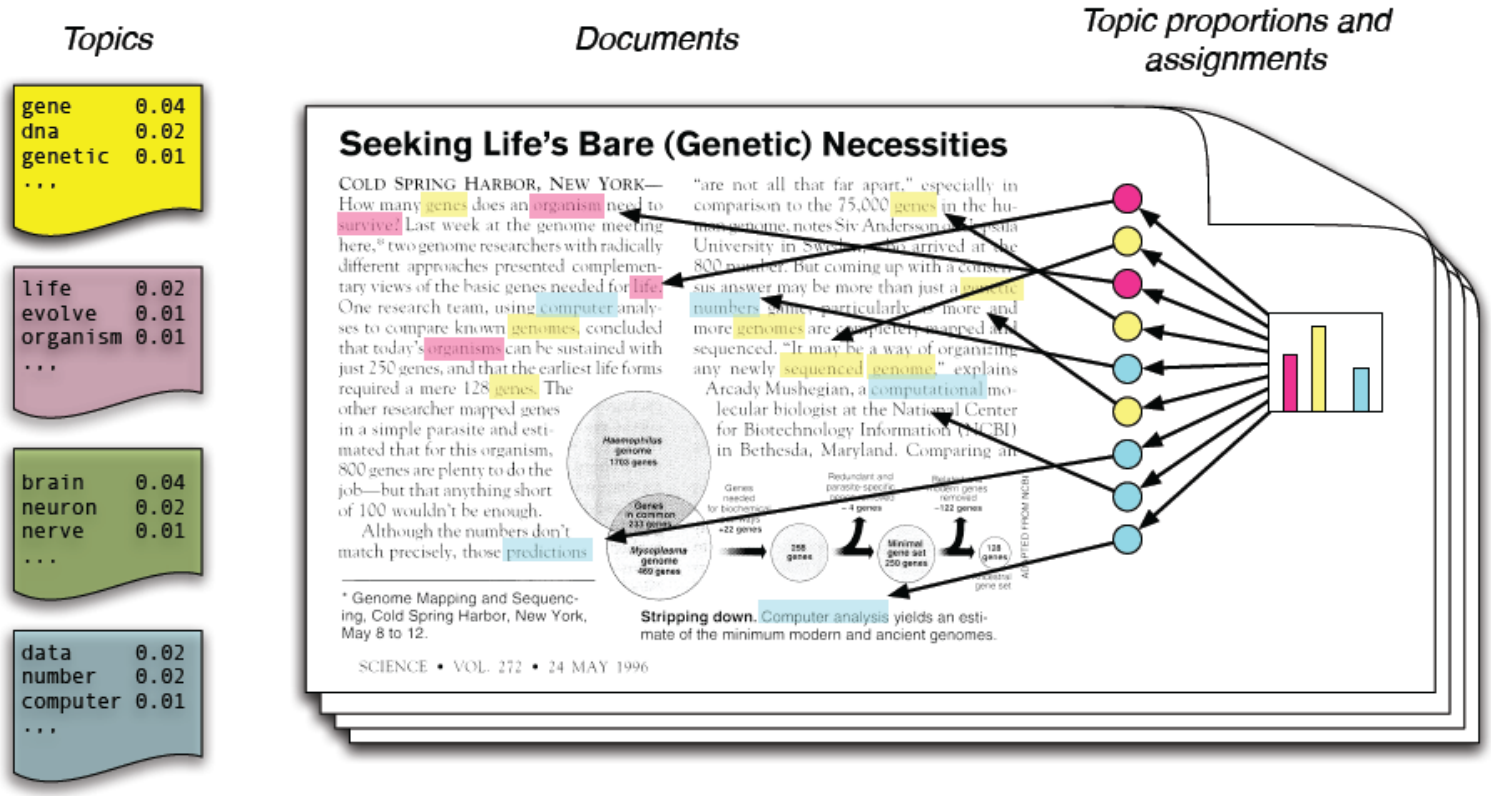
Text clustering

- Identifying structures in the text corpus



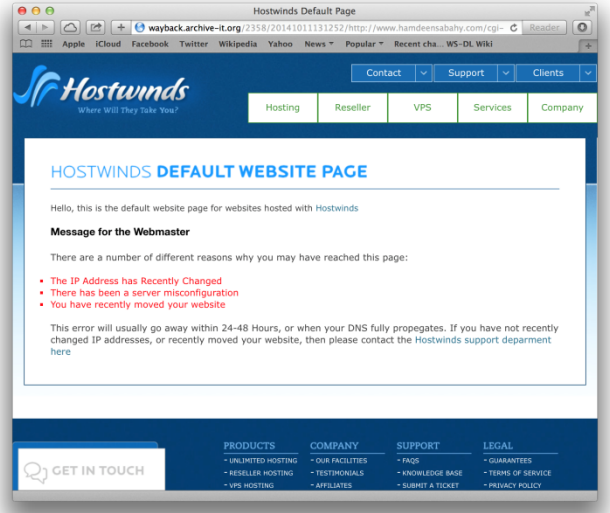
Topic modeling

- Identifying structures in the text corpus



Document similarity

Method	Similarity
cosine	0.7
TF-Intersection	0.6
Jaccard	0.5



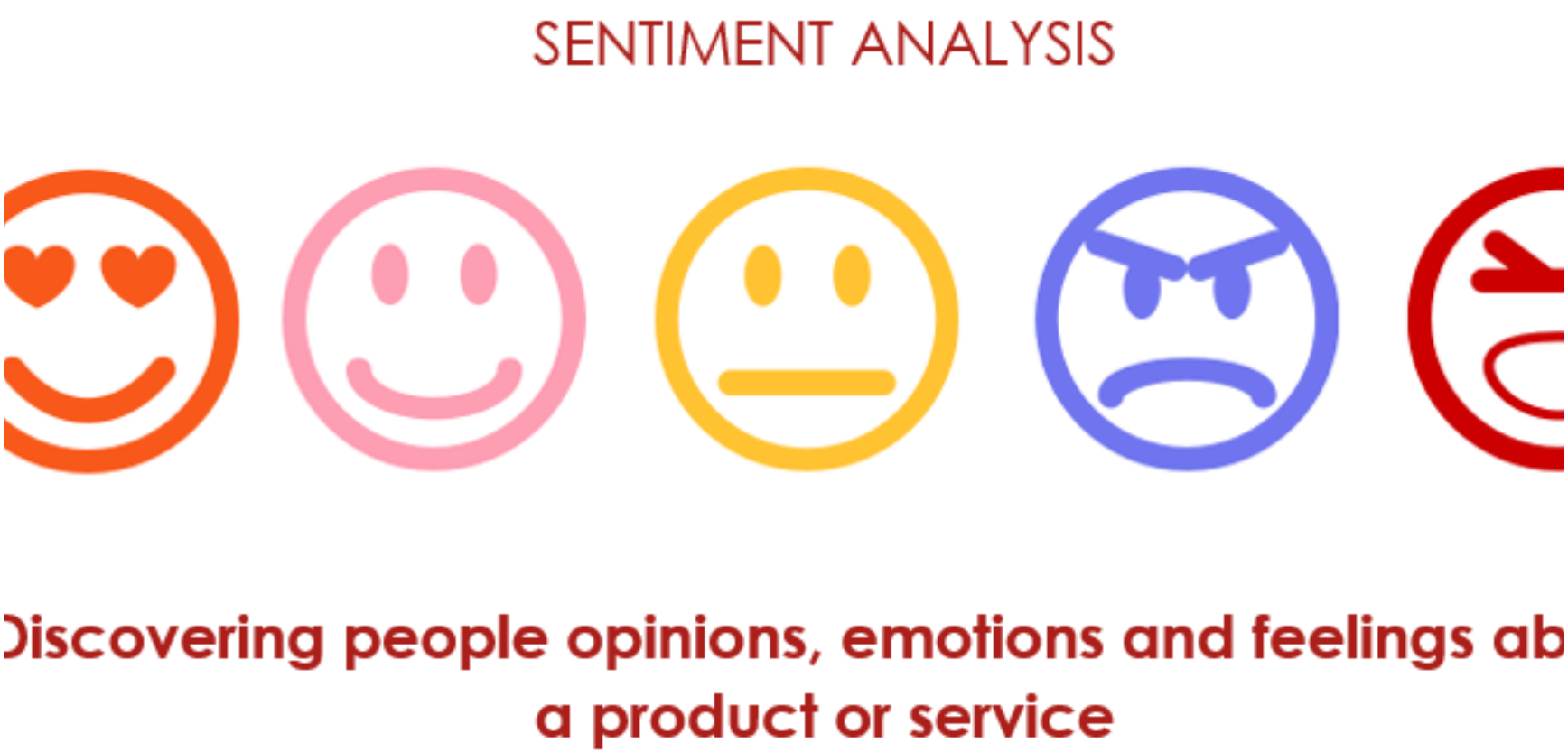
Method	Similarity
cosine	0.0
TF-Intersection	0.0
Jaccard	0.0

Information Retrieval

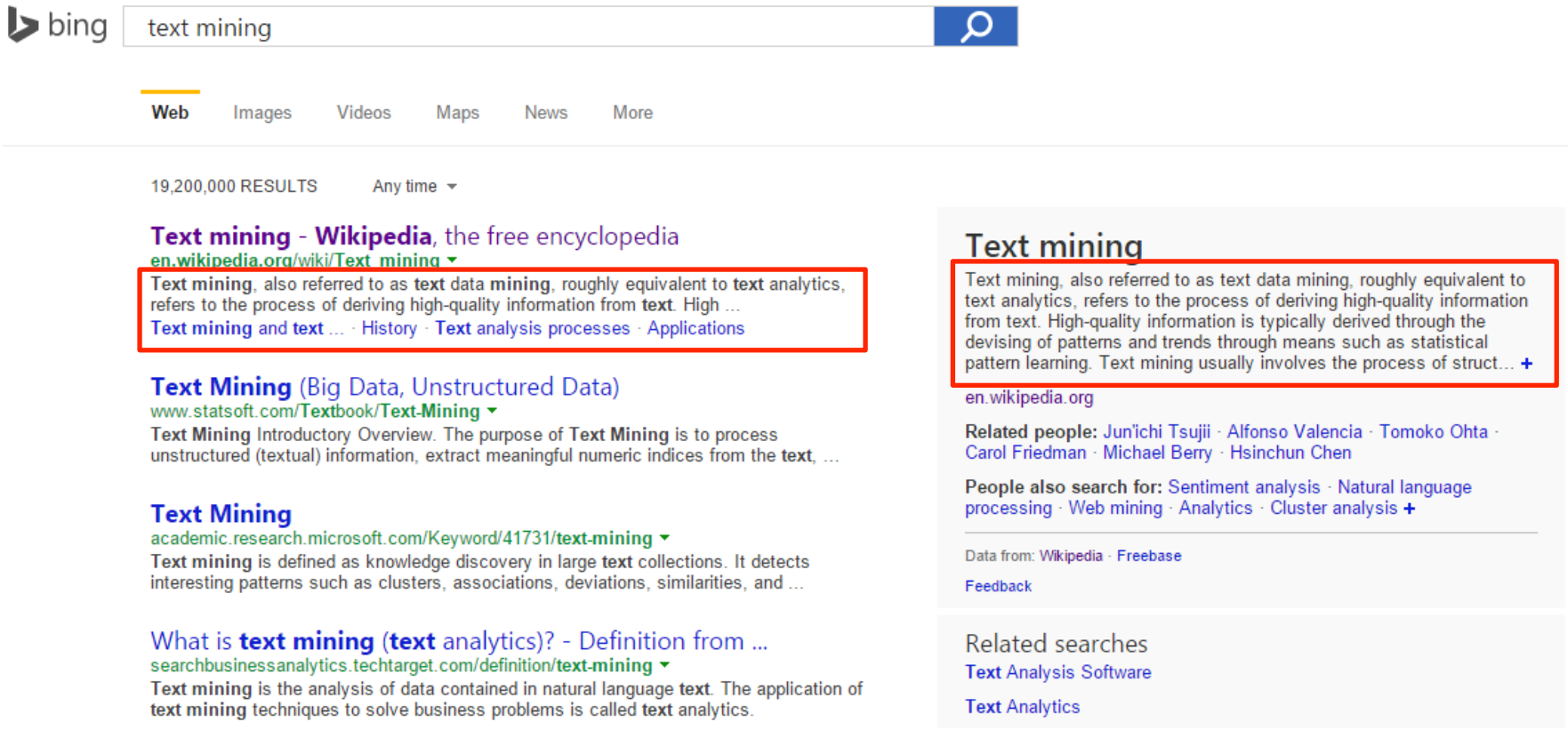
- Information Retrieval (IR) is the search process that locates particular entities based on selection criteria
  - Google search algorithms
  - Library catalog search
- We will **not** discuss IR algorithms

- Information visualization can help to
  - Understand what can be retrieved
  - Understand what has been retrieved
  - Browse
  - Formulate more precise queries

Sentiment analysis



Document summarization



News recommendation

All Stories **News** Entertainment Sports Business More

Flying high: Airstream can't keep up with demand

JACKSON CENTER, Ohio (AP) — Bob Wheeler still gets the question sometimes when people find out he runs the company that builds those shiny aluminum campers: "Airstreams? They still make those?"

