Text and Document Visualization 1

CS 4460 – Intro. to Information Visualization
November 13, 2017
John Stasko

Learning Objectives

- Explain key challenges in visualizing a large document or body of text
- Identify and explain different techniques for representing words and concepts in a document
  - Word cloud, Wordle, Parallel tag cloud, SeeSoft, PhraseNet
- Understand the positives and limitations of word clouds and Wordles
- Describe SeeSoft-style miniature visual representations

Text is Everywhere

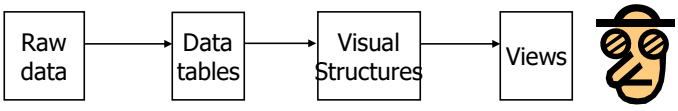
- We use documents as primary information artifact in our lives
- Our access to documents has grown tremendously in recent years due to networking infrastructure
  - WWW
  - Digital libraries
  - …

Big Question

- What can information visualization provide to help users in understanding and gathering information from text and document collections?

Challenge

- What's the big challenge here?
- Text is nominal data
  - Does not seem to map to geometric/graphical presentation as easily as ordinal and quantitative data
- The "Raw data --> Data Table" mapping now becomes more important

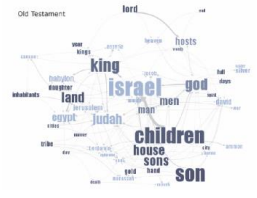


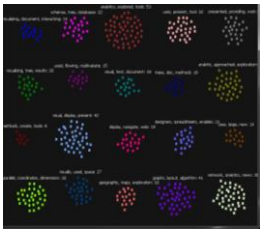
This Week's Agenda

Visualizing text
Showing words, combinations, and context

→

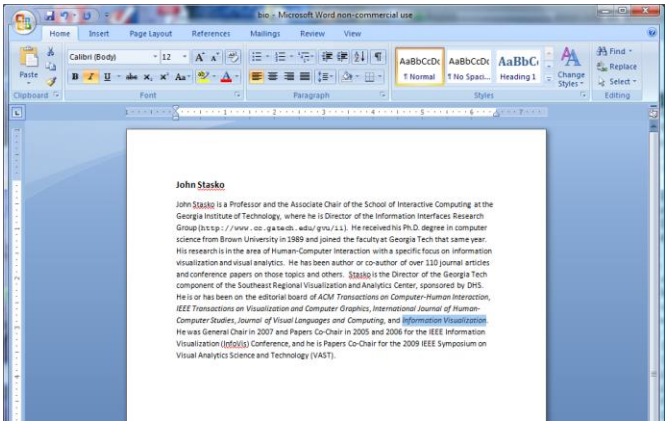
Visualizing document sets
Words & sentences
Analysis metrics
Concepts & themes





What's the simplest text visualization you know?

One Text Visualization



Uses:
Layout
Font
Style
Color
…

Design Challenge

- How would you visualize one of the presidential debates?
- Brainstorm for a few minutes

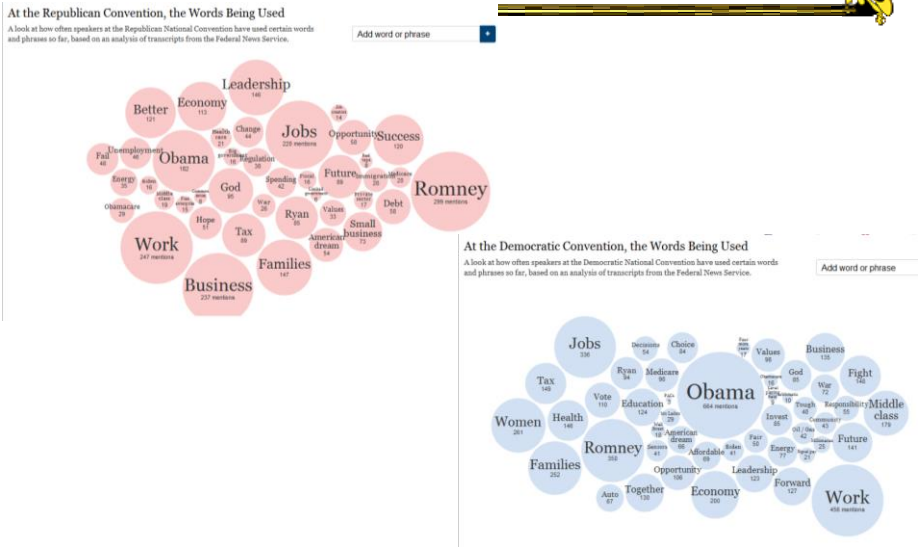
What was implicit in this exercise?

