Text and Document Visualization - Part 1

CS 4460 - Intro to Information Visualization

Spring, 2019

Alex Endert

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
  - email
  - books
  - news articles
  - blogs
  - speeches (after converted to text)
  - …

Big Questions

- What can information visualization provide to help users in understanding and gathering information from text and document collections?
- What questions/tasks can visualization help people answer/perform?
- What are some basic visualization techniques specifically for textual data?
- We will cover these over the next 2 lectures.

Tasks/Goals

- What kinds of analysis questions might a person ask about text & documents?

Example Tasks & Goals

- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- Which documents have an angry tone?
- How are certain words or themes distributed through a document?
- Identify “hidden” messages or stories in this document collection.
- How does one set of documents differ from another set?
- Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
- Understand the history of changes in a document.
- Find connections between documents.

Text as data

what are the attributes? items?

Challenge

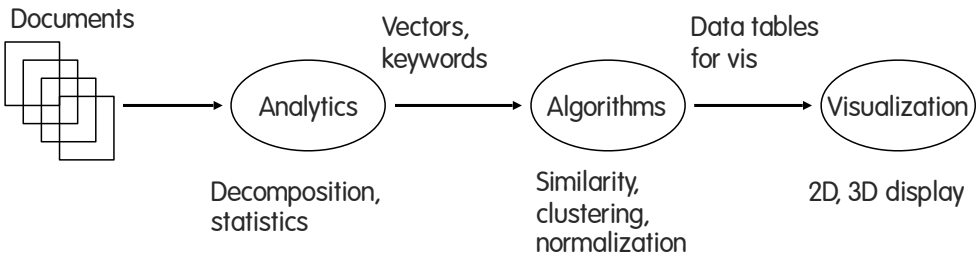
- Text is nominal data
  - Does not seem to map to geometric/graphical presentation as easily as ordinal and quantitative data
  - No natural “order” or “rank” of documents or words
- The “Raw data --> Data Table” mapping now becomes more important
  - For text visualization, *part of the challenge is actually creating this mapping in your system*.
  - i.e., how do you go from a folder of .txt files, to something you can visualize?
- Before we start talking about InfoVis techniques, let’s do a quick overview of how we define the data items and attributes

basic data terminology

- Data **items**
  - typically a single document
  - often the rows in your table
- Data **attributes**
  - a.k.a, “entities” or “terms”
  - words that are extracted from your entire dataset
  - often the columns in your dataset
- Often creates a very sparse data table
  - not every document contains every word in your dataset of documents
  - a set of documents creates a very large set of words

let's walk through an example

- Text data can be represented as hypervariate data



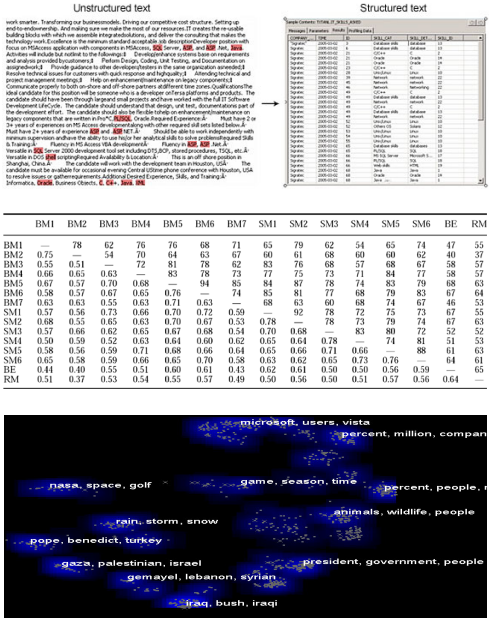
Unstructured Text

Extract Entities

High Dimensional Data

Dimension Reduction Model

Choice of Visual Encodings/ Metaphor



More complex examples in Visual Analytic lectures

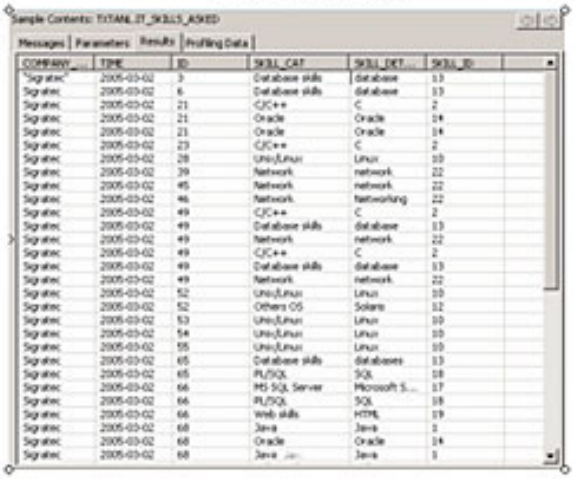
Unstructured Text

Extract Entities

Unstructured text

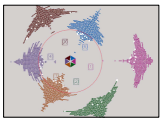
work smarter. Transforming our businessmodels. Driving our competitive cost structure. Setting up end-to-endownership. And making sure we make the most of our resources.IT creates the re-usable building blocks with which we assemble integratedsolutions, and deliver the consulting that makes the technology work.Excellence is the minimum standard acceptable.Job descriptionDeveloper position with focus on MSAccess application with components in MSAccess, SQL Server, ASP, and ASP .Net, Java. Activities will include but notlimit to the followings:‡ Develop/enhance systems base on requirements and analysis provided bycustomers;‡ Perform Design, Coding, Unit Testing, and Documentation on assignedwork;‡ Provide guidance to other developers/testers in the same organization asneeded;‡ Resolve technical issues for customers with quick response and highquality;‡ Attending technical and project management meetings;‡ Help on enhancement/maintenance on legacy components;‡ Communicate properly to both on-shore and off-shore partners atdifferent time zones.QualificationsThe ideal candidate for this position will be someone who is a developer onTersa platforms and products. The candidate should have been through largeand small projects and have worked with the full IT Software Development LifeCycle. The candidate should understand that design, unit test, documentations part of the development effort. The candidate should also be flexible tohelp on enhancement/maintenance on legacy components that are written in Pro*C, PL/SQL, Oracle.Required Experience:Â· Must have 2 or 3+ years of experiences on MS Access developmentalong with other required skill sets listed below.Â· Must have 2+ years of experience ASP and .ASP NET.Â· Should be able to work independently with minimum supervision andhave the ability to use his/or her analytical skills to solve problemsRequired Skills & Training:Â· Fluency in MS Access VBA developmentÂ· Fluency in ASP, ASP .Net.Â· Versatile in SQL Server 2000 development tool set including DTS,BCP, stored procedures, TSQL, etc.Â· Versatile in DOS shell scriptingRequired Availability & Location:Â· This is an off shore position in Shanghai, China.Â· The candidate will work with the development team in Houston, USÂ· The candidate must be available for occasional evening Central UStime phone conference with Houston, USA to resolve issues or gatherrequirements.Additional Desired Experience, Skills, and Training:Â· Informatica, Oracle, Business Objects, C, C++, Java, XML

Structured text

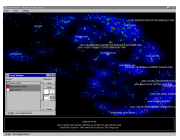


This Week's Agenda

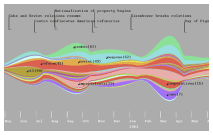
Visualization for IR
Helping search



Visualizing text
Showing words, phrases, and sentences



Visualizing document sets
Words, entities & sentences
Analysis metrics
Concepts & themes



InfoVis → Visual Analytics

InfoVis for Information Retrieval for Textual Data

Information Retrieval

- Can InfoVis help text IR?
  - YES!
- Assume there is some active search or query
  - Show results visually
  - Show how query terms relate to results
  - …

