

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

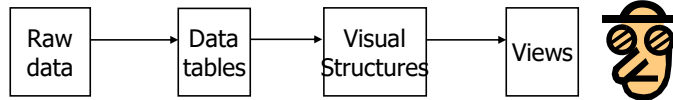


- What can information visualization provide to help users in understanding and gathering information from text and document collections?

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



Fall 2017

CS 4460

5

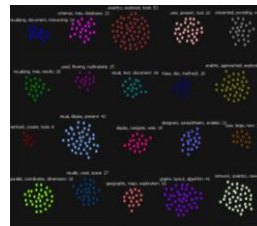
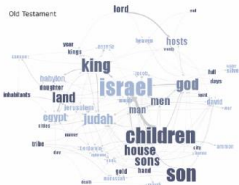
This Week's Agenda



Visualizing text
Showing words,
combinations, and
context



Visualizing document sets
Words & sentences
Analysis metrics
Concepts & themes



Fall 2017

CS 4460

6



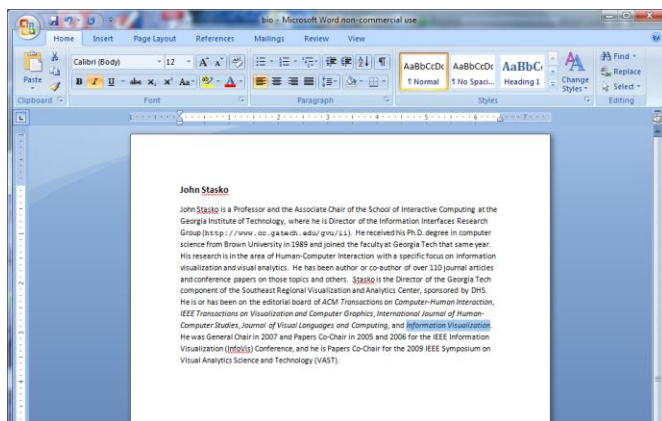
What's the simplest text visualization you know?

Fall 2017

CS 4460

7

One Text Visualization



Uses:
Layout
Font
Style
Color

...

Fall 2017

CS 4460

8

Design Challenge



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

Fall 2017

CS 4460

9



What was implicit in this exercise?

Fall 2017

CS 4460

10

Tasks



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

Fall 2017

CS 4460

11

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

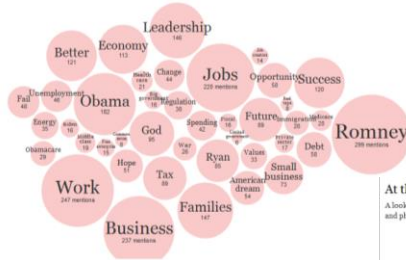
Word Counts



At the Republican Convention, the Words Being Used

A look at how often speakers at the Republican National Convention have used certain words and phrases so far, based on an analysis of transcripts from the Federal News Service.

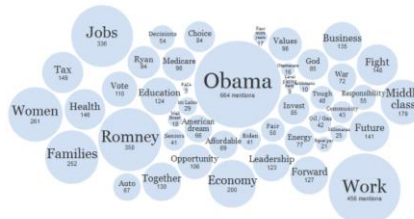
Add word or phrase



At the Democratic Convention, the Words Being Used

A look at how often speakers at the Democratic National Convention have used certain words and phrases so far, based on an analysis of transcripts from the Federal News Service.