Tasks

- What kinds of questions or tasks would someone want to do with such a visualization?

http://www.nytimes.com/interactive/2012/08/28/us/politics/convention-word-counts.html

Word Counts



More Word Counting

http://www.wordcount.org

Tag/Word Clouds

- Currently very "hot" in research community
- 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

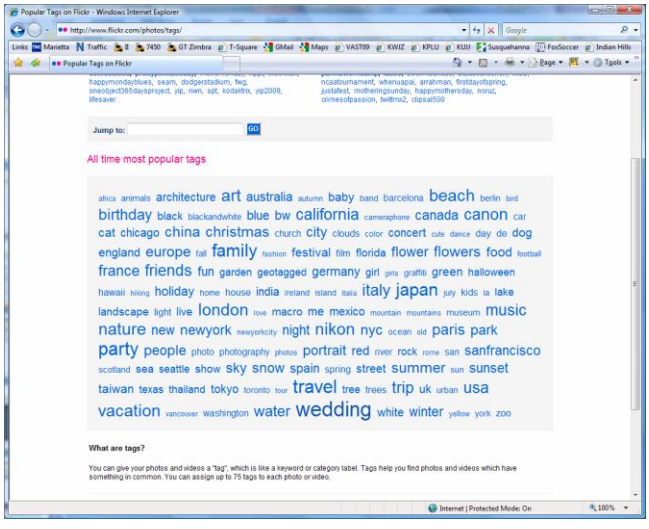
History

- 90-year old Soviet Constructivism
- Milgram's '76 experiment to have people label landmarks in Paris

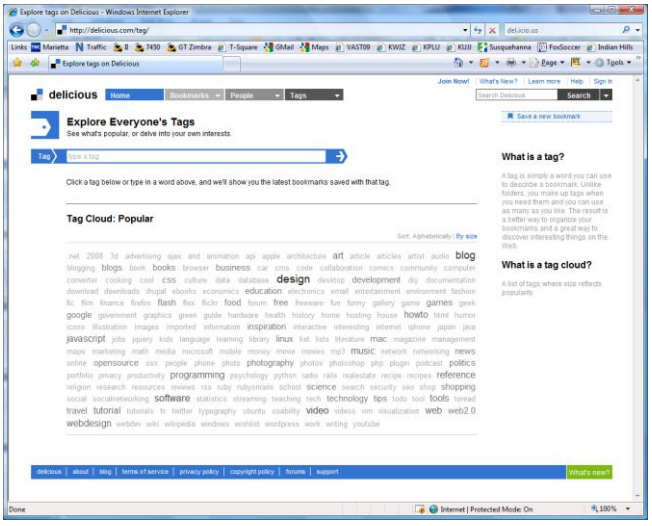
- Flanagan's '97 "Search referral Zeitgeist"
- Fortune's '01 Money Makes the World Go Round