Improving Text Searches

- Lots of tasks for people have to do with searching for text information, then trying to make sense of the results
- Visualizing the results of search queries is one potential important area of text infovis

What Hearst Thinks is Wrong

- Query responses do not include:
  - How strong the match is
  - How frequent each term is
  - How each term is distributed in the document
  - Overlap between terms
  - Length of document
- Document ranking is opaque
- Inability to compare between results
- Input limits term relationships

Hearst
CHI '95

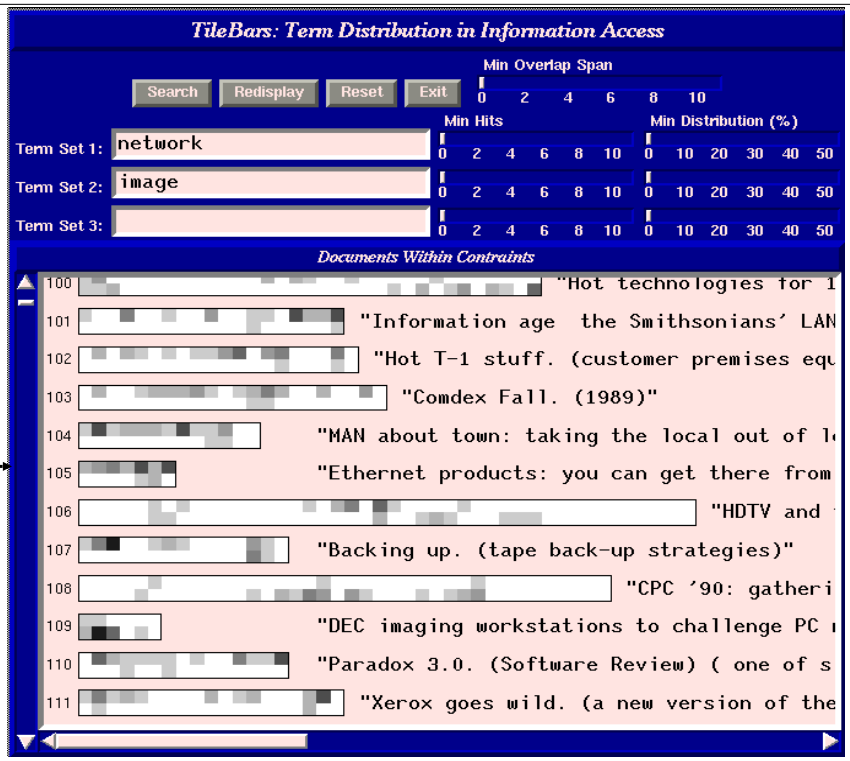
TileBars

- Graphical representation of term distribution and overlap
- Simultaneously indicate:
  - Relative document length
  - Frequency of term sets in document
  - Distribution of term sets with respect to the document and each other

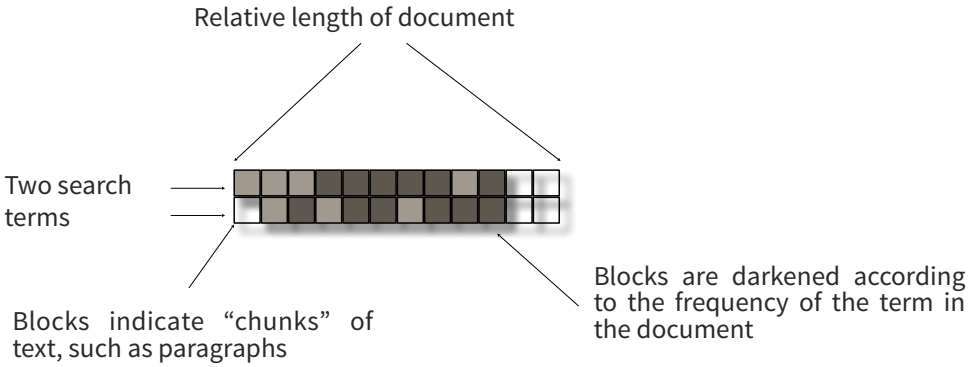
Interface

Search terms

Presentation



Technique



Issues

- Horizontal alignment doesn't match mental model of length of documents
  - i.e., documents don't really get "wider" (horizontal), they get "longer" (vertical)
- May not be the best solution for web searches
  - Images? Apps?

Generalize More

- How about the “holy grail” of a visual search engine?
  - Hot idea for a while, maybe not so much now?

Search Visualization

http://www.kartoo.com

pretty much done for…

Sparkler

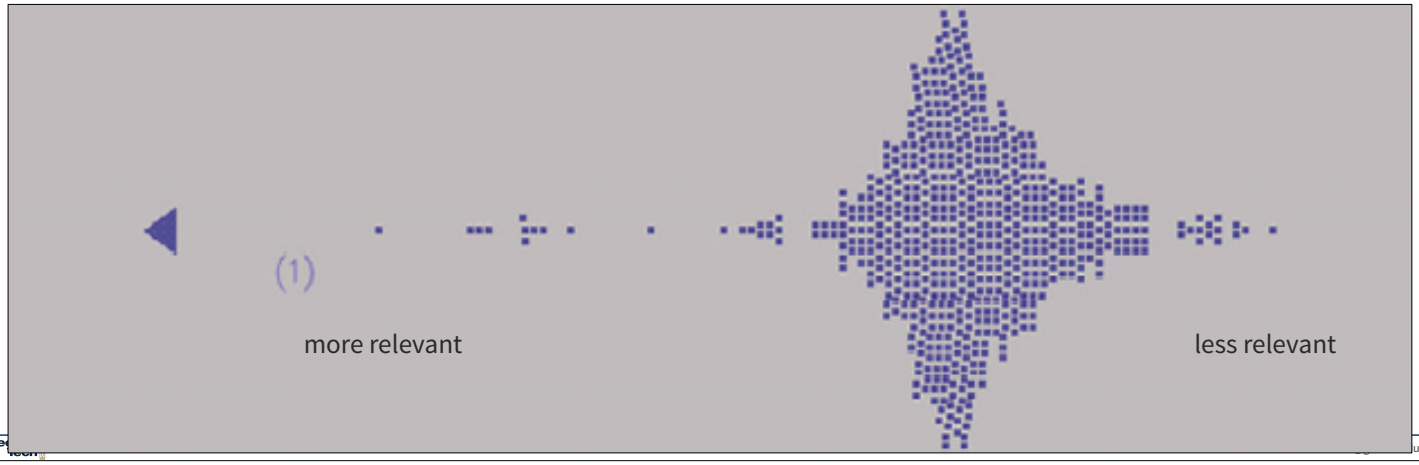
- *How good is the quality or similarity of the search result to my query?*

- Show “distance” from query in order to give user better feel for quality of match(es)
- Also shows documents in responses to multiple queries