Add word or phrase



Fall 2017

CS 4460

12

More Word Counting



WORDCOUNT

◀ PREVIOUS WORD NEXT WORD ▶

the of and to a in that is was for on you be with by he the at to be with the

1 2 3 4 5 6

CURRENT WORD

FIND WORD: BY RANK: REQUESTED WORD: THE RANK: 1 88800 WORDS IN ARCHIVE ABOUT WORDCOUNT

WordCount™ ©2003 Jonathan Harris | Number27 | Help

<http://www.wordcount.org>

Fall 2017

CS 4460

13

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

Fall 2017

CS 4460

14

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



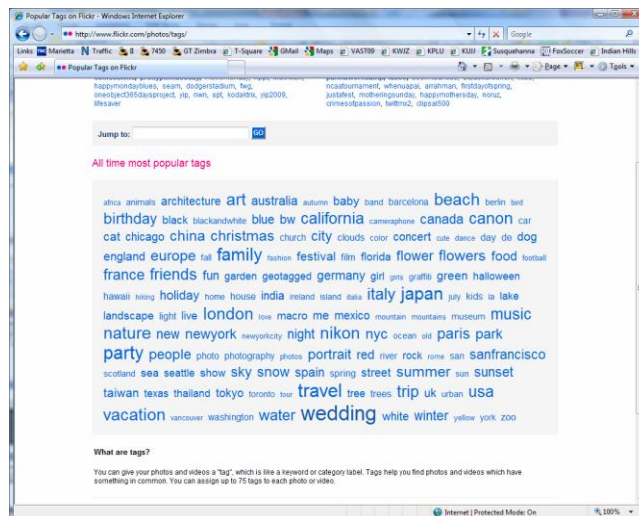
Viégas & Wattenberg
interactions '08

Fall 2017

CS 4460

15

Flickr Tag Cloud

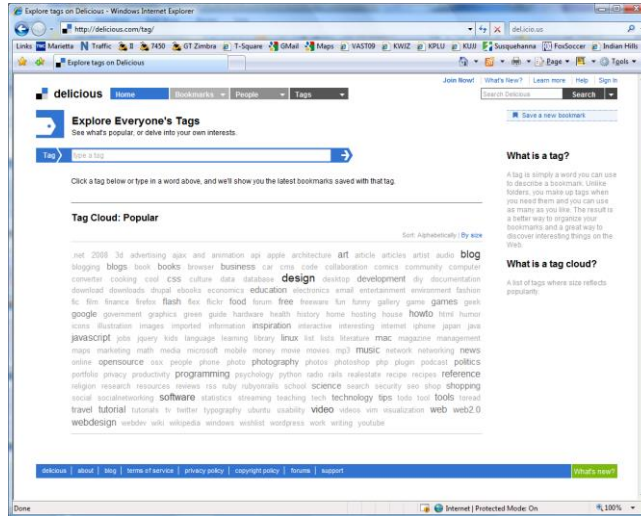


Fall 2017

CS 4460

16

delicious Tag Cloud

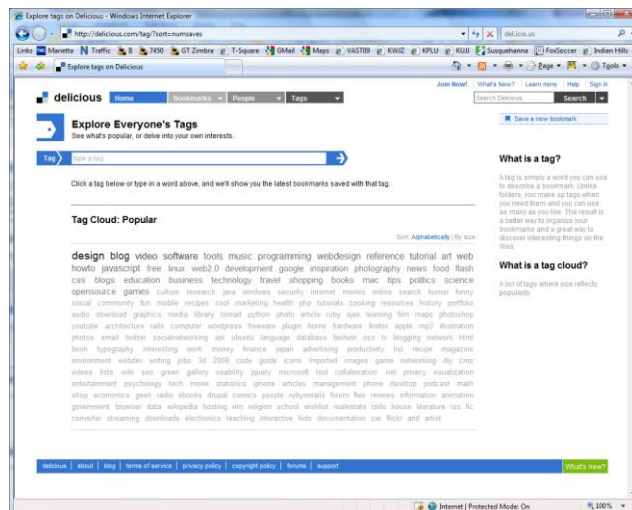


Fall 2017

CS 4460

17

Alternate Order



Fall 2017

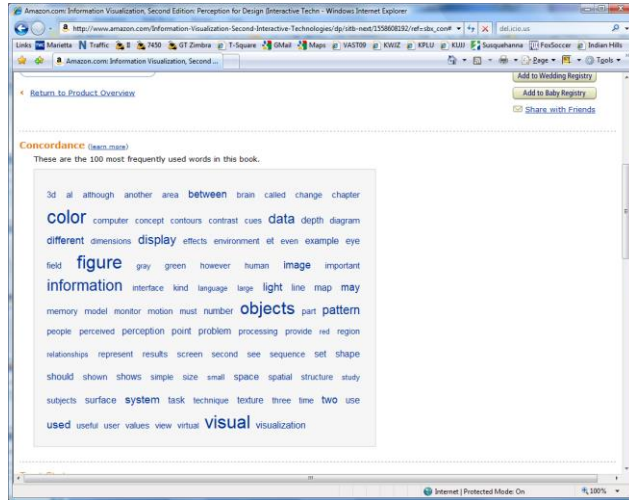
CS 4460

18

Amazon's (old) Product Concordance



Maybe now a
"word cloud"