Associated Press

North Korea's Internet down again. US spooks at work?

North Korea's web connection to the rest of the world – always sketchy and limited at best – went on the blink again Saturday. Most North Koreans wouldn't have noticed, of course. But

Christian Science Monitor 45 mins ago

Wisconsin man keeps 40-year-old Christmas tree up until son returns

By Brendan O'Brien (Reuters) - A Wisconsin man will refuse for about the 40th time to partake in the annual after-holiday chore of putting Christmas

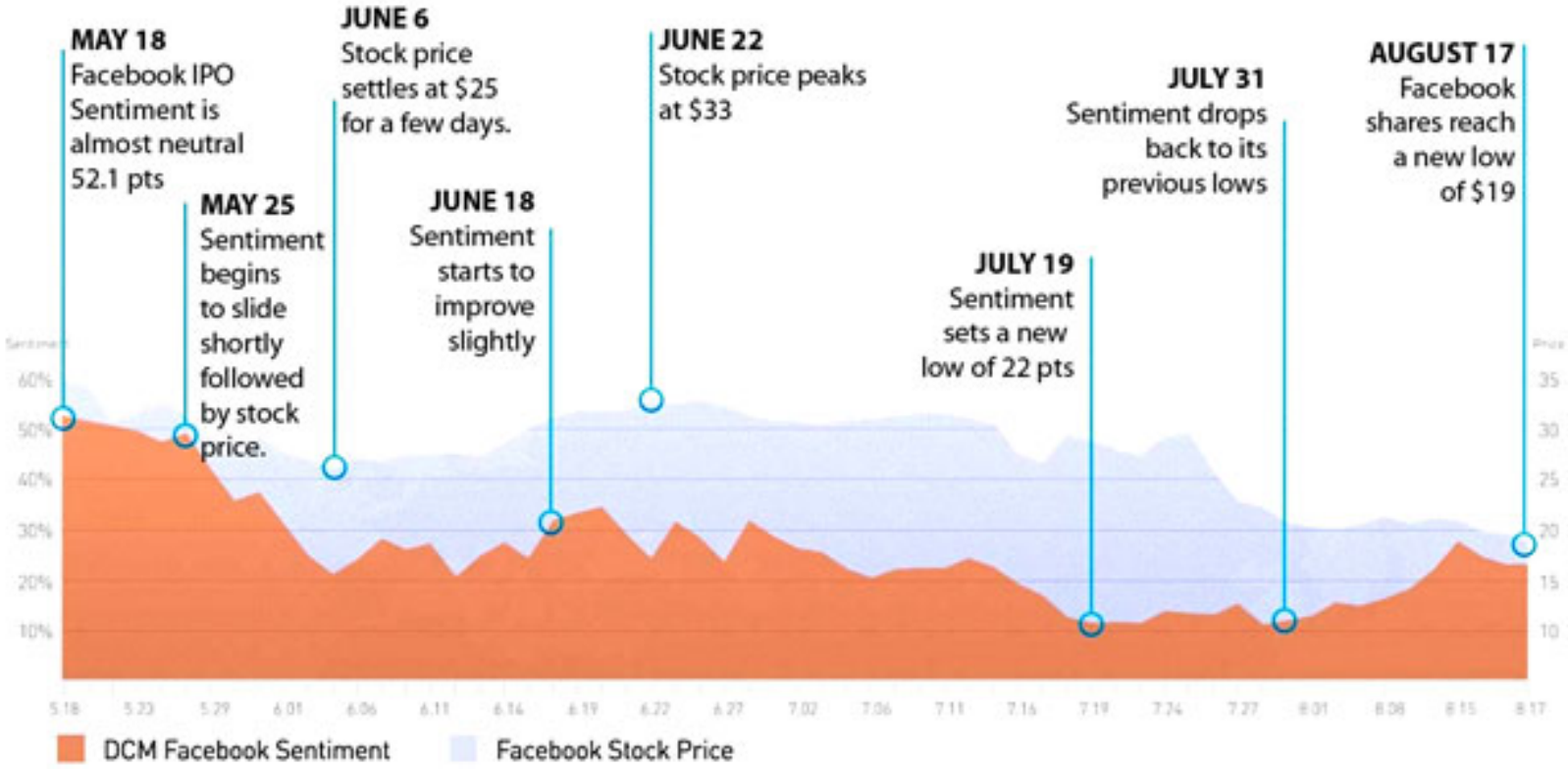
Reuters

Navy Helicopter Drone Completes First Round of Testing

Imagine trying to land a remote-controlled helicopter on top of a motorboat that's speeding across a lake. Navy pilots recently had to contend with just such a scenario as they tested the U.S. military's newest drone, the MQ-8C

LiveScience.com

Text analytics in financial services

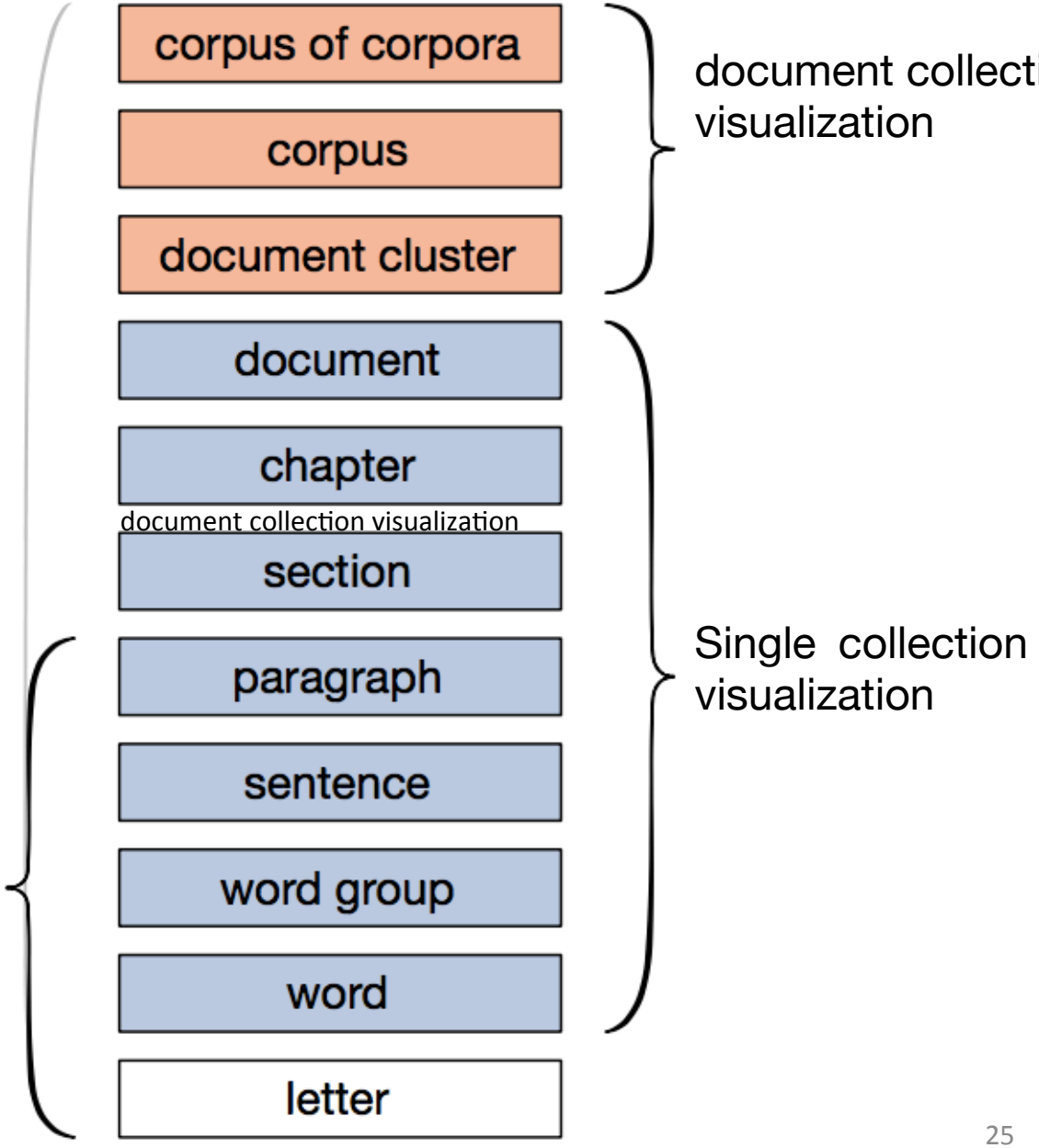


Text visualization

Text data

- Text is Everywhere
  - News articles
  - Speech transcriptions
  - Legal documents
- Documents
  - Articles, books and novels E-mails
  - web pages, blogs
  - Tags, comments
- Collection of documents
  - Messages (e-mail, blogs, tags, comments)
  - Social networks (personal profiles)
  - Academic collaborations (publications)

Text units hierarchy



Typical way of doing text visualization

- Typical visualization scenarios:
  - Visualization of document collections
  - Visualization of search results
  - Visualization of document timeline

Text/Document attributes

- corpus is collection of documents
  - objects within - words, sentences, paragraphs, documents, or collection of documents
  - plural: corpora

- Metadata
  - author
  - date of creation
  - date of modification
  - comments
  - size

Computing statistics about documents

- number of words or paragraphs
- word distribution
- word frequency
- relationships between paragraphs or documents

Text features are complicated

- Be aware!! text understanding can be hard:
  - *Toilet out of order. Please use floor below.*
  - *"One morning I shot an elephant in my pajamas. How he got in my pajamas, I don't know."*
  - *"Did you ever hear the story about the blind carpenter who picked up his hammer and saw?"*

http://en.wikipedia.org/wiki/List_of_linguistic_example_sentences

Challenges

- Text is nominal
  - High dimensionality
- The step "Raw data → Data Table" mapping is important/central
- Unstructured text does NOT have any explicit meta-data
  - Meta-data is sometimes extracted from raw text
    - Google News extracts dates

- Contrast to structured text of an online library with explicit meta-data such as
    - Author name
    - Year of publication
    - Title
    - ISBN number
    - Library of Congress umber
    - Publisher name
    - etc.

Levels of text representations

- Lexical - transforming a string of characters into a sequence of atomic entities, tokens
    - Character (character n-grams and sequences)
    - Words (stop-words, stemming, lemmatization)
    - Phrases (word n-grams, proximity features)

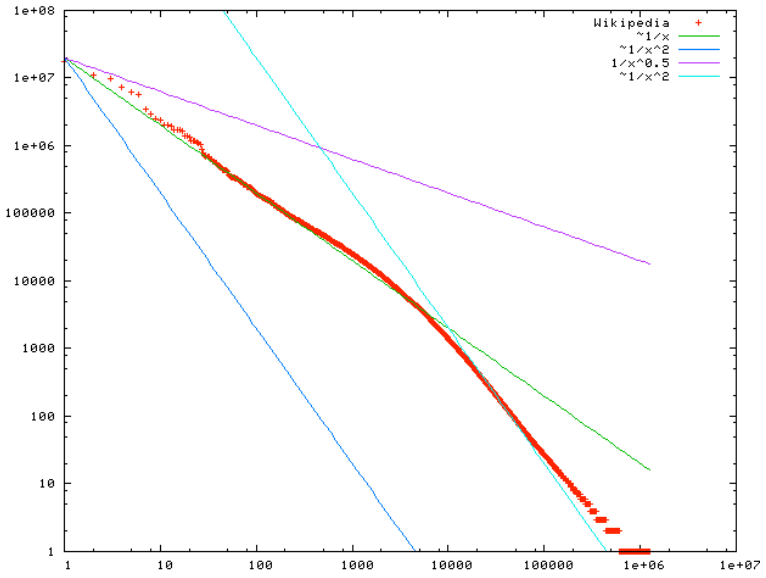
- Syntactic - identifying and tagging each token's function, named entity recognition (NER)
    - Vector-space model
    - Named entities recognition
    - Language models

- Semantic - extraction of meaning and relationships between pieces of knowledge derived from the structures identified in the syntactical level
    - Collaborative tagging / Web2.0
    - Templates / Frames
    - Ontologies

Word level

- The most common representation of text used for many techniques

- Word frequencies in texts have power distribution:
  - small number of very frequent words
  - big number of low frequency words
  - small number of words describe most of the key concepts in small documents



Zipf's law

Phrase level

- Instead of having just single words we can deal with phrases
- The main effect of using phrases is to more precisely identify sense
- Google Ngrams (ngram is a contiguous sequence of n items):
  - https://books.google.com/ngrams

Named entities recognition

verb
Stop!

noun	verb
John	works.

noun	verb	verb
John	is	working.

pronoun	verb	noun
She	loves	animals.

noun	verb	adjective	noun
Animals	like	kind	people.

noun	verb	noun	adverb
Tara	speaks	English	well.

noun	verb	adjective	noun
Tara	speaks	good	English.

pronoun	verb	preposition	adjective	noun	adverb
She	ran	to	the	station	quickly.

pron.	verb	adj.	noun	conjunction	pron.	verb	pron.
She	likes	big	snakes	but	I	hate	them.