Havre et al
InfoVis ‘01

Visualizing One Query

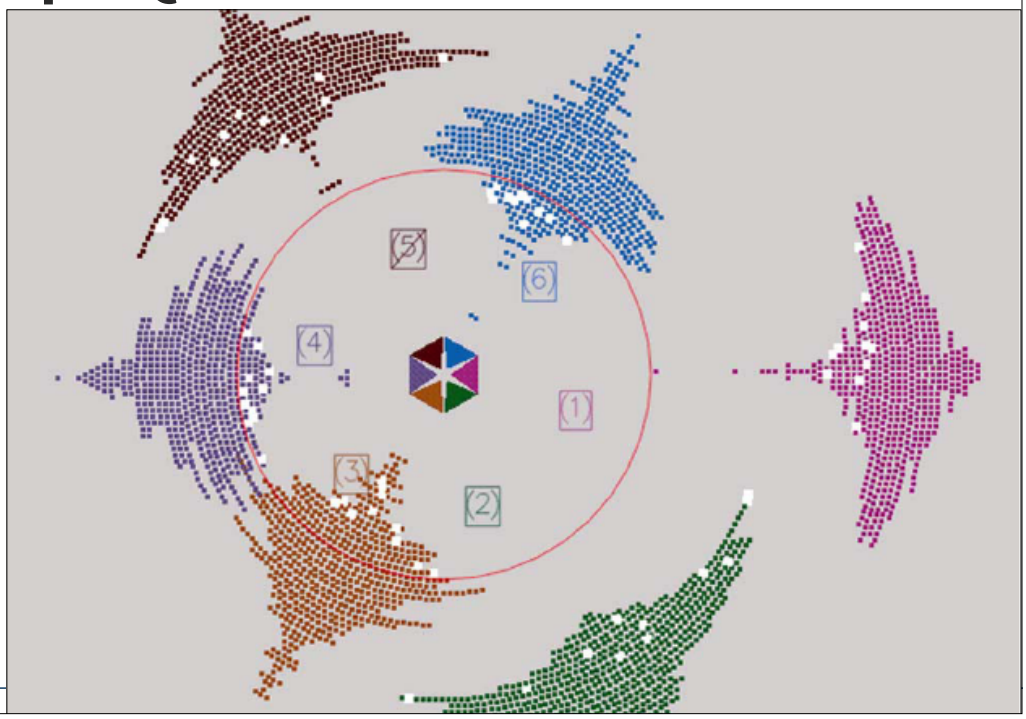
- Triangle – query
- Square – document
- Distance between query and documents represents their relevance



Visualizing Multiple Queries

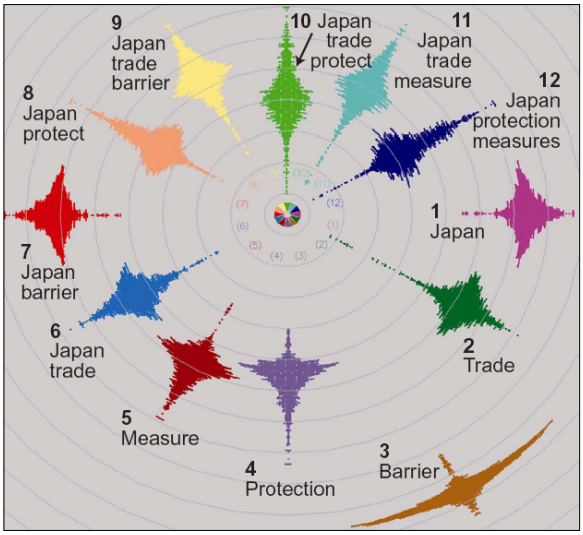
Six queries here

Bullseye allows viewer to select quality results

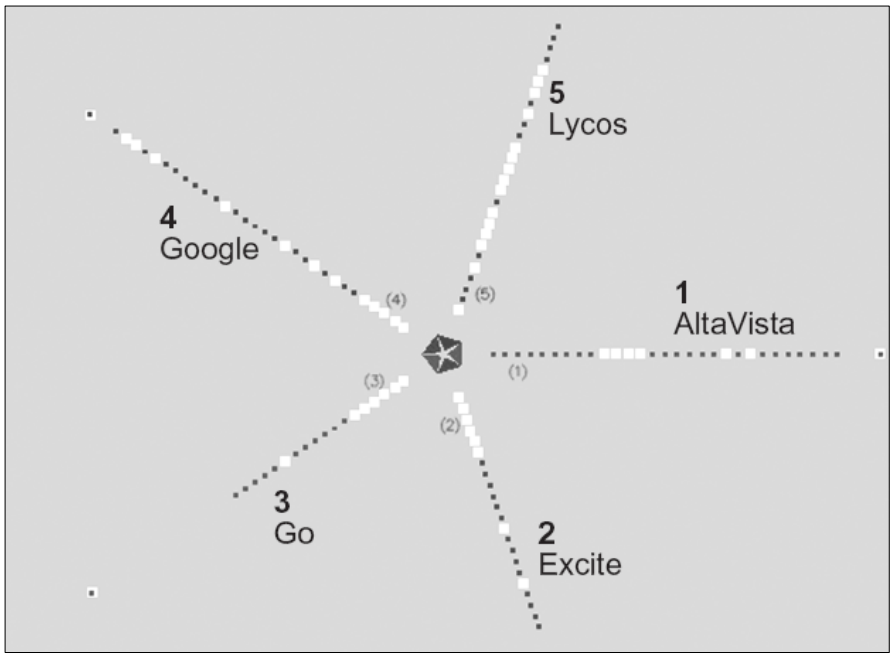


Test Example

- Text Retrieval Conference test document collection
- AP news stories from June 24–30, 1990
- TREC topic: Japan Protectionist Measures
- Sparkler found 16 of 17 relevant documents