Fall 2017

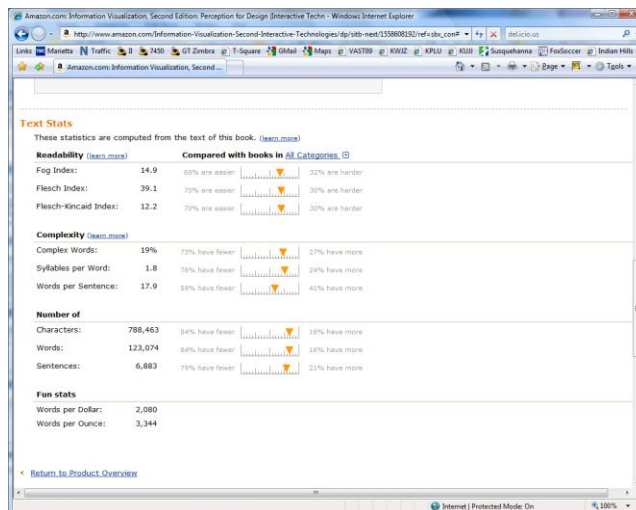
CS 4460

19

More (old) Info



There are
other types
of info about
a document
on Amazon



Fall 2017

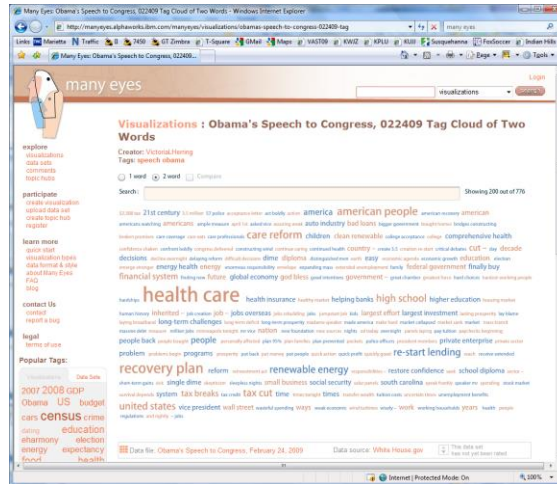
CS 4460

20

Many Eyes Tag Cloud



Here, pairs of words are shown



Fall 2017

CS 4460

21

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

Fall 2017

CS 4460

22

The screenshot shows a web browser displaying the NiemanLab website. The article title is "Word clouds considered harmful" by Jacob Harris, dated Oct 13, 2011. A red arrow points to a sentence in the article: "Every time I see a word cloud presented as insight, I die a little inside." Below the screenshot, there is a quote: "is a shoddy visualization that fails all the principles I hold dear." To the right of the quote is a paragraph of text: "For starters, word clouds support only the crudest sorts of textual analysis, much like figuring out a protein by getting a count only of its amino acids. This can be wildly misleading; I created a word cloud of Tea Party feelings about Obama, and the two largest words were implausibly 'like' and 'policy,' mainly because the importuned word 'don't' was automatically excluded."

Fall 2017

CS 4460

23

Why So Popular?

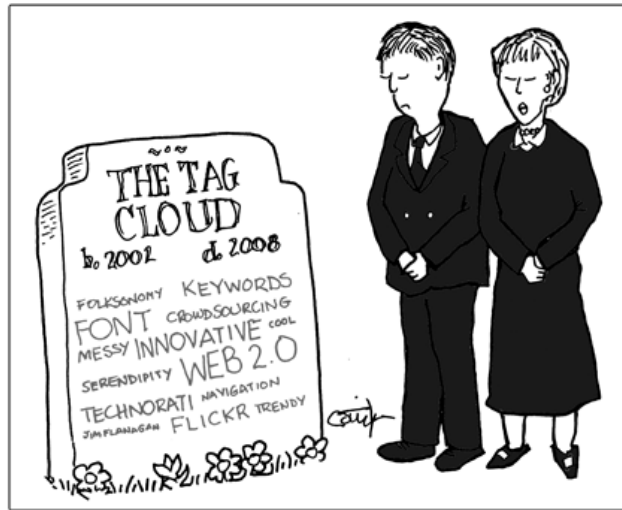
- Serve as social signifiers that provide a friendly atmosphere that provide a point of entry into a complex site
- Act as individual and group mirrors
- Fun, not business-like