Viégas & Wattenberg
interactions '08

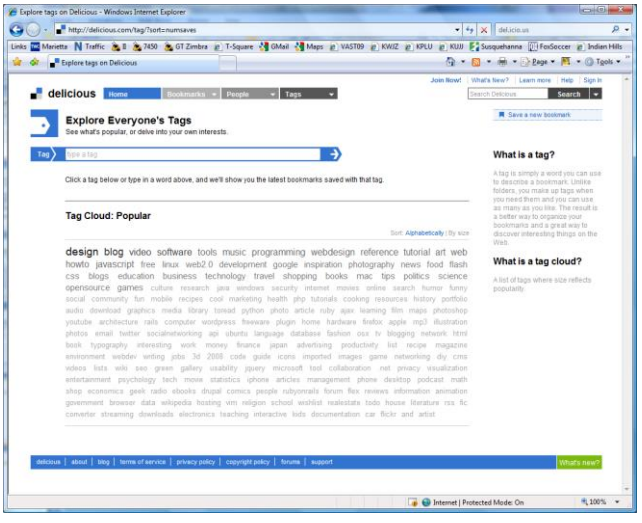
Flickr Tag Cloud



delicious Tag Cloud

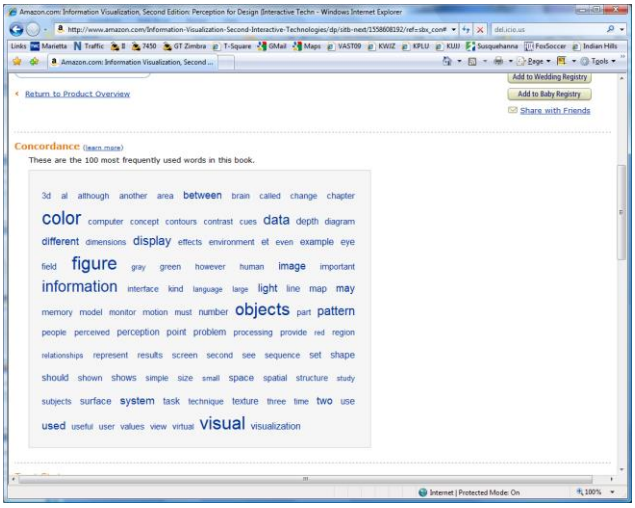


Alternate Order



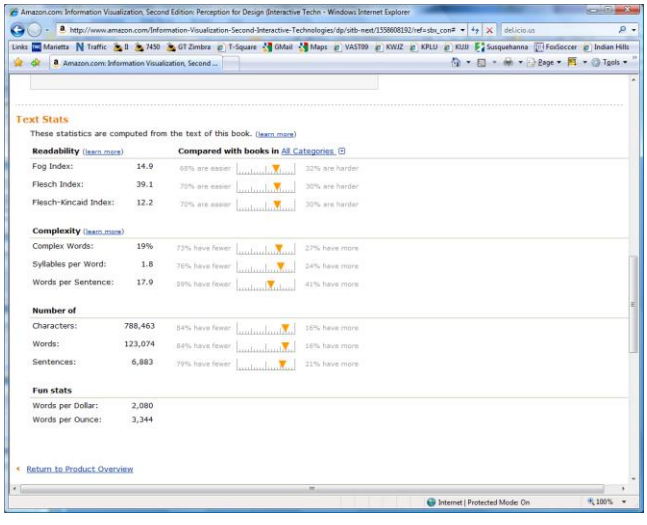
Amazon's (old) Product Concordance

Maybe now a "word cloud"