Here is a sentence that contains every part of speech:

interjection	pron.	conj.	adj.	noun	verb	prep.	noun	adverb
Well,	she	and	young	John	walk	to	school	slowly.

Source:
Date: Nov 16, 2004

Color Legend:
person, place, organization, date, money, time

Alderwood to probe voting machines

Story by: Ellie Olmsen

Republicans in Alderwood joined Democrats yesterday in criticizing the performance of the city's costly new high-tech voting system, saying that it may have disenfranchised voters in the Nov. 4 election.

The Republican commission scolded the city board of elections for minimizing problems with the touch-screen machines that the city purchased this year for $1.5 million and asked Mayor Rex Luthor to investigate what went wrong before the machines are pressed into service again.

Alderwood's touch-screen voting machines, which resemble laptop computers without keyboards, were supposed to simplify voting and tabulating results. But in a debut that mirrored many of the problems experienced last year in areas across the country, some voters found the machines confusing, and the reporting of vote tallies was delayed almost a day.

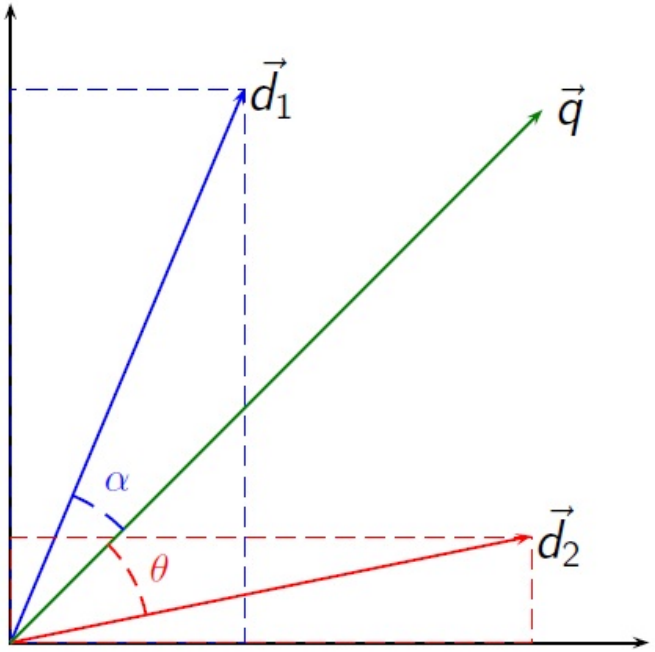
Luthor responded that he would try to address the board's concerns. He said he has called for a public meeting of the three-member board of elections to go over the requests at 5 p.m. today.

"I pledge that I will answer every question as soon as I possibly can in the proper fashion," he said.

http://www.englishclub.com/grammar/parts-of-speech_2.htm

Vector Space Model

- How does one compare the similarity of two documents?
- Vector-space model
  - make list of each unique word in document
    - throw out common words (a, an, the, ...) - *stop words*
    - make different forms the same (bake, bakes, baked) – *stemming*
  - store count of how many times each word appeared!
  - alphabetize, make into a vector
- Applications:
  - Classification
  - Clustering
  - Visualization



https://en.wikipedia.org/wiki/Vector_space_model

Not all words are equally useful

- Often apply tf-idf
  - **term frequency** − how frequent if the term in a document
  - **inverse document frequency** − the more frequent the word in other documents the less important it is

$$tfidf(w) = tf.\log(\frac{N}{df(w)})$$

  - Tf(w) – term frequency
    - number of word occurrences in a document
  - Df(w) – document frequency
    - number of documents containing the word
  - N – number of all documents
  - TfIdf(w) – relative importance of the word in the document

Term Vectors with tf-idf Weights

term vector - each dimension represents the weight of a given word in that document

id	men	entered	bank	charlotte	missiles	masks	aryan	guns	witnesses	reported	silver	suv	august
seg1.txt	0.239441	0	0.153457	0.195243	0	0.237029	0	0.195243	0.237029	0.140004	0.195243	0.237029	0
seg13.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg14.txt	0	0.192197	0	0	0	0	0	0	0	0	0	0	0.172681
seg15.txt	0	0	0	0	0	0	0	0	0	0	0	0	0.149652
seg16.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg17.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg18.txt	0	0.158432	0	0	0	0	0	0	0	0	0	0	0
seg19.txt	0	0	0	0.197255	0	0	0	0	0	0.141447	0	0	0.155038
seg2.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg20.txt	0	0.234323	0	0	0	0	0	0	0	0	0	0	0
seg21.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg22.txt	0	0	0	0	0.139629	0	0.127389	0	0	0	0	0	0
seg23.txt	0	0	0	0	0	0	0	0	0	0.180656	0	0	0
seg24.txt	0	0	0	0	0	0	0.117966	0	0	0.117966	0	0	0
seg25.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg26.txt	0	0	0	0	0	0	0	0	0	0	0	0	0
seg27.txt	0	0	0.235418	0	0	0	0.214781	0	0	0	0	0	0
seg28.txt	0	0	0	0	0.151753	0	0	0	0	0	0	0	0
seg29.txt	0	0	0	0	0	0	0.129852	0	0	0	0	0	0.142329
seg3.txt	0	0	0	0	0.18432	0	0	0	0	0	0	0	0
seg30.txt	0.078262	0	0	0	0	0	0	0	0	0	0	0	0
seg31.txt	0	0	0.213409	0	0	0	0.194701	0	0	0	0	0	0
seg32.txt	0	0	0	0	0	0	0	0	0	0	0	0	0

Similarity between document vectors

- Each document is represented as a vector of weights
- Cosine similarity (dot product) is the most widely used similarity measure between two document vectors
  - calculates cosine of the angle between document vectors
  - efficient to calculate (sum of products of intersecting words)
  - similarity value between 0 (different) and 1 (the same)

$$Sim(D_1, D_2) = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_j x_j^2} \sqrt{\sum_k x_k^2}}$$

Typical steps of processing to derive text features

- Cleaning
  - Removing punctuations
- Tokenization - sentence splitting
- Change to lower case
- Stop word removal
  - most frequent words in a language ("a", "the", "and")
- Stemming - demo porter stemmer

Stop Words

- Stop-words are words that from non-linguistic view do not carry information
  - they have mainly functional role
  - usually we remove them to help the methods to perform better

- Stop words are language dependent – examples:
  - **English**: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ...
  - **Dutch**: de, en, van, ik, te, dat, die, in, een, hij, het, niet, zijn, is, was, op, aan, met, als, voor, ...
  - **Slovenian**: A, AH, AHA, ALI, AMPAK, BAJE, BODISI, BOJDA, BRŽKONE, BRŽČAS, BREZ, CELO, DA, DO, ...

Stemming

- Different forms of the same word are usually problematic for text data analysis, because they have different spelling and similar meaning (e.g. learns, learned, learning,…)

- Stemming is a process of transforming a word into its stem (normalized form)
  - Learns -> learn
  - learning -> learn
  - learning -> learn

Stemming

- For English is mostly used Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer
  - Removing the commoner morphological and inflexional endings from words in English.

- Example cascade rules used in English Porter stemmer
  - ATIONAL -> ATE relational -> relate
  - TIONAL -> TION conditional -> condition
  - IZER -> IZE digitizer -> digitize

Word Count

Tokenization

D1 = 'Julie loves me more than Linda loves me'

D2 = 'Jane likes me more than Julie loves me'

Stop word removal

D1_tokens =
['Julie', 'loves', 'me', 'more', 'than', 'Linda', 'loves', 'me']
D2_tokens =
['Jane', 'likes', 'me', 'more', 'than', 'Julie', 'loves', 'me']

Stemming

D1_tokens_no_sw = ['Juli', 'love', 'Linda', 'love']
D2_tokens_no_sw = ['Jane', 'like', 'Julie', 'love']

Word count

Words	Frequency
Jane	1
like	1
Juli	1
love	1

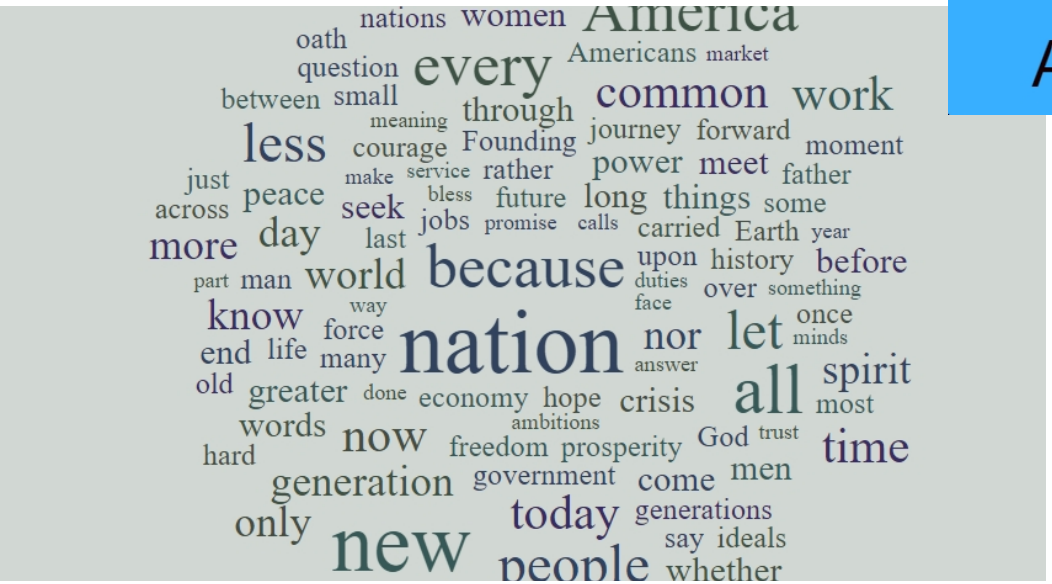
Words	Frequency
loves	2
Juli	1
Linda	1

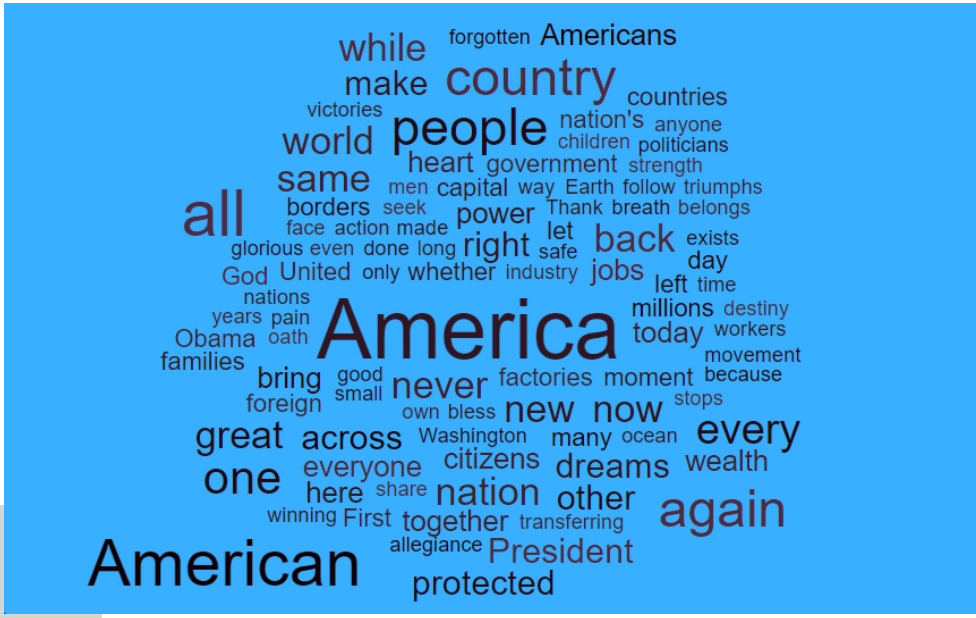
Demo
Moved to next lecture

Why visualize text?