Hearst & Rosner
HICSS '08

Fall 2017

CS 4460

24



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

Fall 2017

CS 4460

25

Wordle

<http://www.wordle.net>

can do volunteering
 can do volunteering from Scope, Leonard Cheshire and Russell
 Commission. Available at www.scope.org.uk —
<https://www.facebook.com/leonardcheshire.org/> 7 minutes ago



Women's Rights
 Women have rights too! — macdoodle11 11 minutes ago



Fall 2017

"verschreibbar" by Daniela 5 minutes ago



"Generals Douglas McArthur's Speech" by Bob the Builder 31 minutes ago



CS 4460

26

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

Fall 2017

CS 4460

27

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

Fall 2017

CS 4460

28

Fun Uses



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

Fall 2017

CS 4460

29

2-day Survey in Jan. 09



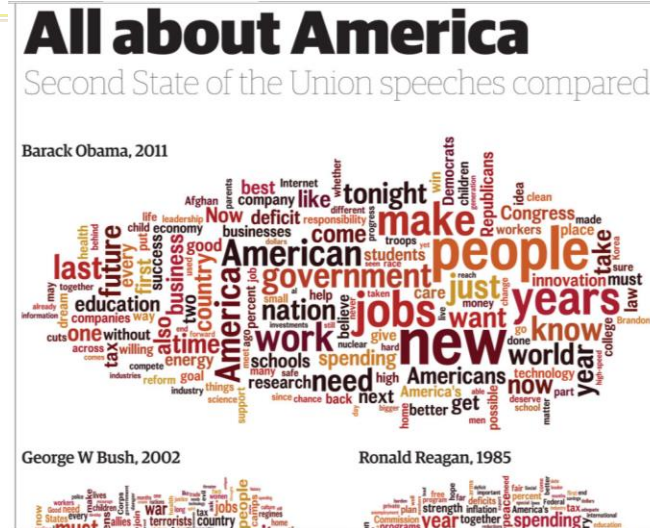
- 2/3 respondents were women
- Interest came from design, visual appeal, beauty
- Why preferred over word clouds:
 - Emotional impact
 - Attention-keeping visuals
 - Organic, non-linear
- Fair percentage didn’t know what size signified

Fall 2017

CS 4460

30

SoTU Wordles



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

Fall 2017

CS 4460

31

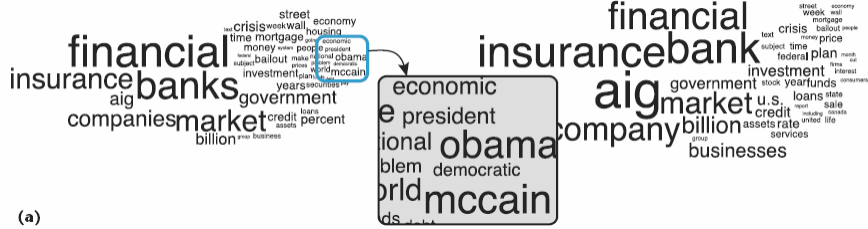
What variations of a word cloud/wordle can you think of?