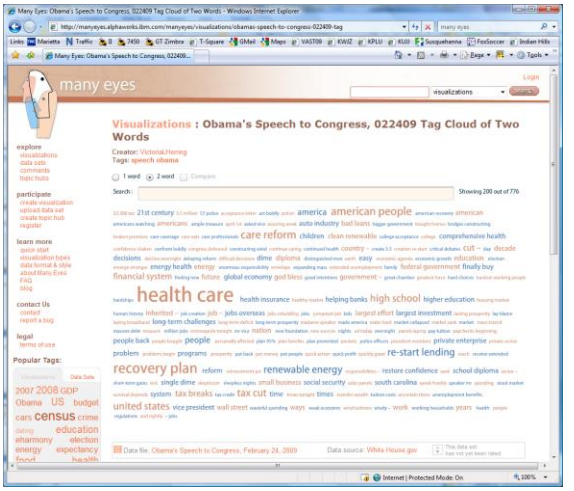
More (old) Info

There are other types of info about a document on Amazon



Many Eyes Tag Cloud

Here, pairs of words are shown

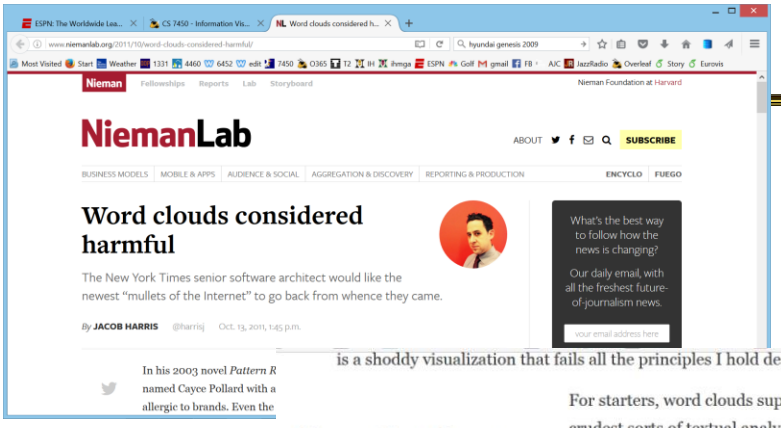


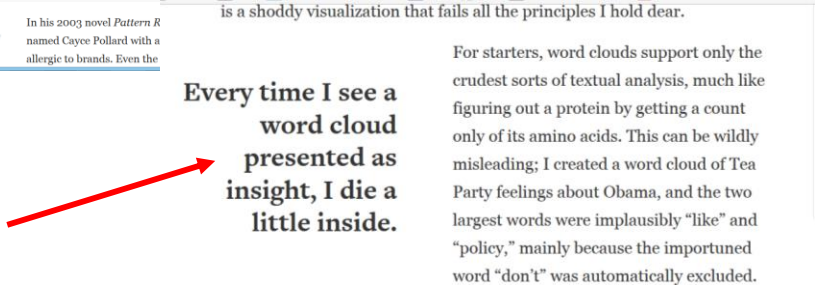
Problems

- 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

Gruen et al
CHI '06

http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/

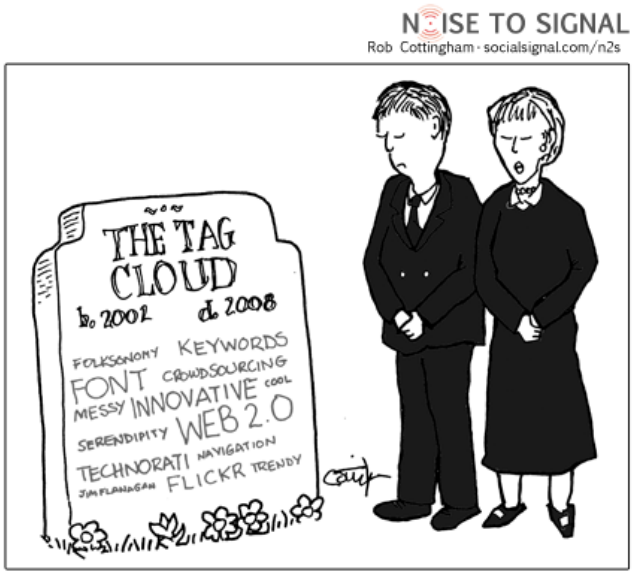




Why So Popular?

- Serve as social signifiers that provide a friendly atmosphere that provide a point of entry into a complex site
- Act as individual and group mirrors
- Fun, not business-like

Hearst & Rosner
HICSS '08



http://www.socialsignal.com/system/files/images/2008-08-01-tagcloud.gif

Wordle

http://www.wordle.net

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

Viegas, Wattenberg, & Feinberg
TVCG (InfoVis) '09

Layout Algorithm

- Details not published
- Idea:
  - sort words by weight, decreasing order
    for each word w
      w.position := makeInitialPosition(w);
      while w intersects other words:
        updatePosition(w);
  - Init position randomly chosen according to distribution for target shape
  - Update position moves out radially