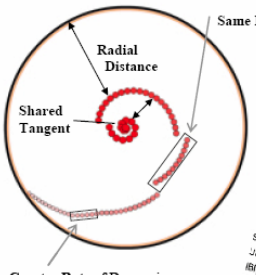


Another Idea



Use it to compare search results from different search engines

RankSpiral



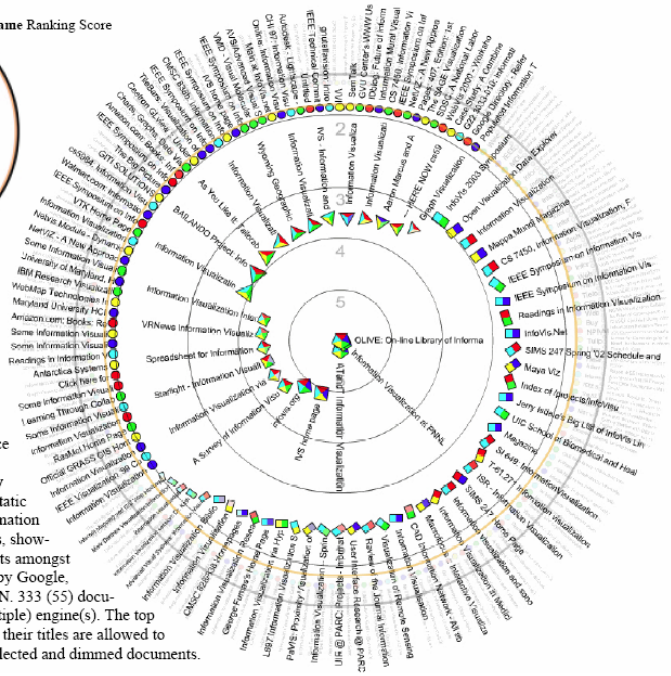


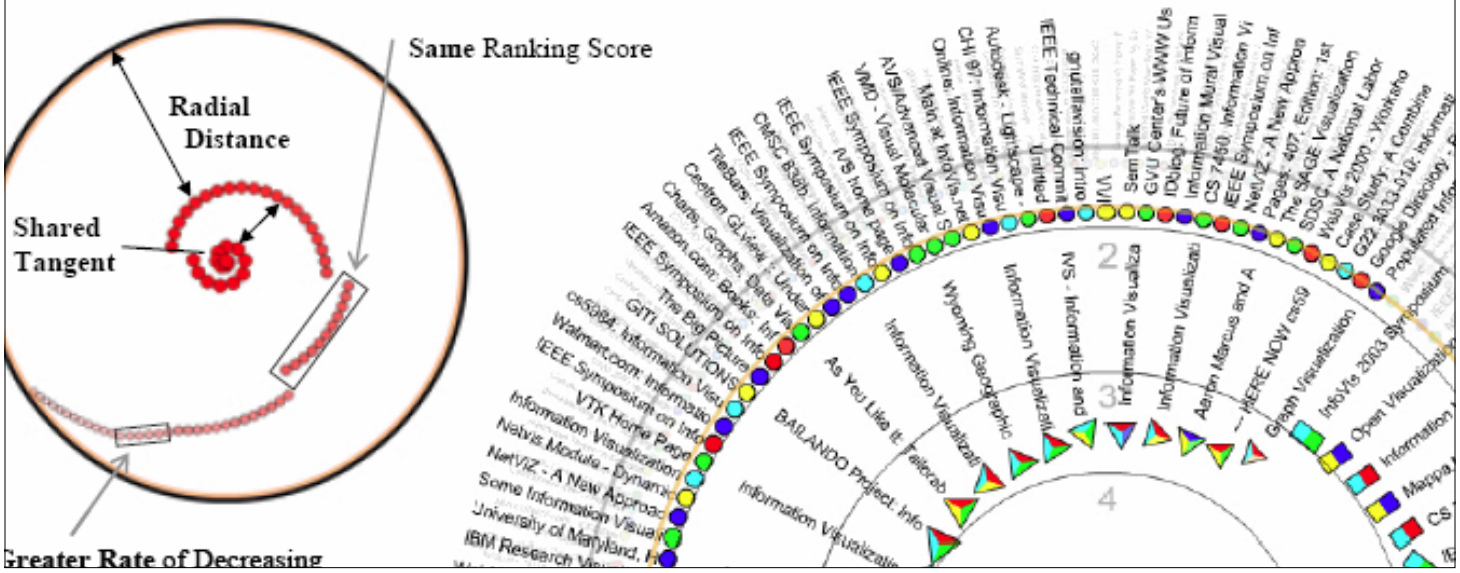
Figure 1: (Top) *RankSpiral* places consecutive document icons next to each other so that they do not overlap. Total ranking score of documents increases to-ward the center. Radial distance between documents that have the same angle can be used to display title fragments. (Right) shows a static RankSpiral that maximizes information density and minimizes occlusions, showing here the 388 unique documents amongst the top 100 documents retrieved by Google, Teoma, AltaVista, Lycis and MSN. 333 (55) documents were found by single (multiple) engine(s). The top 100+ documents are selected and their titles are allowed to extend across the remaining unselected and dimmed documents.

Color represents different search engines

Spoerri
InfoVis '04 poster

RankSpiral

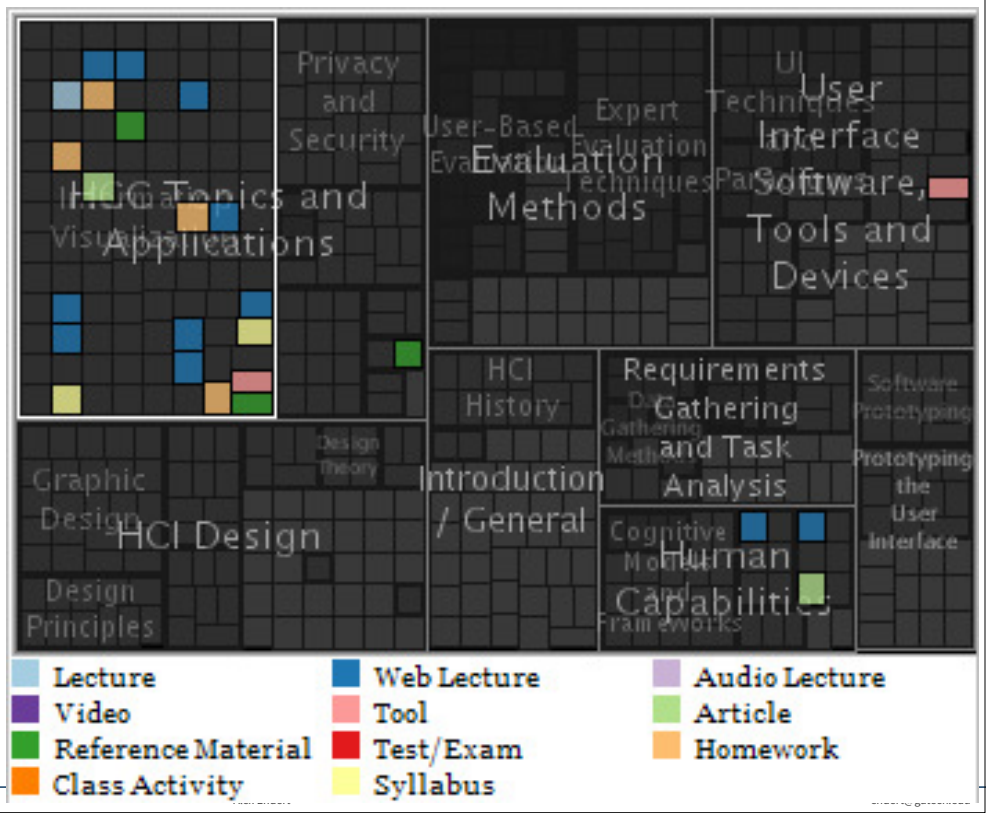
Color represents different search engines



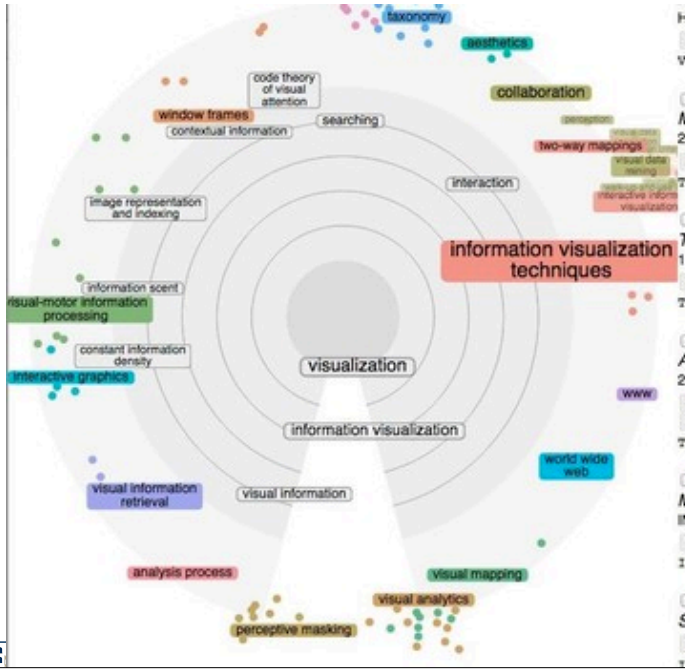
ResultMaps

Clarkson, Desai & Foley
TVCG (InfoVis) '09

Treemap-style vis for showing query results in a digital library



Interactive Intent Modeling for Text Search



Human Factors in Computing Systems (Summary), 1998-01-01)
visualization | world wide web | information visualization
Visual representation of information req...

A scalable framework for information visualization
M Kreuseler, N Lopez, H Schumann (IEEE SYMPOSIUM ON INFORMATION VISUALIZATION 2000, 2000-01-01)
focus+context | visualization | information visualization
This paper describe major concepts of a ...