Fall 2017

CS 4460

32

A Little More Order



Order the words more by frequency

Cui et al
IEEE CG&A '10

Fall 2017

CS 4460

33

Semantic/Context Word Clouds

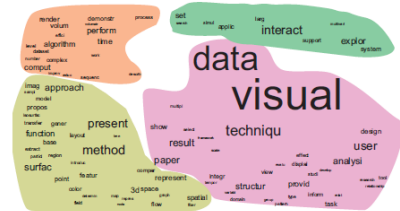
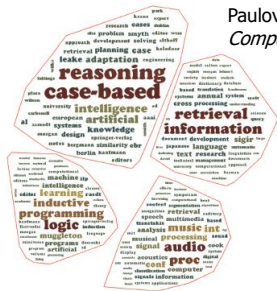
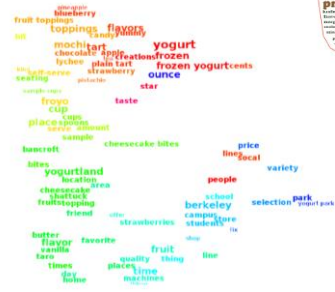


Group by
related concepts

Paulovich et al
Computer Graphics Forum '12

Wang et al
Graphics Interface '14

Wu et al
Computer Graphics Forum '11



Fall 2017

CS 4460

34

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

Fall 2017

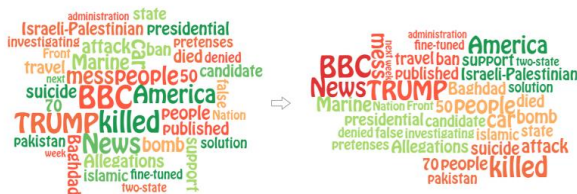
CS 4460

35

Systems



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



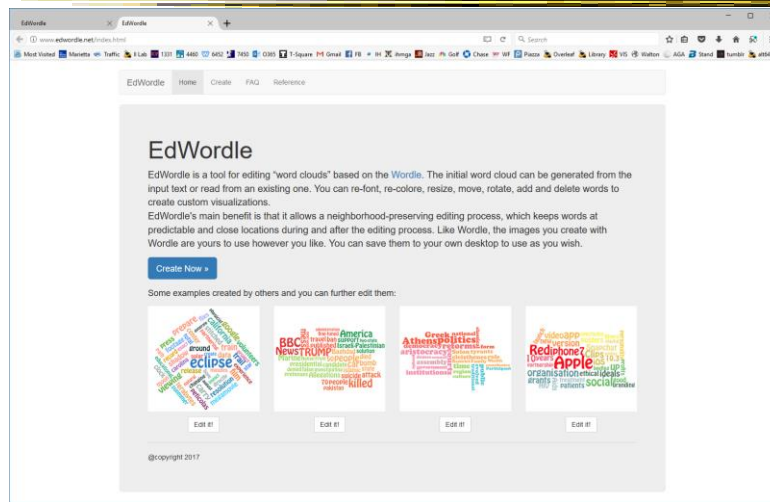
EdWordle
Wang et al
TVCG (InfoVis 17) '18

Fall 2017

CS 4460

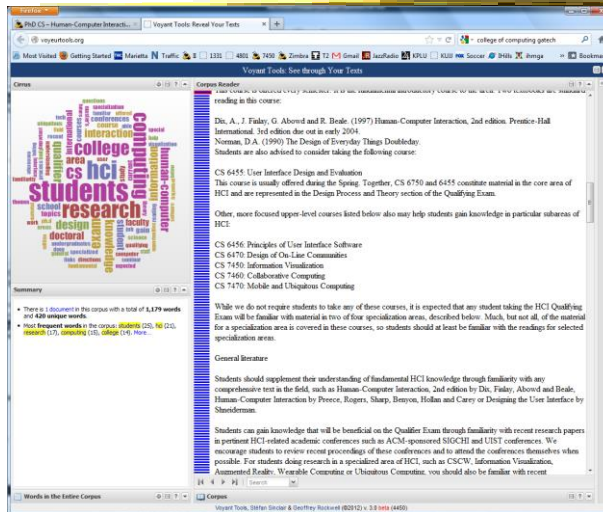
36

Example



Demo

Text Analysis on Web



Demo

Multiple Documents?