- **Understanding** – get the "gist" of a document
- **Grouping** – cluster for overview or classification
- **Compare** – compare document collections, or inspect evolution of collection over time
- **Correlate** – compare patterns in text to those in other data, e.g., correlate with social network

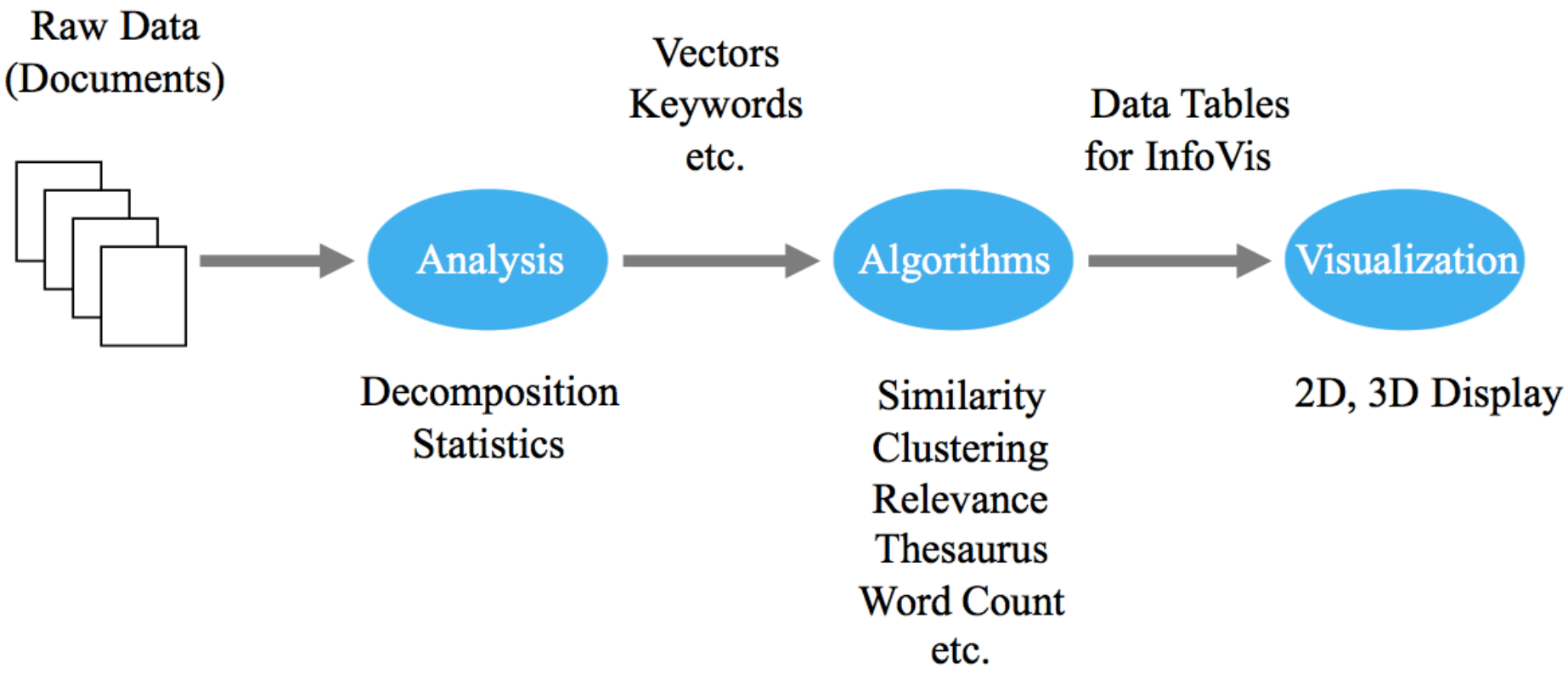
Obama Speech versus Trump speech





https://www.usatoday.com/story/news/politics/onpolitics/2016/01/12/obama-state-of-the-union-word-clouds/78712780/

Process for text/document in InfoVis



Single Document Visualization

Tag/Word clouds

The font size and darkness are proportional to the frequency of the word in the document

Tag/Word clouds

- Have proven to be very popular on web
- Idea is to show word/concept importance through visual means
  - tags: user-specified metadata (descriptors) about something
  - sometimes generalized to just reflect word frequencies

Wordle

The size of the text corresponds to the frequency of the word in the document

Wordle

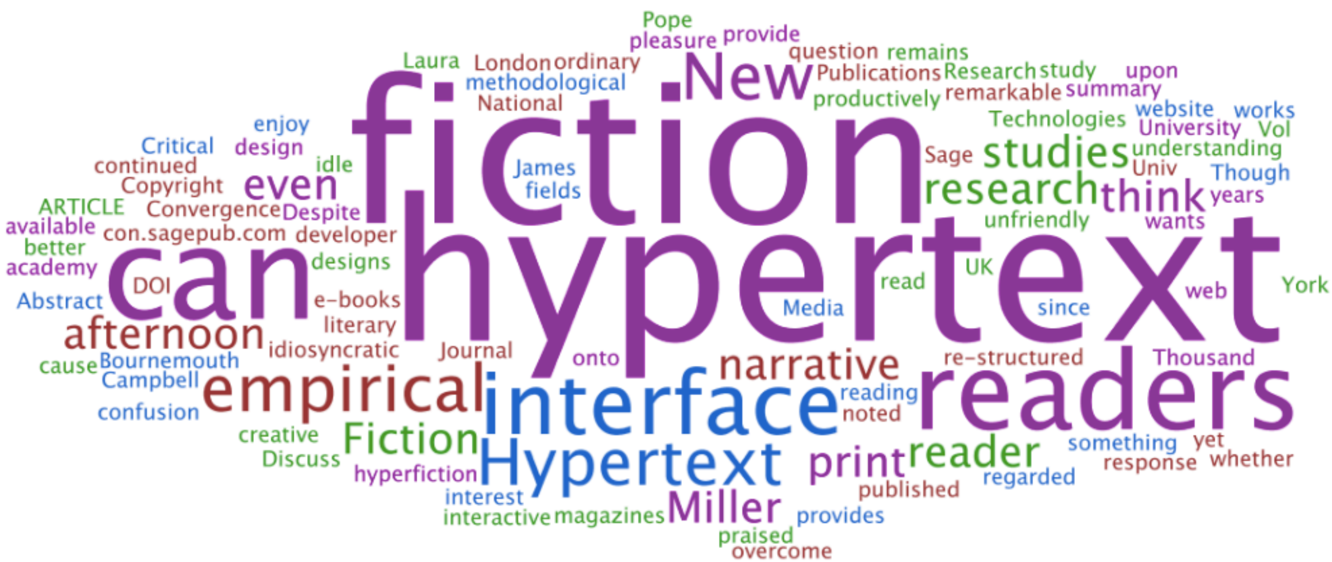
- Tightly packed words, sometimes vertical or diagonal
- Word size is linearly correlated with frequency (typically square root in cloud)
- Multiple color palettes
- User gets some control

ICA06

Word cloud design parameters

- alphabetical order
- same orientation or different orientation
- font
- color
- Foreground or background

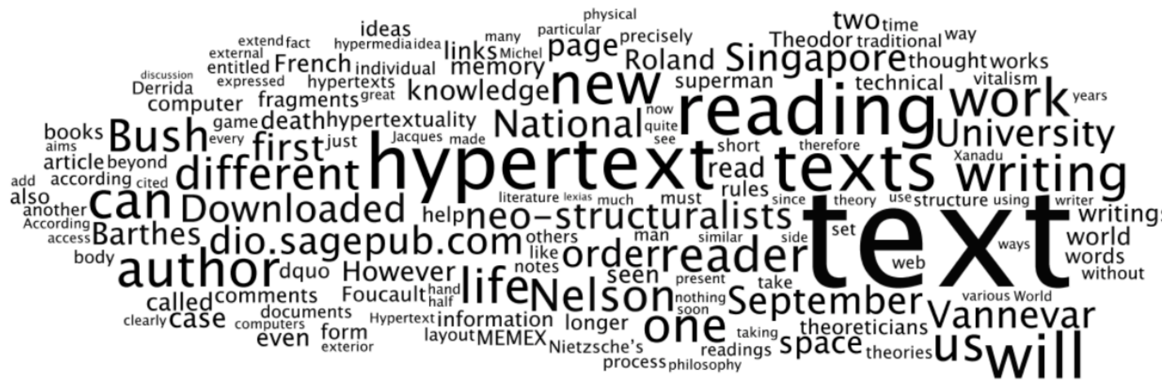
Alpha order

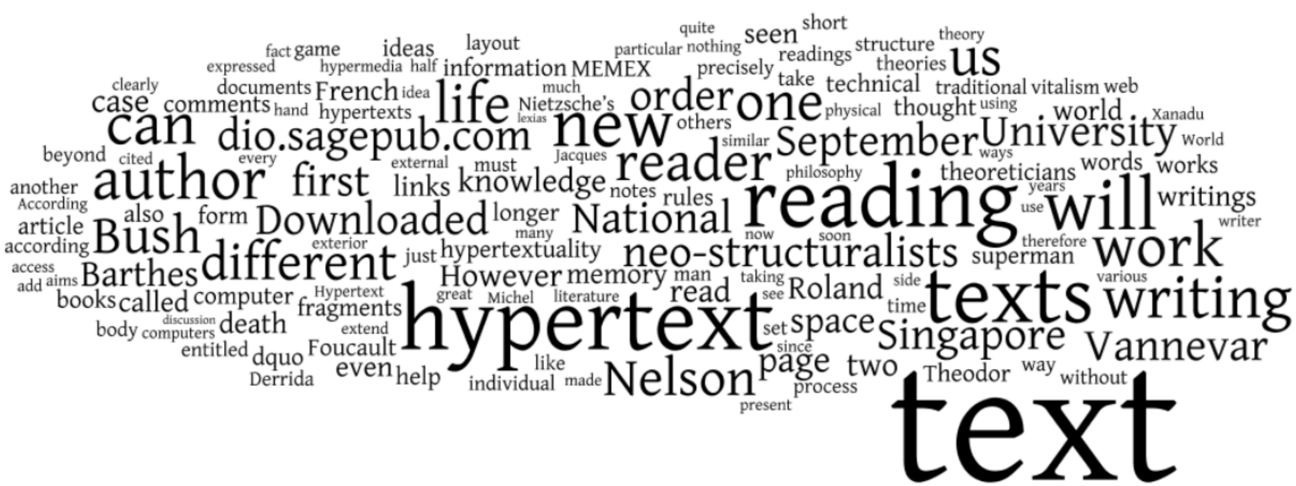

Pope
pleasure provide
Laura London ordinary
question remains
methodological
National
New
Publications Research study upon
remarkable summary
productively
fiction
enjoy
Critical design
continued idle
Copyright even
ARTICLE Convergence Despite
available
con.sagepub.com developer
better
academy
James
fields
Technologies website works
University Vol
Sage studies understanding
Univ Though
research think years
unfriendly wants
can
DOI
Abstract
designs
e-books
literary
afternoon
idiosyncratic Journal
hypertext
read UK
Media since web York
onto
narrative re-structured Thousand
cause Bournemouth
Campbell
empirical
interface
readers
reading noted
confusion
creative
Discuss
Fiction
hyperfiction
Hypertext print reader
something yet whether
response
regarded
published
interest
interactive magazines Miller provides
praised
overcome


Journal
pleasure Discuss whether summary e-books available National London interest
Though
re-structured ordinary interactive Technologies website Univ
Bournemouth University
readerscan
Despite upon
remains Critical years
think Copyright
continued Thousand overcome works Pope
noted academy better
Miller idiosyncratic hyperfiction developer print provides
unfriendly
read literary
empirical
Fiction since response studies
designs fields
fiction
study Hypertext
UK productively Laura
onto
Sage Vol
York Convergence
interface Media idle New
yet research
cause
magazines
afternoon even wants methodological
web con.sagepub.com something understanding
DOI creative ARTICLE regarded design enjoy
reading narrative
reader Publications
Campbell
hypertext
question remarkable
confusion Abstract Research
James
praised provide published