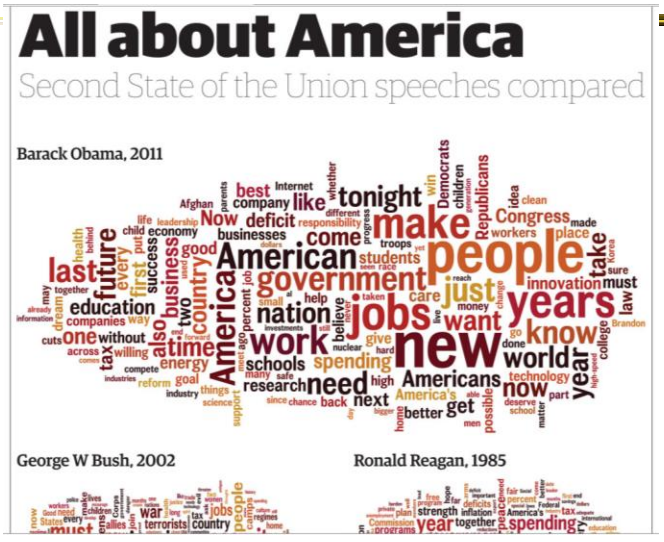
Fun Uses

- Political speeches
- Songs and poems
- Love letters (for "boyfriend points")
- Wedding vows
- Course syllabi
- Teaching writing
- Gifts

2-day Survey in Jan. 09

- 2/3 respondents were women
- Interest came from design, visual appeal, beauty
- Why preferred over word clouds:
  - Emotional impact
  - Attention-keeping visuals
  - Organic, non-linear
- Fair percentage didn't know what size signified

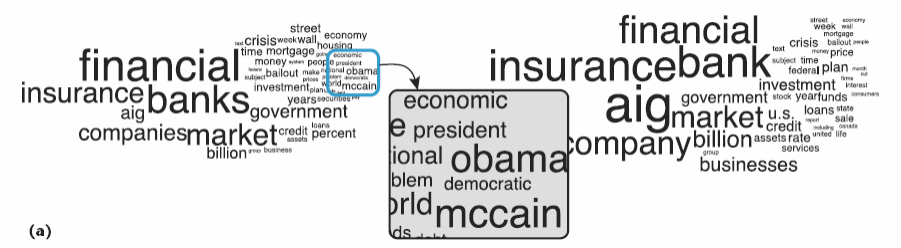
SoTU Wordles



http://www.guardian.co.uk/news/datablog/2011/jan/25/state-of-the-union-text-obama#

What variations of a word cloud/wordle can you think of?

A Little More Order



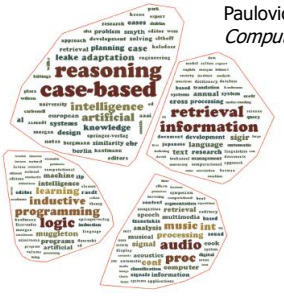
Order the words more by frequency

Cui et al
IEEE CG&A '10

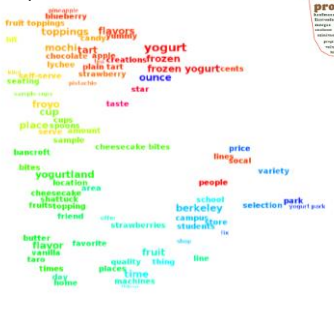
Semantic/Context Word Clouds

Group by
related concepts

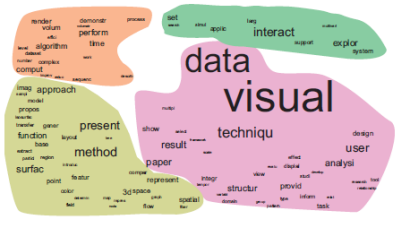
Paulovich et al
Computer Graphics Forum '12



Wang et al
Graphics Interface '14



Wu et al
Computer Graphics Forum '11



Wordle Characteristics

- Layout, words are automatic
- If you had some control, what would you like to change or alter?
  - Alter color (within a palette)
  - Pin words, redo the rest
  - Move and rotate words
  - Smooth animation and collision detection for tracking changes