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

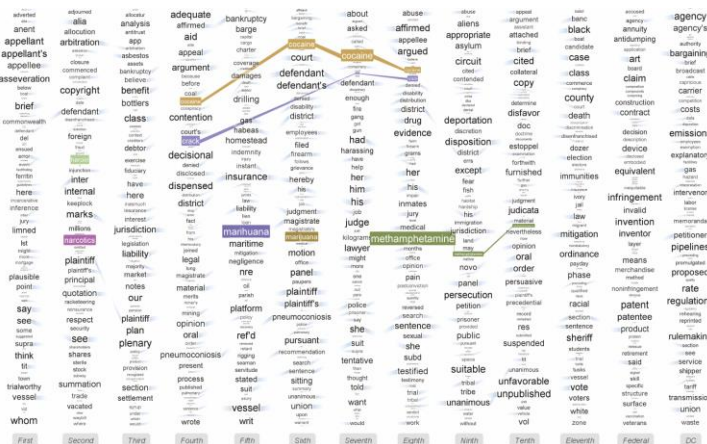
Ideas?

Fall 2017

CS 4460

39

Parallel Tag Clouds



Video

Different circuit courts

Collins et al
VAST '09

Fall 2017

CS 4460

40

Analytic Support



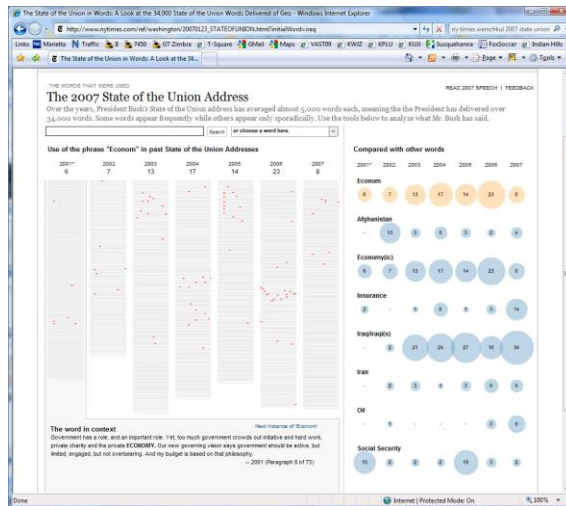
- Note: Word Clouds and Wordles are really more overview-style visualizations
 - Don't really support queries, searches, drill-down
- How might we also support queries and search?

Fall 2017

CS 4460

41

Overview & Timeline



State of the Union Addresses

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

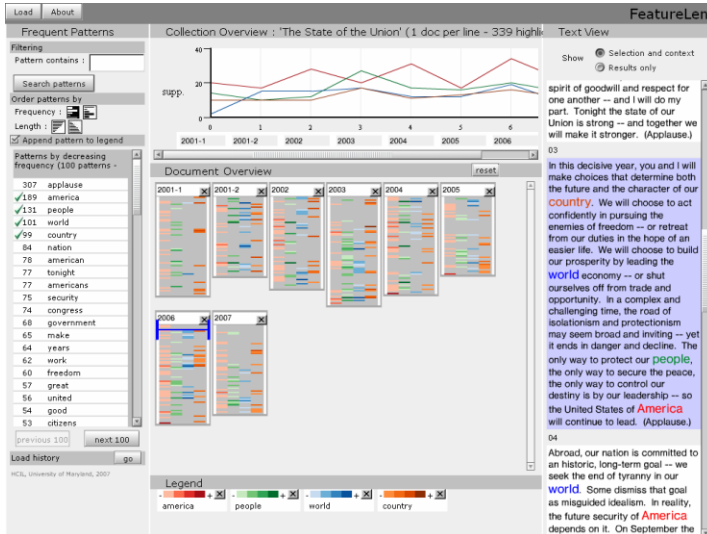
Fall 2017

CS 4460

42

FeatureLens

Video



Show patterns of words or n-grams

Don et al
CIKM '07

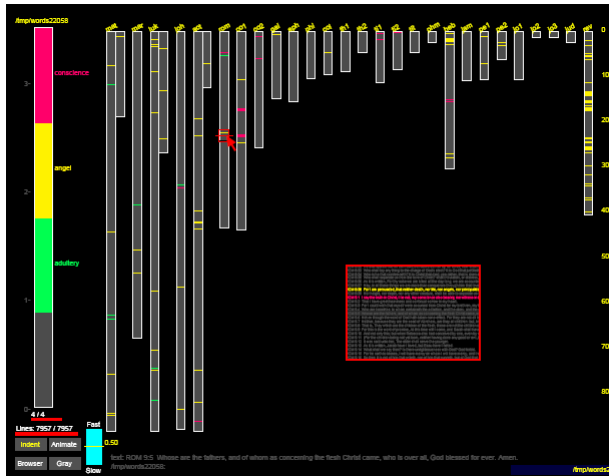
<http://www.cs.umd.edu/hcil/textvis/FeatureLens/>

Fall 2017

CS 4460

43

SeeSoft Display



Like taping text to the wall and walking far away

New Testament

Eick
Journal Comput. & Graph. Stats '94

Fall 2017

CS 4460

44

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?