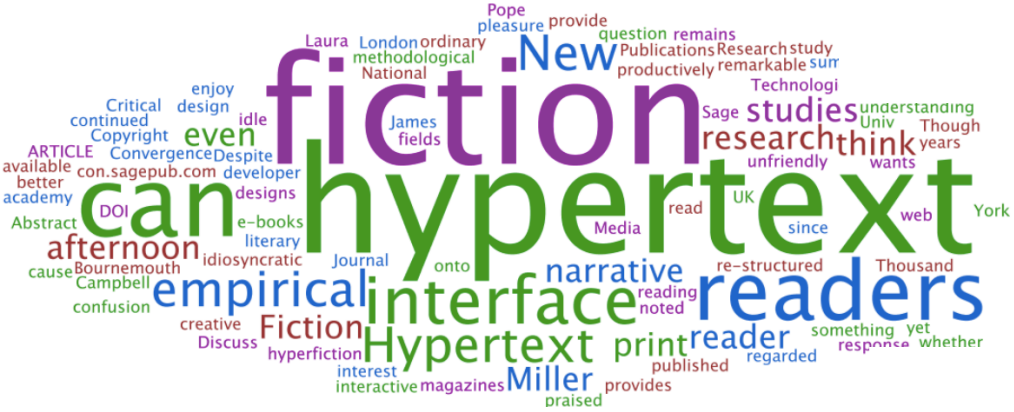
Same/different orientation

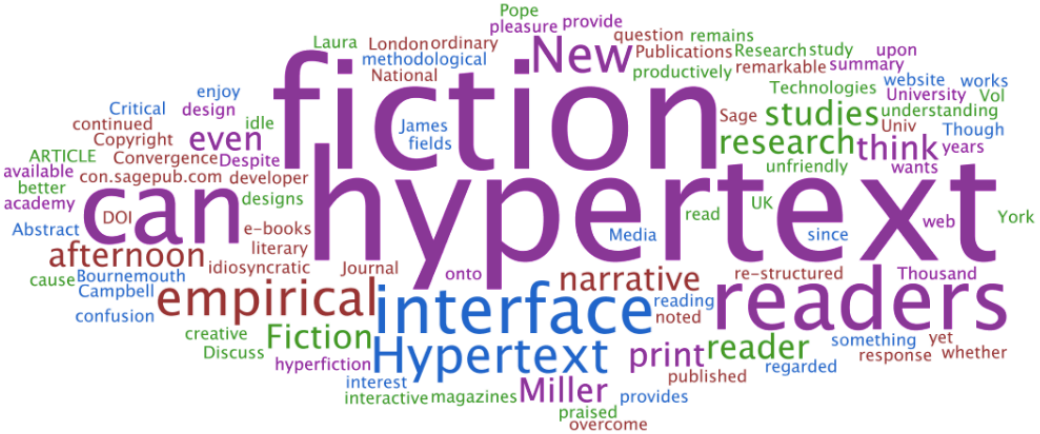
Font

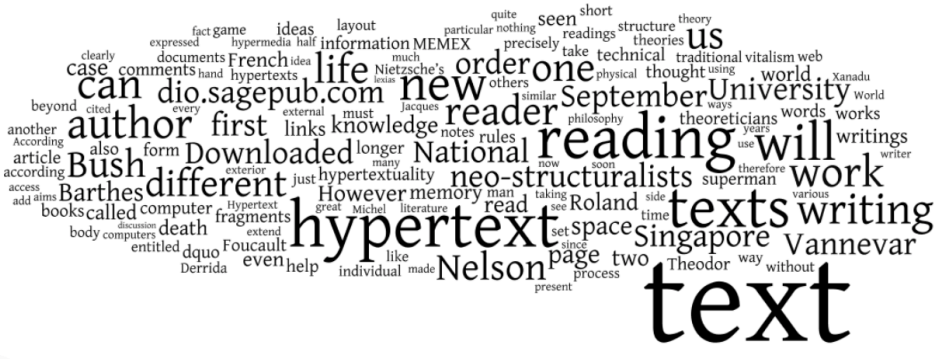

text
hypertext
reading
texts
writing
work
new
Singapore
University
National
different
Downloaded
neo-structuralists
dio.sagepub.com
reader
author
Bush
can
first
life
Nelson
September
Vannevar
one
will
us
space
order
Barthes
However
knowledge
Roland
page
memory
links
ideas
French
death
hypertextuality
fragments
computer
Theodor
thought
works
technical
vitalism
superman
read
rules
writings
world
words
without
theoreticians
readings
information
MEMEX
Nietzsche's
Foucault
comments
documents
called
case
even
form
process
philosophy
layout
longer
seen
two
time
traditional
way


text
hypertext
reading
texts
writing
will
work
new
author
can
different
Bush
Downloaded
neo-structuralists
dio.sagepub.com
National
Nelson
reader
September
University
Singapore
Vannevar
one
order
life
first
us
knowledge
Barthes
However
memory
hypertextuality
Roland
space
page
two
Theodor
superman
theoreticians
writings
world
works
words
French
hypertexts
information
MEMEX
comments
case
links
ideas
layout
short
seen
structure
theories
technical
traditional
vitalism
web
thought
Nietzsche's
death
fragments
computer
called
books
article
Foucault
dquo
even
help
individual
Derrida
entitled
process
without
way

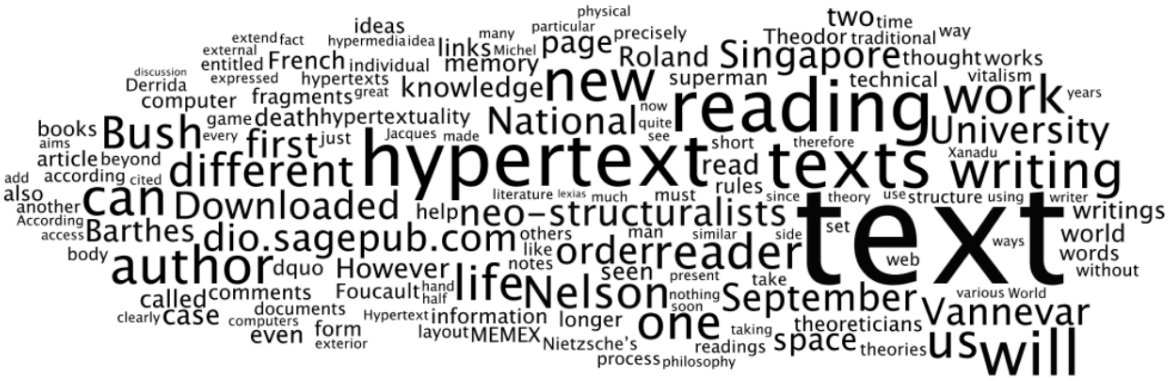
Color palette

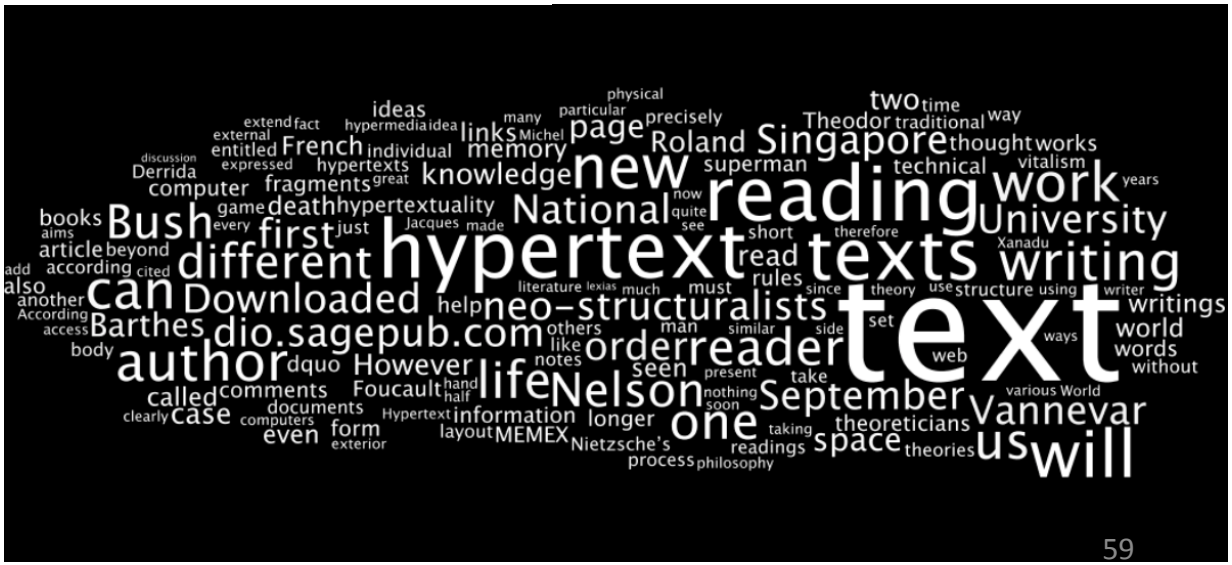






Foreground/Background





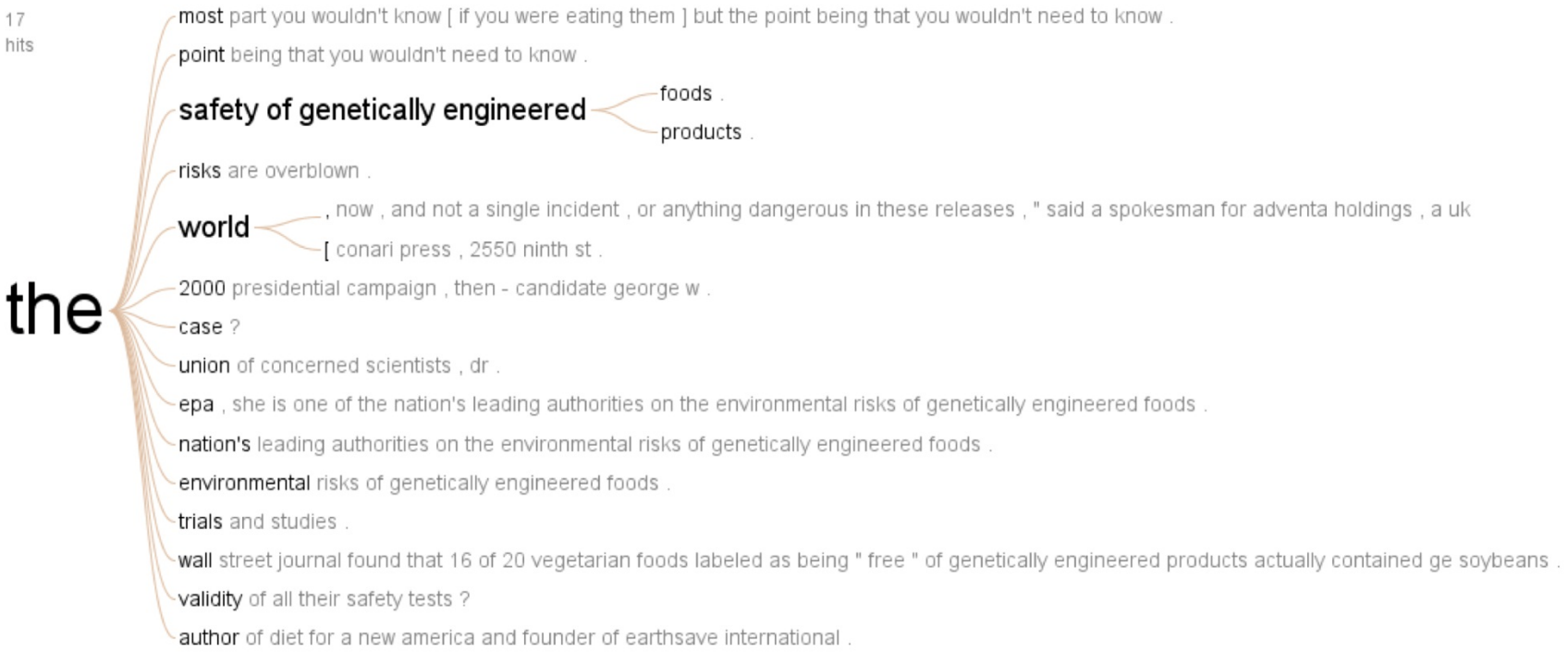
Problems with tag clouds

- Actually not a great visualization. Why?
- hard to find a particular word
  - long words get increased visual emphasis
  - font sizes are hard to compare
- alphabetical ordering not ideal for many tasks
- Studies have even shown they underperform

Remember

- Word Clouds and Wordles are really more overview-style visualizations
  - don't really support queries, searches, drill-down