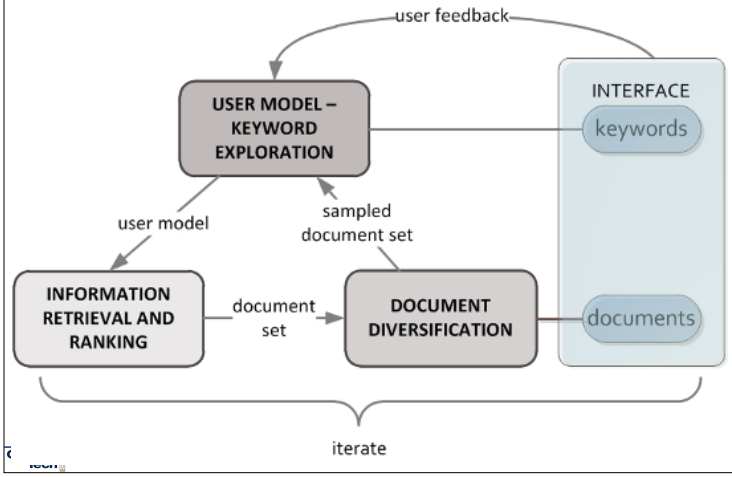
Visualized information sharing for negotiation
T Teraoka, M Akiyoshi (DESIGN OF COMPUTING SYSTEMS: COGNITIVE CONSIDERATIONS, 1997-01-01)
visualization | interaction | collaboration
This paper describes a framework to shar...

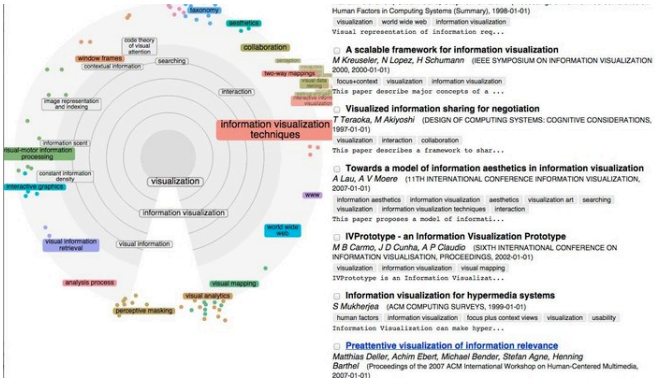
Towards a model of information aesthetics in information visualization
A Lau, A V Moere (11TH INTERNATIONAL CONFERENCE INFORMATION VISUALIZATION, 2007-01-01)
information aesthetics | information visualization | aesthetics | visualization art | searching | visualization | information visualization techniques | interaction
This paper proposes a model of informati...

IVPrototype - an Information Visualization Prototype
M B Carmo, J D Cunha, A P Claudio (SIXTH INTERNATIONAL CONFERENCE ON INFORMATION VISUALISATION, PROCEEDINGS, 2002-01-01)
visualization | information visualization | visual mapping
IVPrototype is an Information Visualizat...

Information visualization for hypermedia systems
S Mukherjea (ACM COMPUTING SURVEYS, 1999-01-01)
human factors | information visualization | focus plus context views | visualization | usability
Information Visualization can make hyper...

Interactive Intent Modeling for Text Search



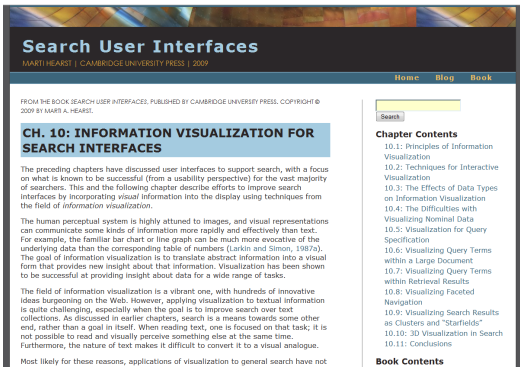


To Learn More, check out

Marti Hearst's Book

Chapter 10

http://searchuserinterfaces.com/book/

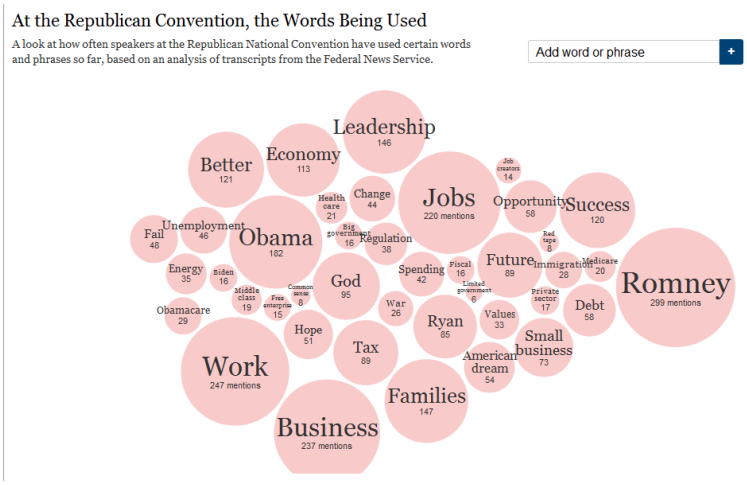


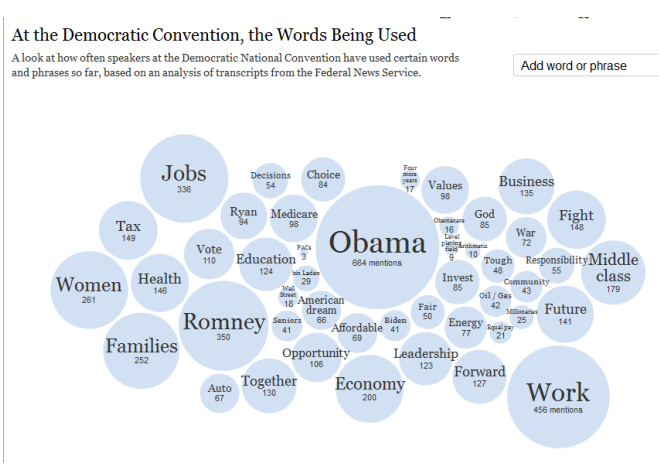
on to next category

- OK, let's move beyond just search/IR
- How do we represent the words, phrases, and sentences in a document or set of documents?
  - Main goal of understanding versus search

InfoVis for sentences, words, phrases

Word Counts





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

Or simply list them

Donald J. Trump

January 20, 2017 | 1459 words

dreams american obama jobs everyone

Barack Obama

January 21, 2013 | 2096 words

journey creed she complete requires

January 20, 2009 | 2395 words

nation new today generation jobs

George W. Bush

January 20, 2005 | 2071 words

freedom liberty tyranny fire history

January 20, 2001 | 1592 words

story civility affirm ideals promise

William J. Clinton

January 20, 1997 | 2155 words

century new promise land nation

January 20, 1993 | 1598 words

world today change idea season

https://qz.com/889611/the-most-frequently-used-words-in-donald-trumps-inauguration-speech-and-every-previous-one/