Fall 2017

CS 4460

45

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

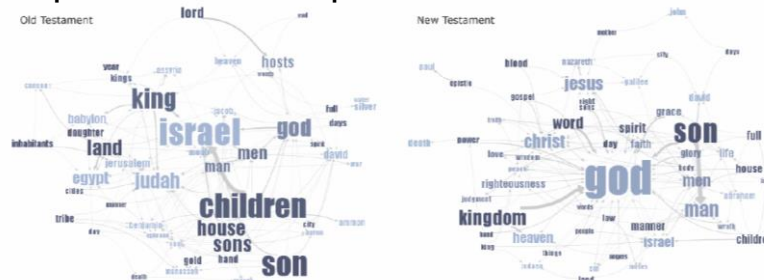


Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

Fall 2017

CS 4460

46

Examples

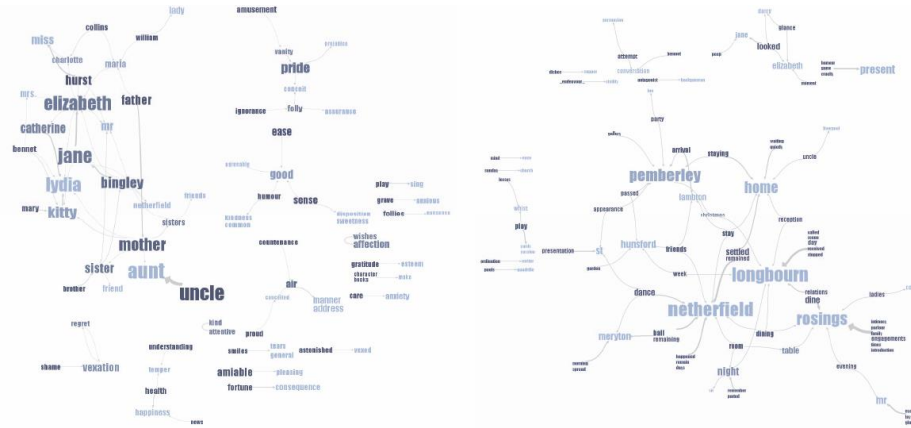


Fig 5. Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

Fall 2017

CS 4460

47

User Interface

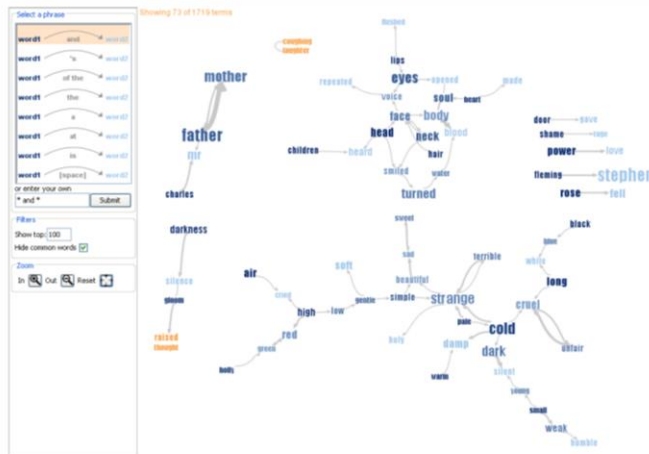


Fig 3. The Phrase Net user interface applied to James Joyce's *Portrait of the Artist as a Young Man*. The user can select a predefined pattern from the list of patterns on the left or define a custom pattern in the box below. This list of patterns simultaneously serves as a legend, a list of presets 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.

Fall 2017

CS 4460

48

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

Fall 2017

CS 4460

51

References



- All referred to papers

Fall 2017

CS 4460

52