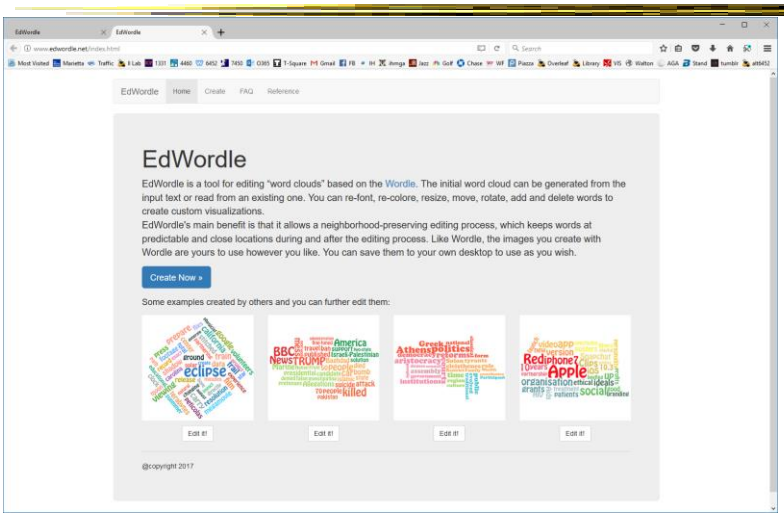
Systems

Mani-Wordle
Koh et al
TVCG (InfoVis) '10

EdWordle
Wang et al
TVCG (InfoVis 17) '18

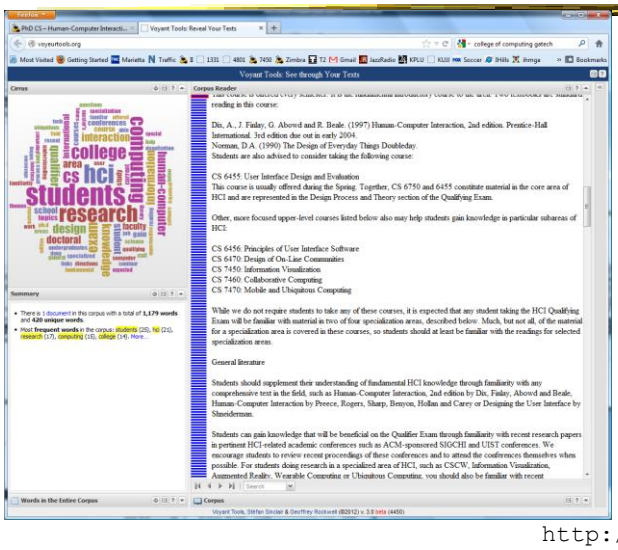
http://www.edwordle.net/

Example



Demo

Text Analysis on Web



Demo

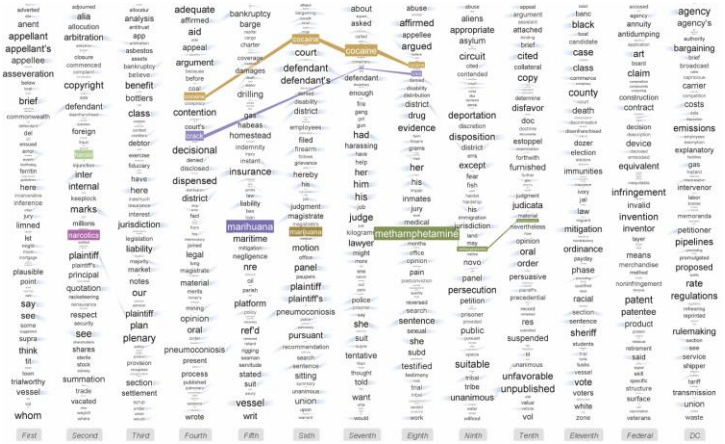
http://voyant-tools.org/

Multiple Documents?

- How do we show word frequencies across multiple related documents?

Ideas?