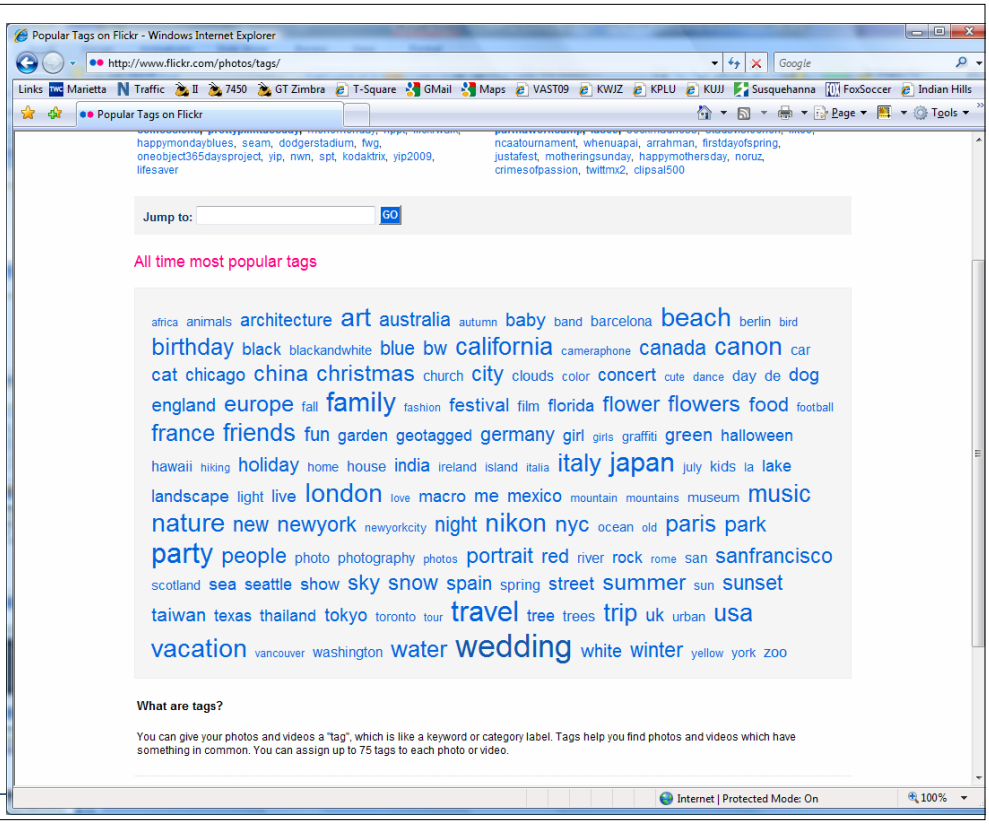
More Word Counting

http://www.wordcount.org

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

Flickr Tag Cloud



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

Wordle

http://www.wordle.net

Wordle

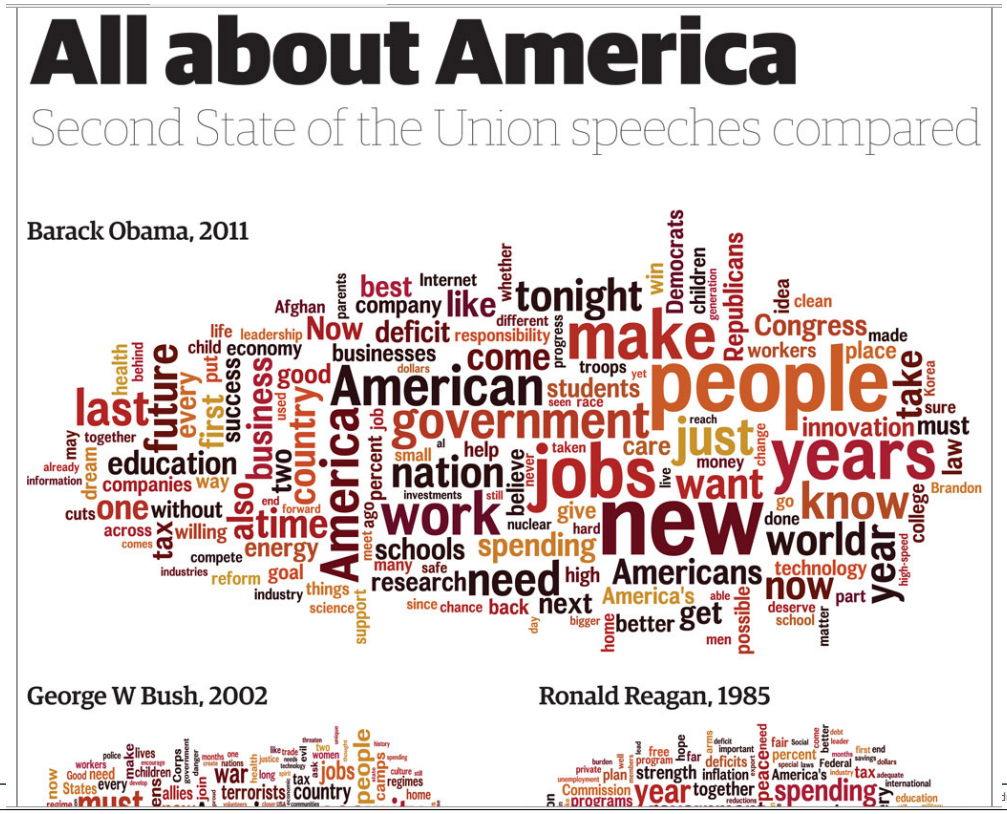
- Tightly packed words, sometimes vertical or diagonal
- Word size is linearly correlated with frequency (typically square root in cloud)
- Multiple color palettes
- User gets some control

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

Layout Algorithm

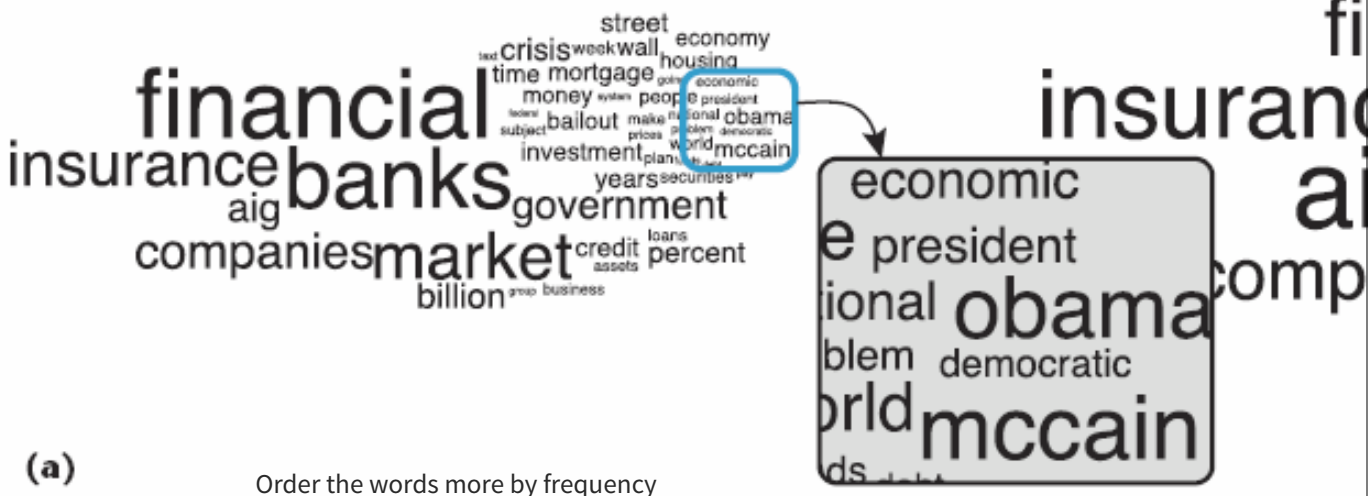
- Details not published
- Idea:
  - sort words by weight, decreasing order
    for each word w
        w.position := makeInitialPosition(w);
        while w intersects other words:
          updatePosition(w);
  - Init position randomly chosen according to distribution for target shape
  - Update position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

SoTU Wordles



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

A Little More Order



Order the words more by frequency
& put more similar words closer together

Cui et al
IEEE CG&A '10

also some fun uses of Wordles

- away messages
- Songs and poems
- Wedding vows
- Course syllabi
- Teaching writing
- Gifts
- art

Problems

- Actually not a great visualization. Why?