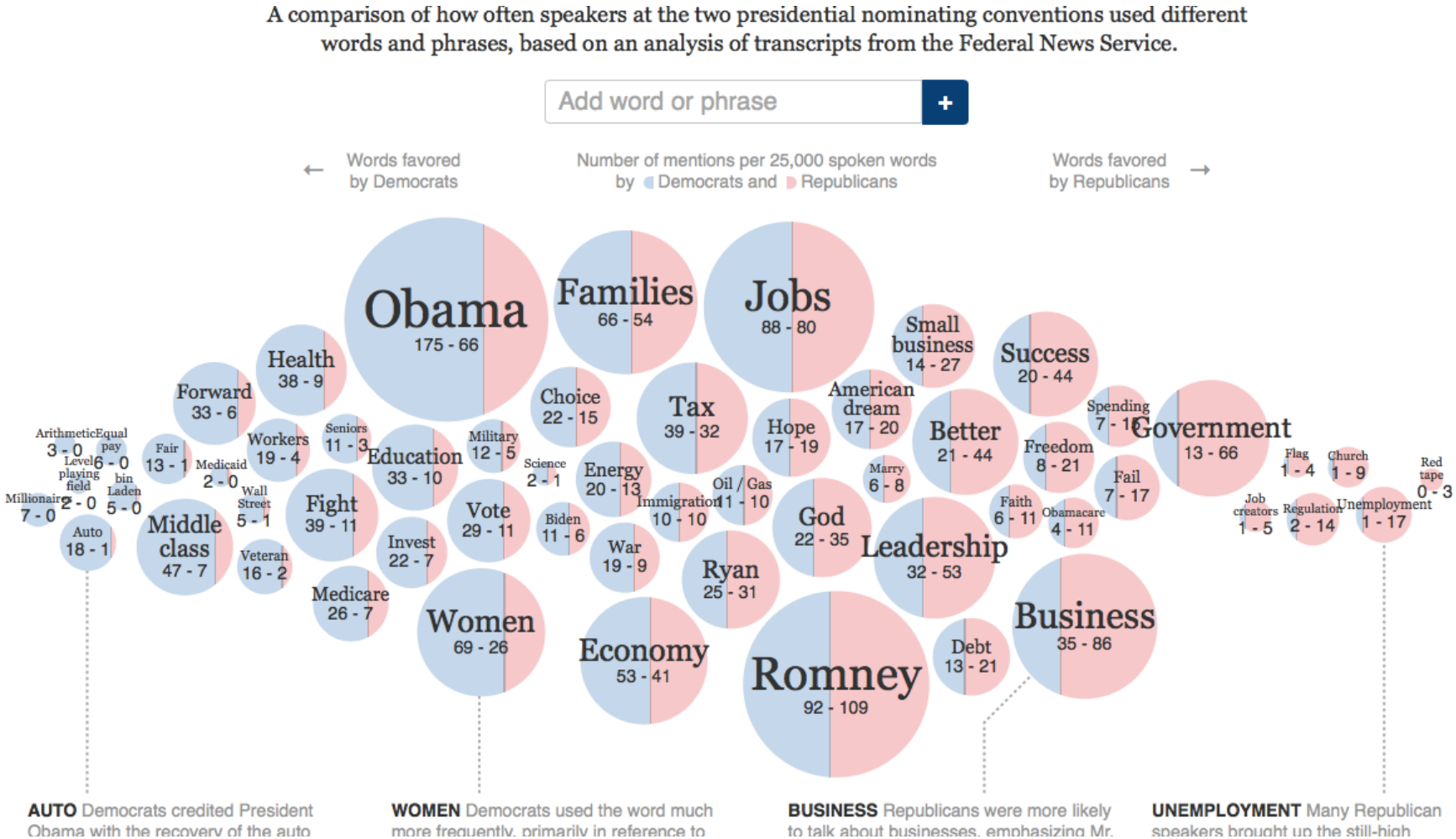
WordTree



ManyEyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)

The branches of the tree represent the various contexts following a root word or phrase in the document.

NY Times example
Word Counts at 2012 Conventions



http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html

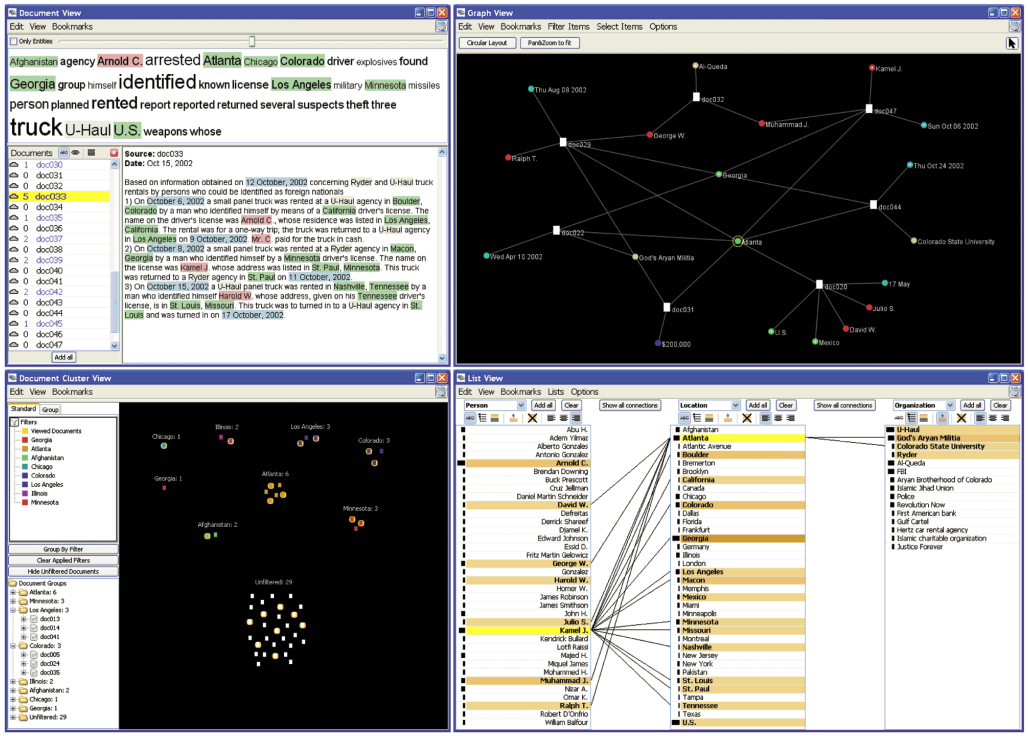
Documents Collections

Visualizing document sets

- Still do words, phrases, sentences
- Add
  - more context of documents
  - document analysis metrics
  - document meta-data
  - document entities
  - connections between documents ! documents concepts and themes

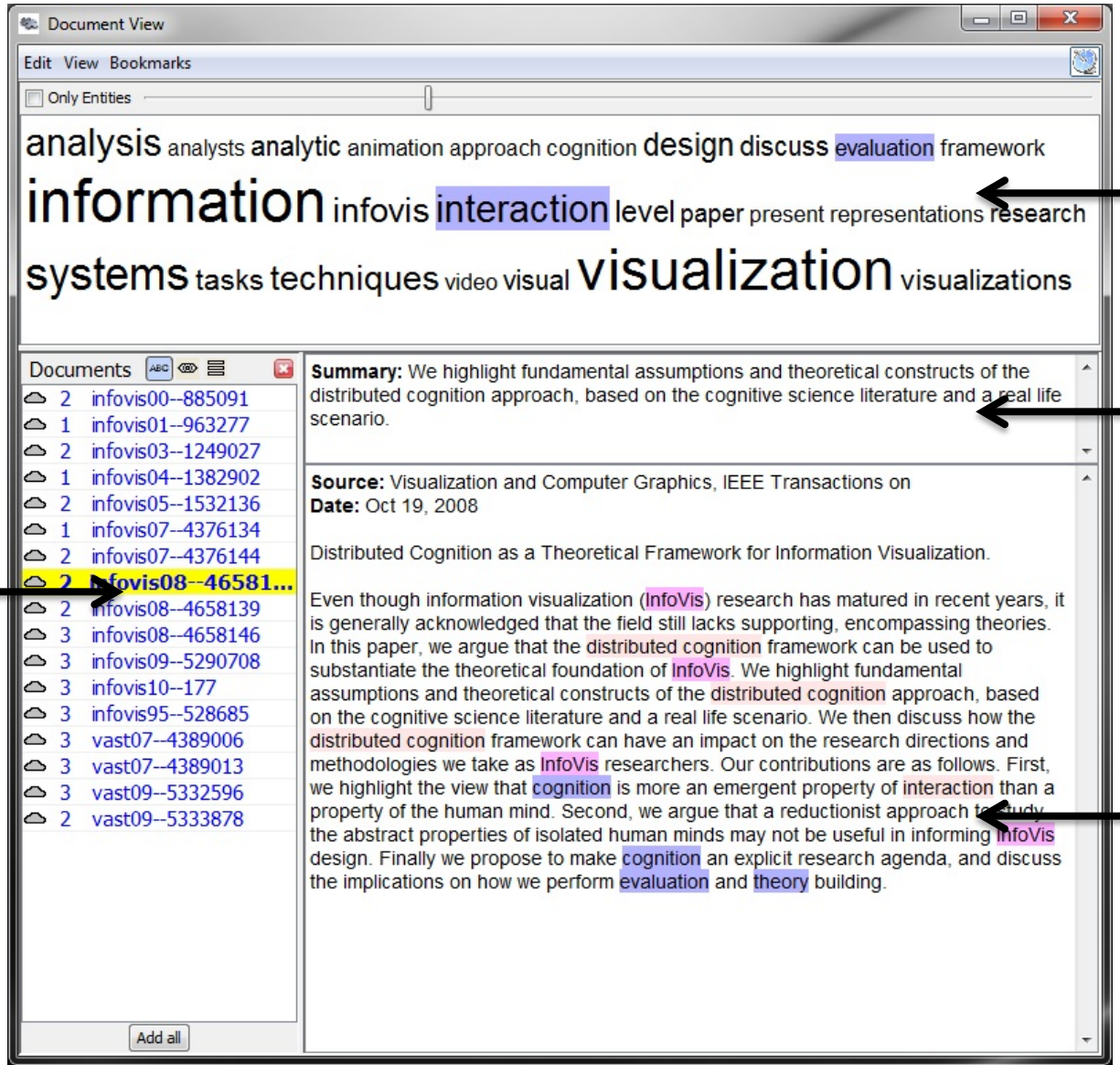
Jigsaw

- Variety of visualizations ranging from word-specific, to entity connections, to document clusters
- Primary focus is on entity-document and entity-entity connection
- Search capability coupled with interactive exploration