Parallel Tag Clouds



Different circuit courts

Video

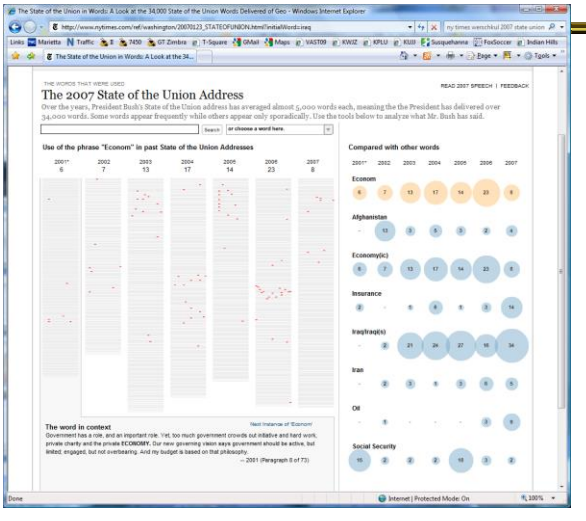
Collins et al
VAST '09

Analytic Support

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

- How might we also support queries and search?

Overview & Timeline

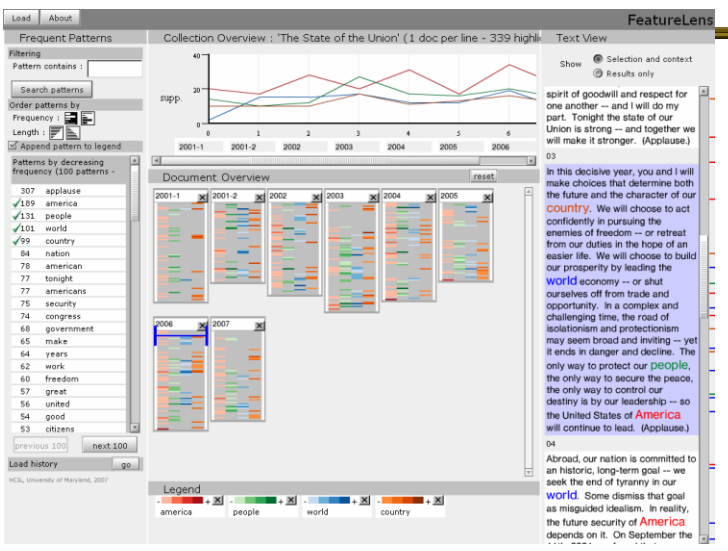


State of the Union Addresses

http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html?initialWord=iraq

FeatureLens

Video

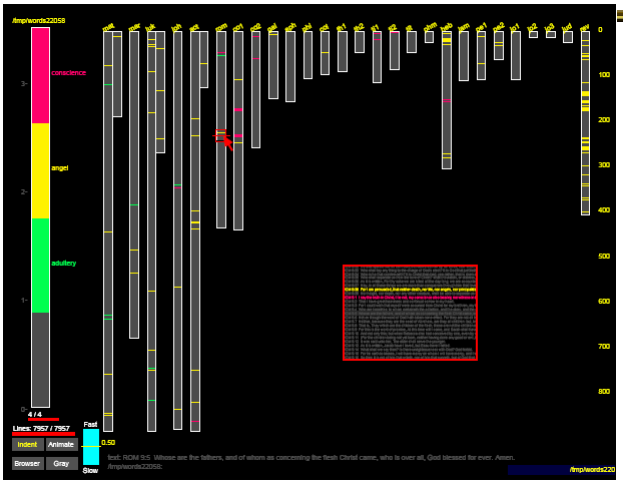


Show patterns of words or n-grams

Don et al
CIKM '07

http://www.cs.umd.edu/hcil/textvis/featurelens/

SeeSoft Display



Like taping text to the wall and walking far away

New Testament

Eick
Journal Comput. & Graph. Stats '94

Combinations

- What if you were interested in pairs of words (typically nouns) in documents, eg
  - X and Y
  - X's Y
  - X at Y
  - X (is|are|was|were) Y
- How visualize that?