Problems

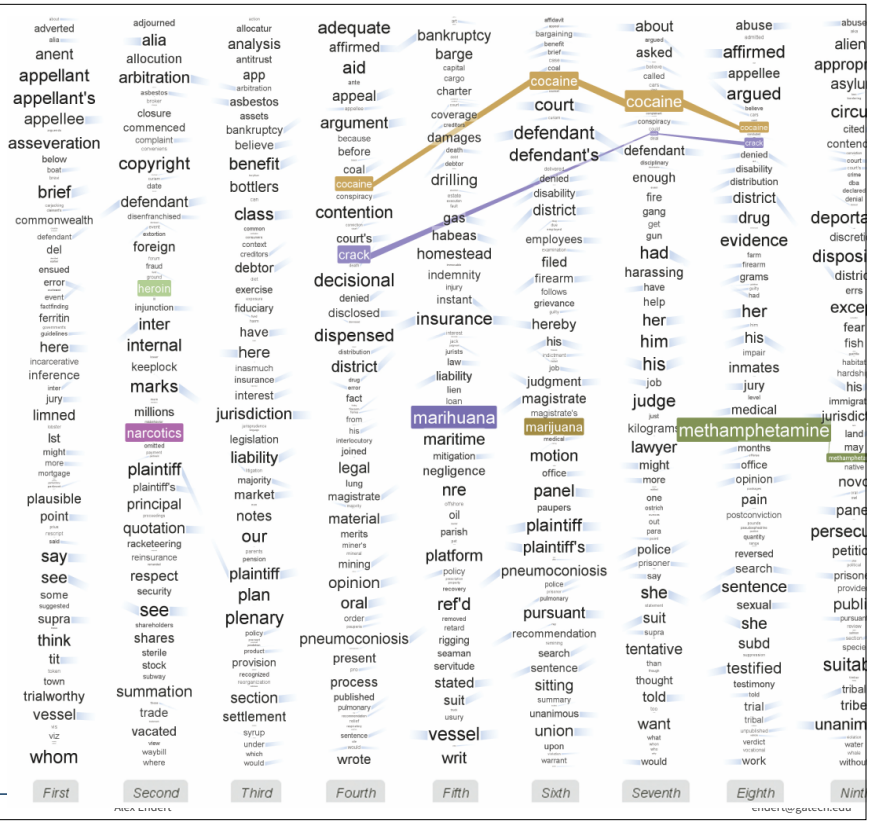
- Actually not a great visualization. Why?
  - Hard to find a particular word
  - *Long words get increased visual emphasis (and longer words are not "more important")*
  - Font **sizes are hard to compare**
  - Alphabetical ordering not ideal for many tasks
- Studies have even shown they underperform alternatives

Gruen et al
CHI '06

Multiple Documents?

- How to show word frequencies across multiple related documents?

Parallel Tag Clouds



Collins et al
VAST ‘09

Video

Different circuit courts

DocuBurst

Uses WordNet, sets of synonyms grouped together

Size – # of leaves in subtree
Hue – diff synsets of word
Shade – frequency of use

Collins et al
EuroVis ‘09

http://faculty.uoit.ca/collins/research/docuburst

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?

Hint: will learn about this during Visual Analytics Week

Words, Phrases, Sentences

Beyond Individual Words

- Can we show combinations of words, phrases, and sentences?

Word Tree

Stack
no visualization

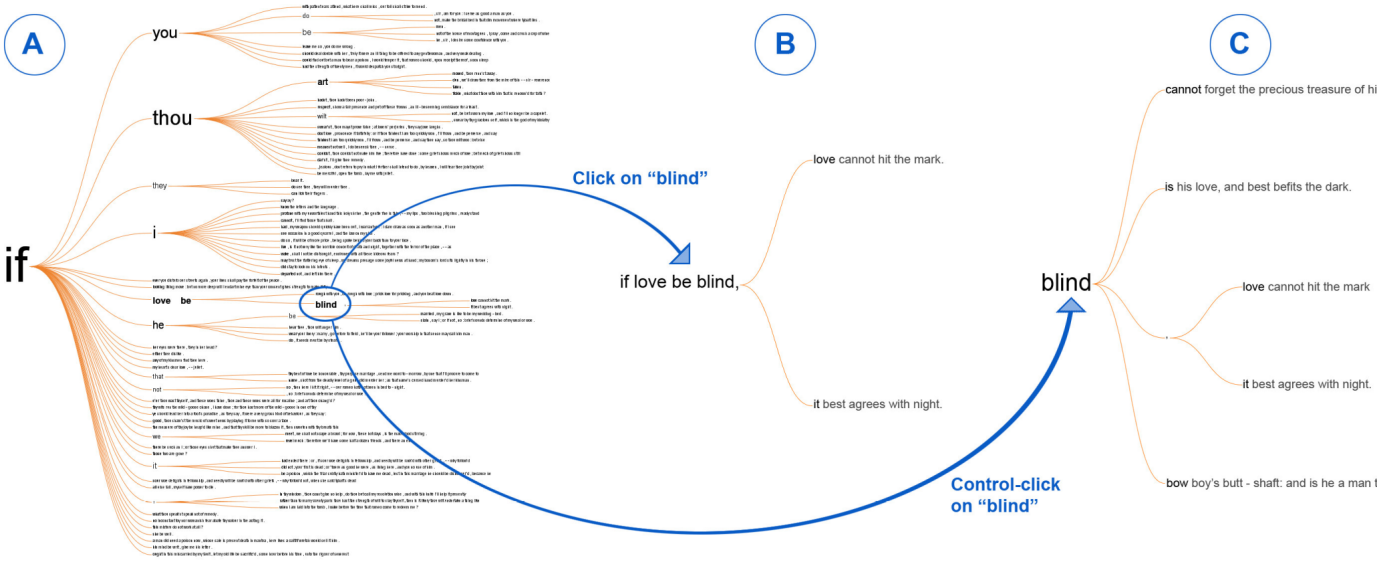
From King James Bible

Word Tree

- Shows context of a word or words
  - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus truncate the tree, only show that subtree begins from that word
- Ordered alphabetically, by frequency, or by first appearance

Wattenberg & Viégas
TVCG (InfoVis) '08

Interaction



Phrase Nets

Nets don't typically have one root. They can have cycles, edges going back and forth, etc.