https://www.cc.gatech.edu/gvu/ii/jigsaw/

Jigsaw
Document view



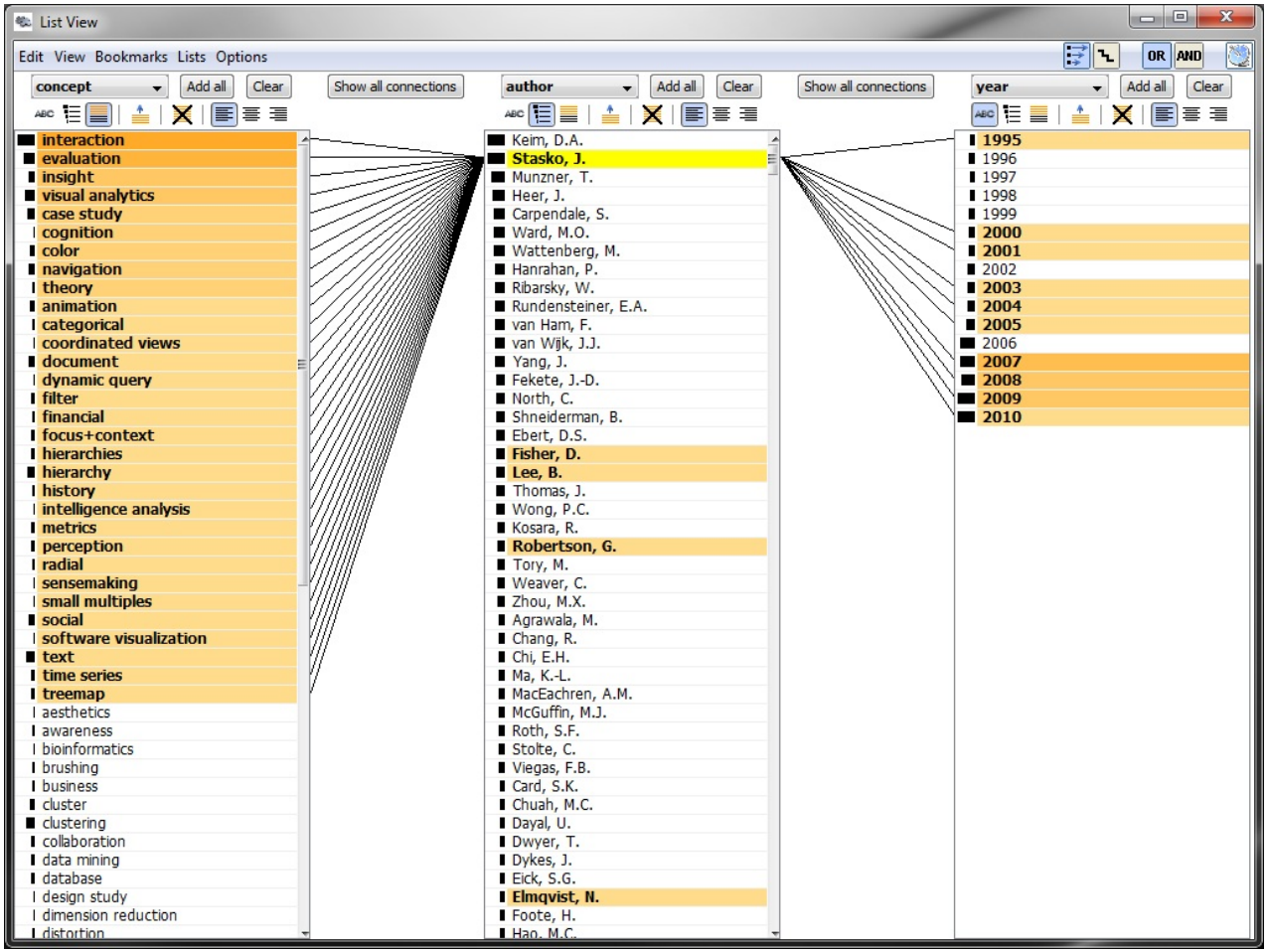
Wordcloud overview

Document summary

Doc list

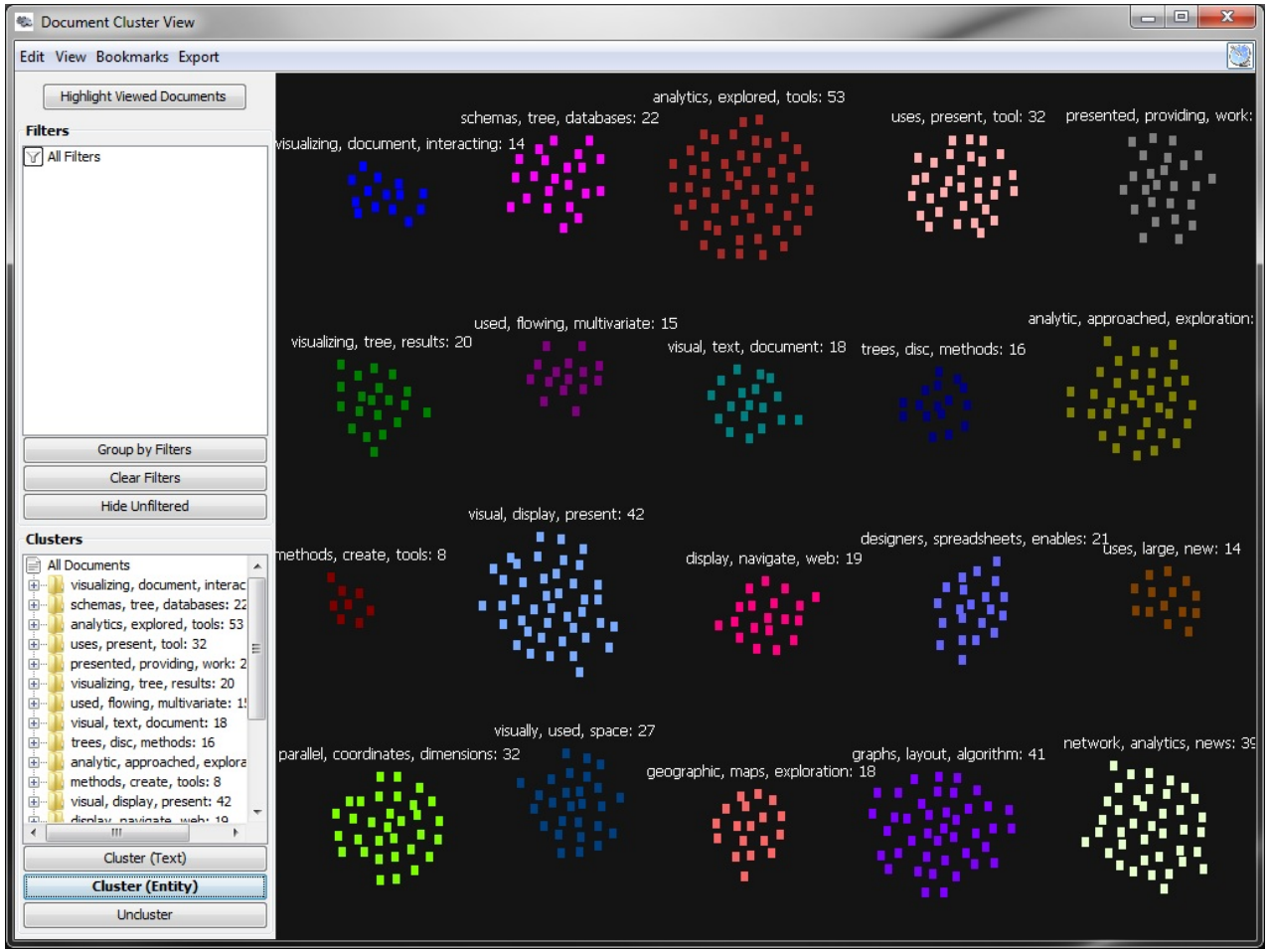
Selected document's text with identified entities

Jigsaw
List view



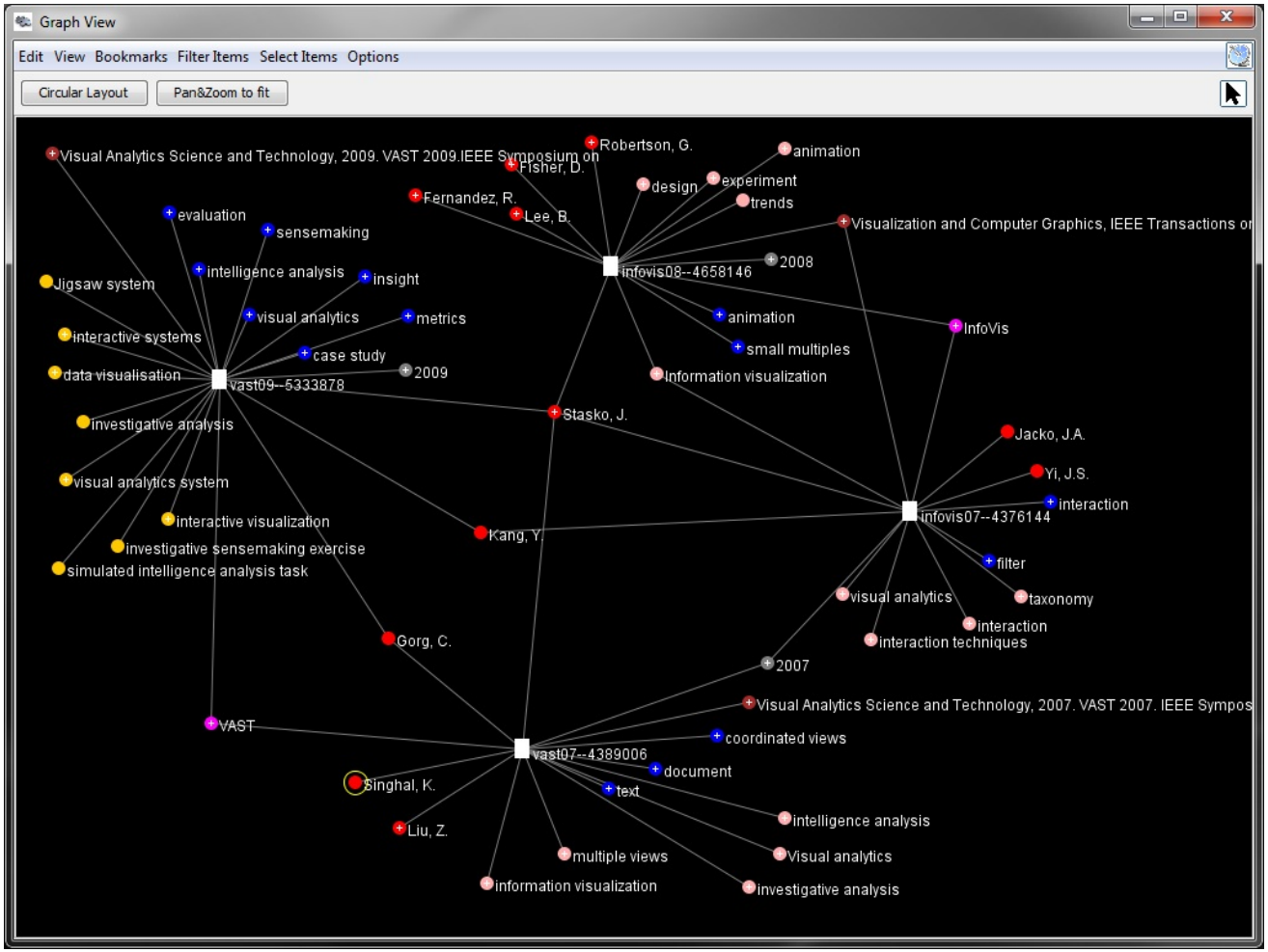
Entities listed by type

Jigsaw
Document cluster view



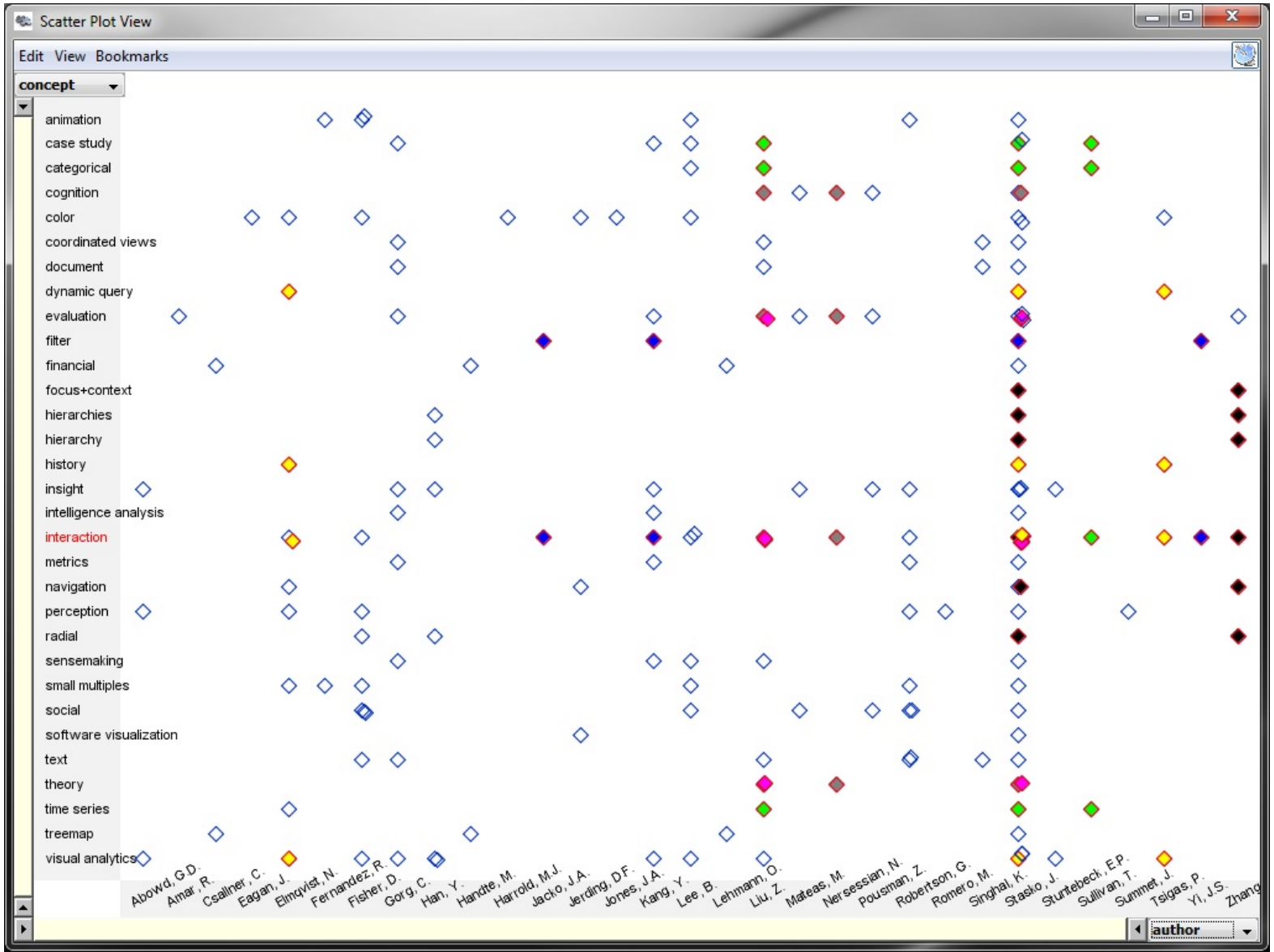
https://www.cc.gatech.edu/gvu/ii/jigsaw/