Was added to Many Eyes

van Ham et al
TVCG (InfoVis) '09

Phrase Nets

- Examine unstructured text documents
- Presents pairs of terms (previous slide)
- Uses special graph layout algorithm with compression and simplification

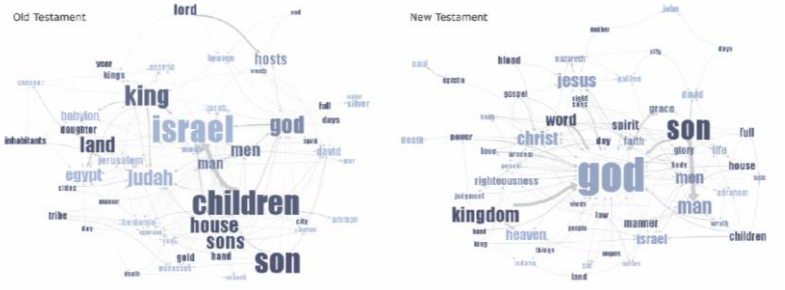


Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

Examples

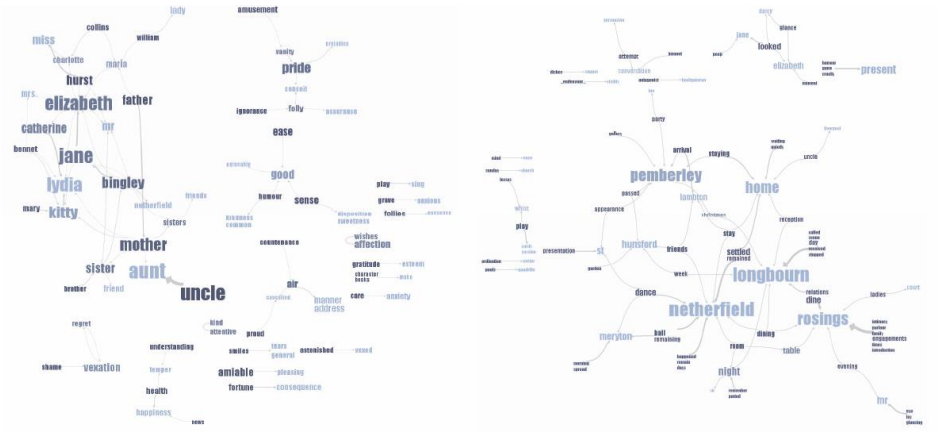


Fig 5. Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

User Interface

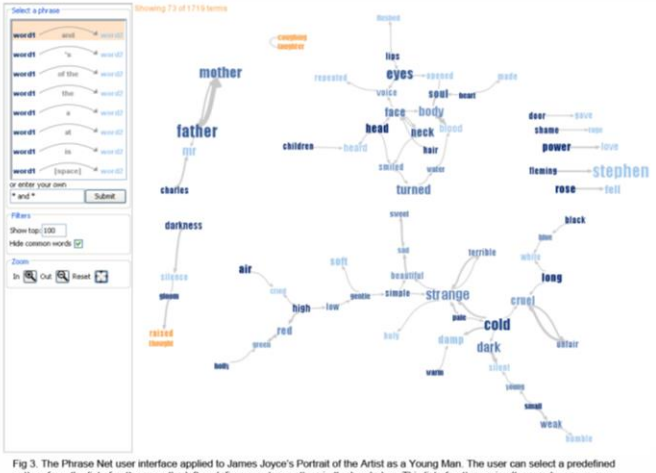


Fig 3. The Phrase Net user interface applied to James Joyce's Portrait of the Artist as a Young Man. The user can select a predefined pattern from the list of patterns on the left or define a custom pattern in the box below. This list of patterns simultaneously serves as a legend, a list of presets and an interactive training mechanism for regular expressions. Here the user has selected "...X and Y...", revealing two main clusters, one almost exclusively consisting of adjectives, the other of verbs and nouns. The highlighted clusters of terms have been aggregated by our edge compression algorithm.

Next Time

- More about text (beyond words) and collections of documents
  - Sentences
  - Analysis metrics
  - Entities
  - Concepts & themes

Learning Objectives

- Explain key challenges in visualizing a large document or body of text
- Identify and explain different techniques for representing words and concepts in a document
  - Word cloud, Wordle, Parallel tag cloud, SeeSoft, PhraseNet
- Understand the positives and limitations of word clouds and Wordles
- Describe SeeSoft-style miniature visual representations

Upcoming

- Text and Documents 2
  - Prep: Watch Bohemian Bookshelf video
- Lab 9: Layout in D3

References

- All referred to papers