Jigsaw
Network view

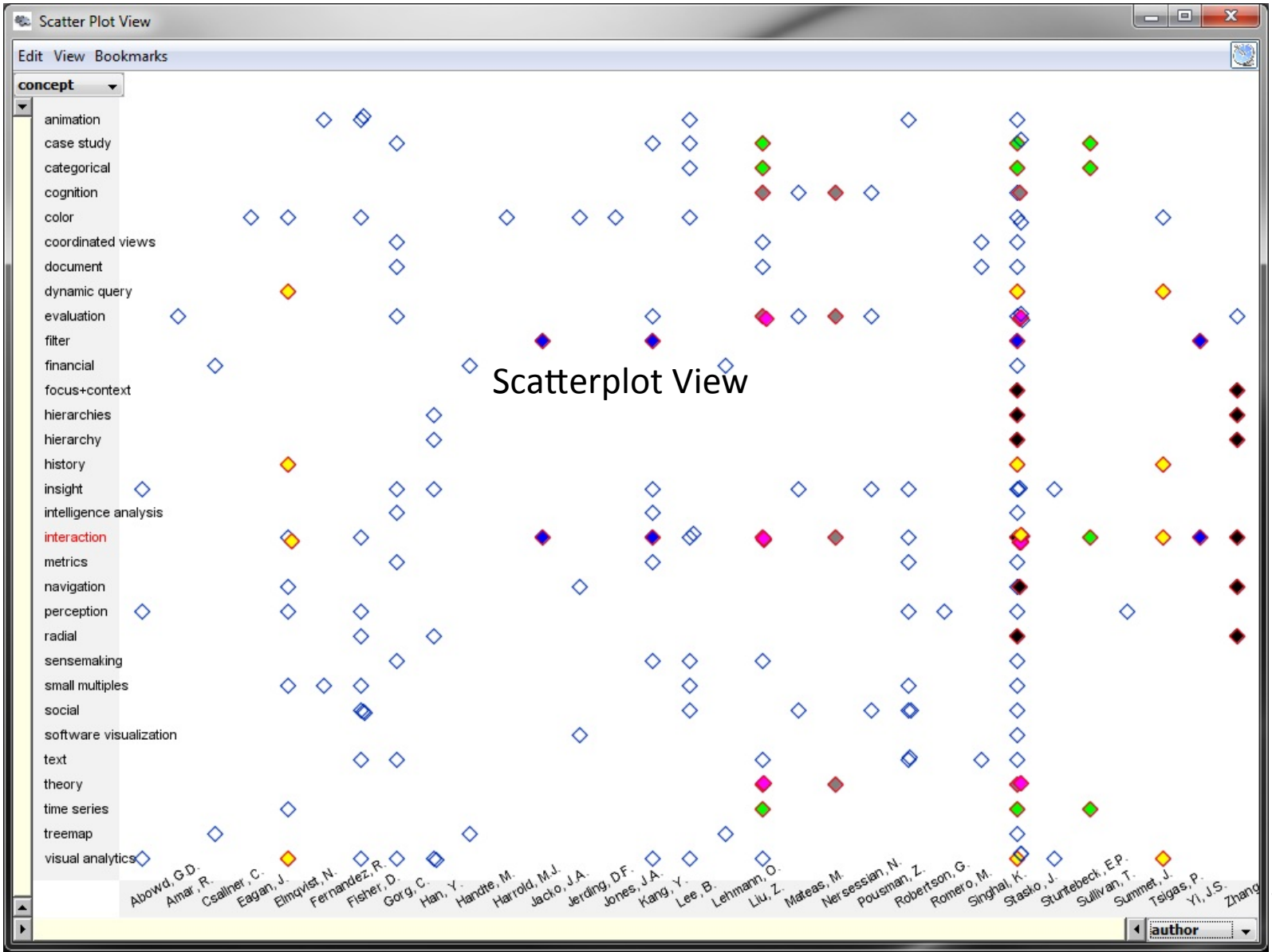


https://www.cc.gatech.edu/gvu/ii/jigsaw/

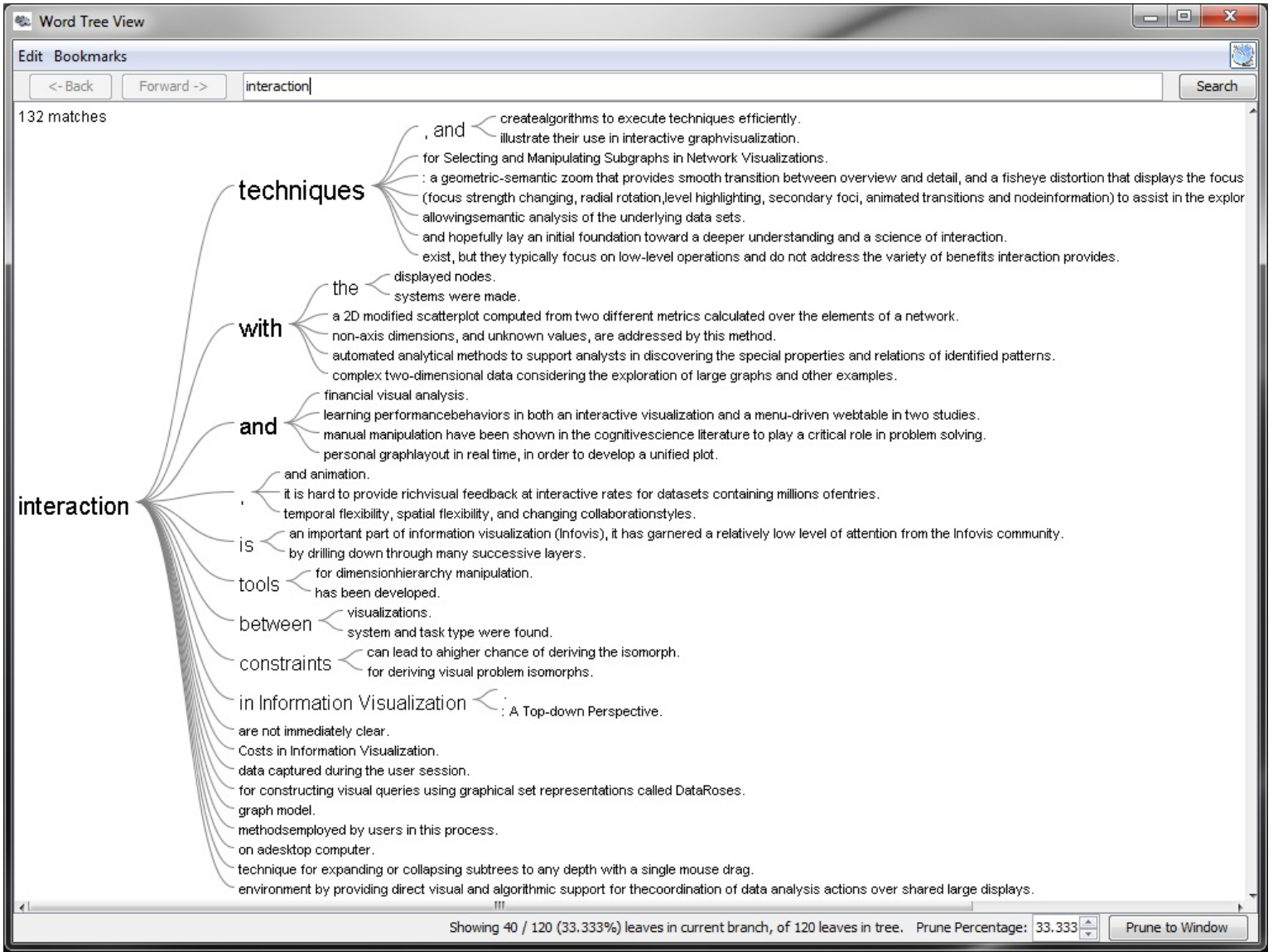
Document Grid View



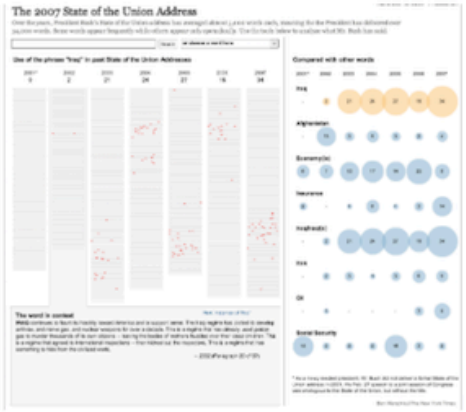
Scatterplot View



WordTree View

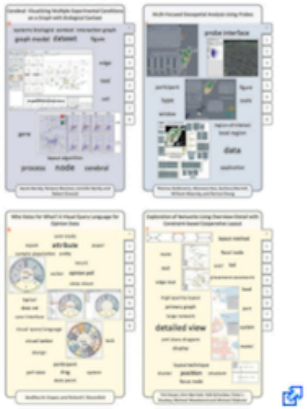


New York Times



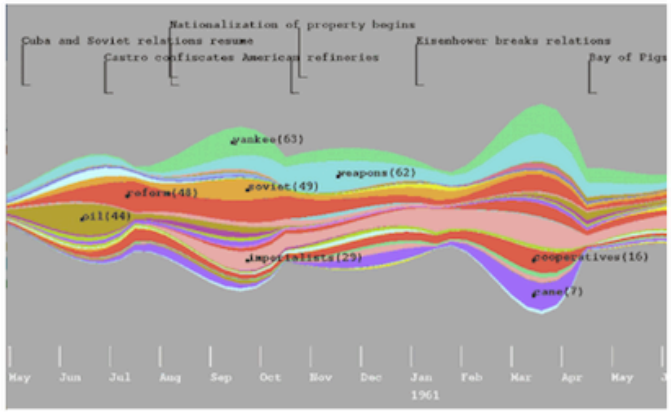
Interactive visualization of President Bush's State of the Union Address in 2007

Document Cards



Document Cards: A Top Trumps Visualization for Documents (demo video)

ThemeRiver

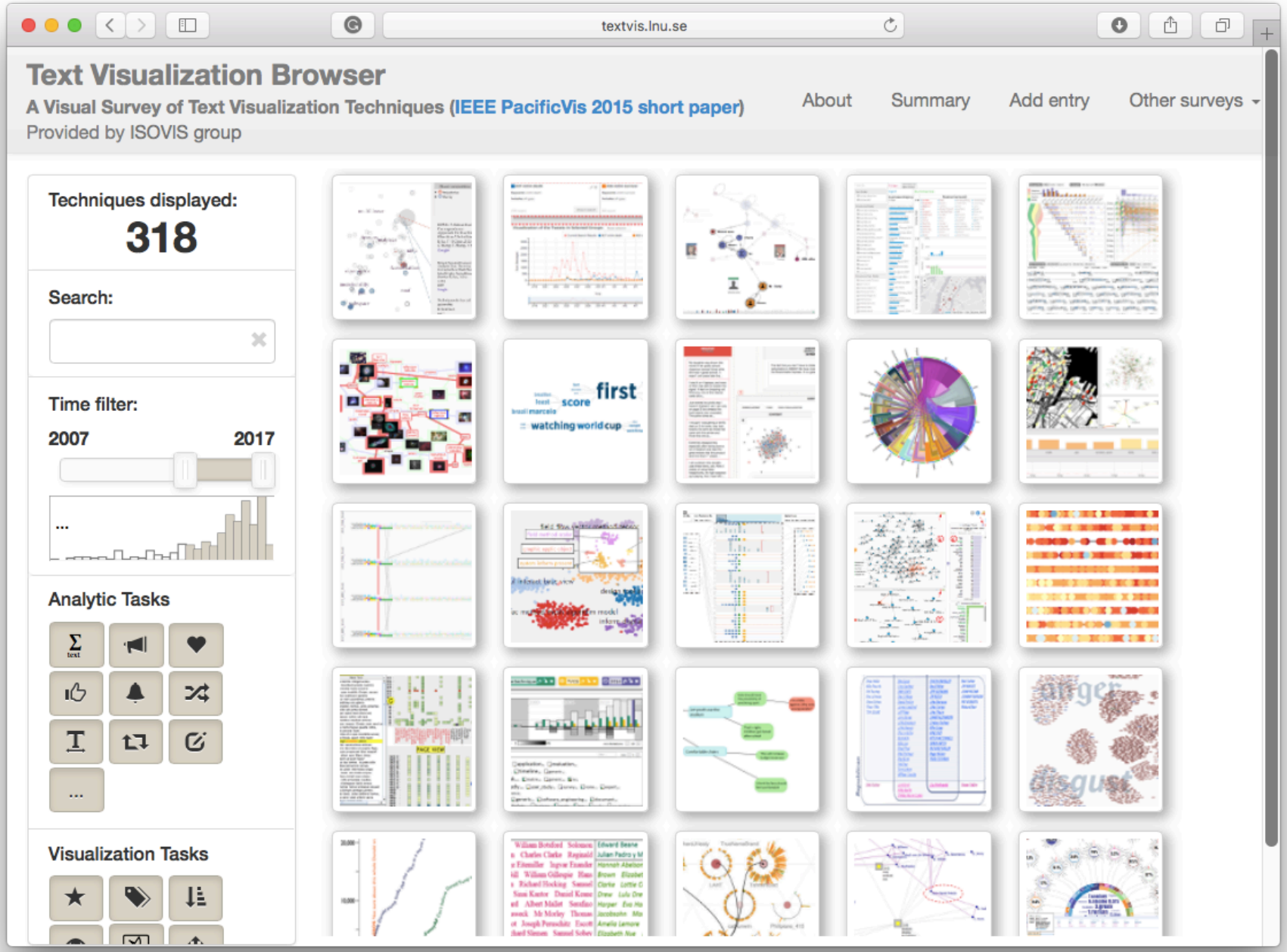


ThemeRiver: Visualizing Thematic Changes in Large Document Collections (demo video)

PaperLens

Understanding Research Trends in Conferences using PaperLens

http://www.cs.odu.edu/~mweigle/Research/Archive-Vis#toc4



http://textvis.lnu.se/

Next

- Visualizing social networks