- Examine unstructured text documents
- Presents pairs of terms from phrases such as
  - X and Y
  - X's Y
  - X at Y
  - X (is|are|was|were) Y
- Uses special graph layout algorithm with compression and simplification

van Ham et al
TVCG (InfoVis) '09

Examples

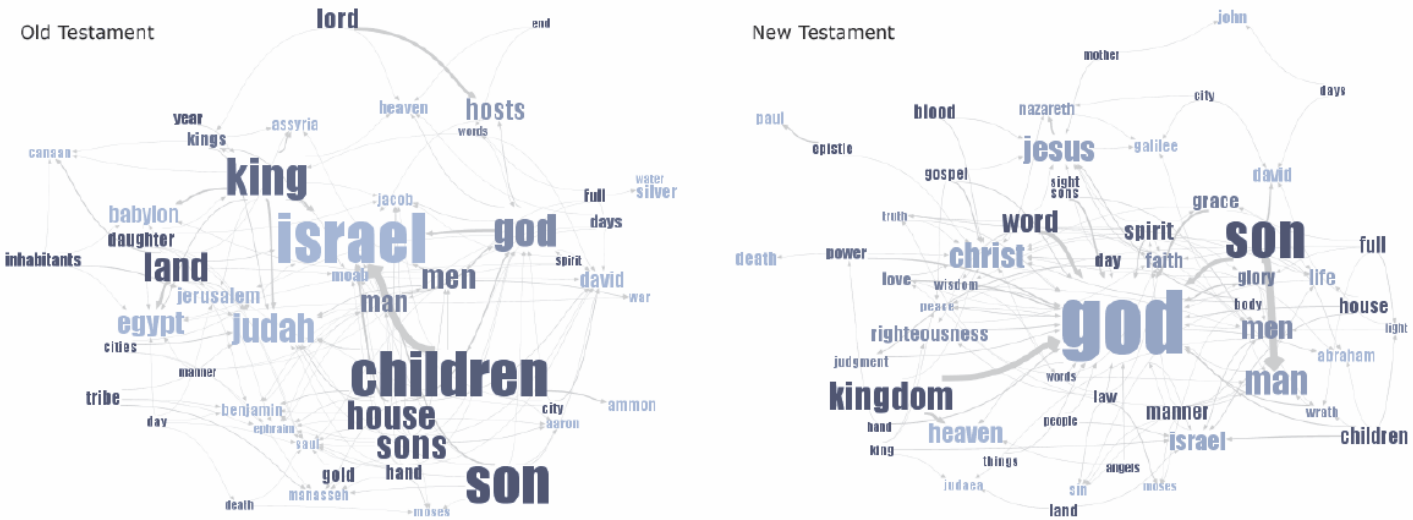


Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

SentenTree

Add grammatical structure to word cloud to give this more meaning.

Think of it as a word tree that goes toward both directions for every word

visualizing tweets for a given event

in this case, the Word Cup

combines word clouds (frequency of words) with sentence structure

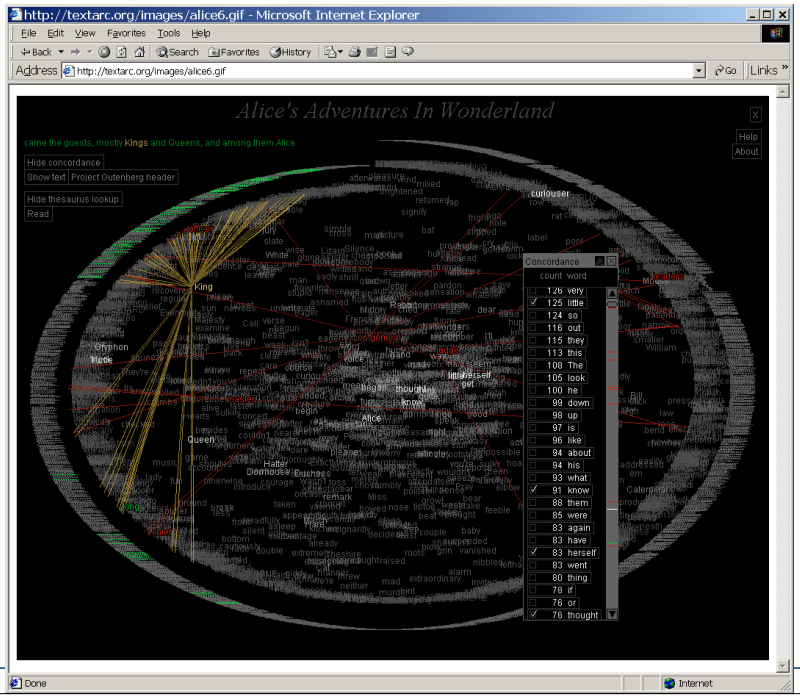
Hu et al., 2016

Another Challenge

- Visualize an entire book
- What does that mean?
  - Word appearances
  - Sentences
  - …

TextArc

http://textarc.org



Sentences laid out
in order of appearance

Words near to where
they appear

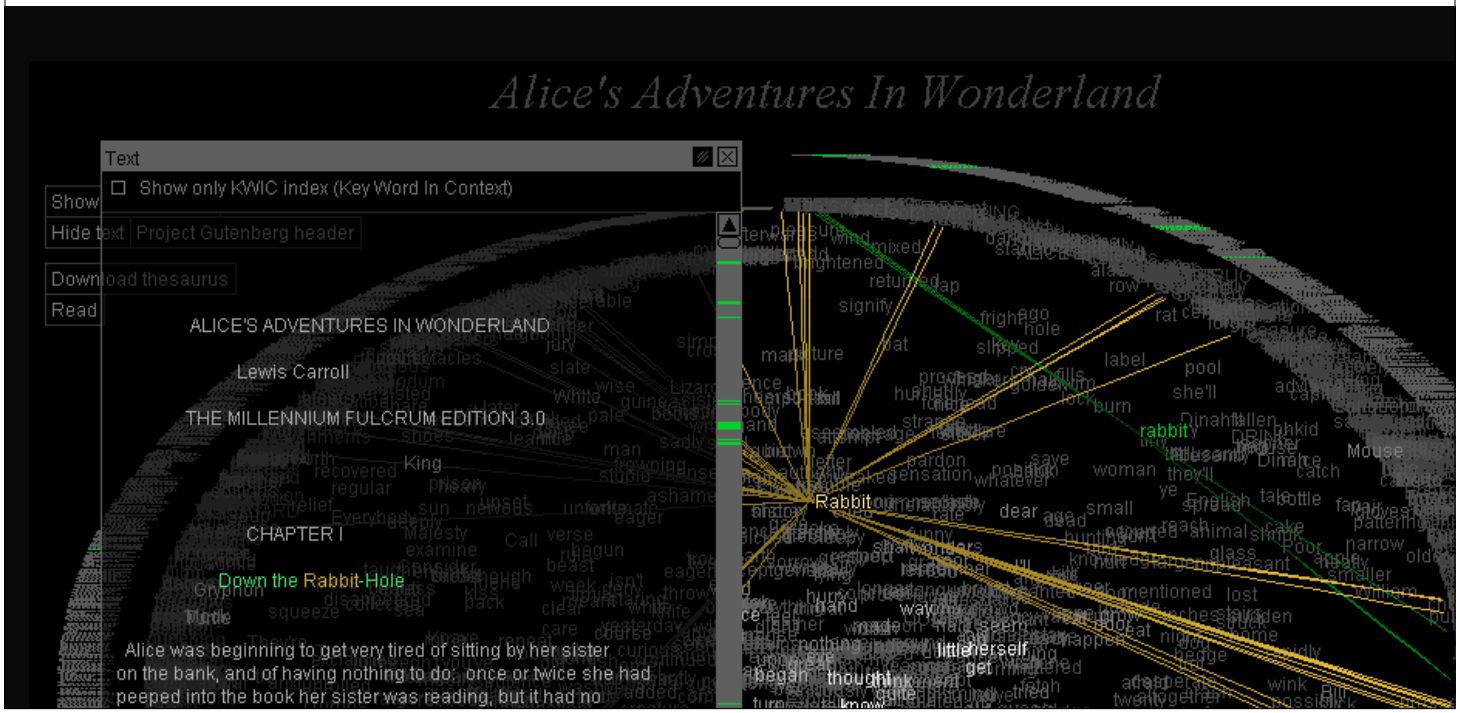
need a lot of interaction to make
sense of this one!

Brad Paley

endert@gatech.edu

TextArc

http://textarc.org



Let's try it

- Get into small groups (3-5 people)
- We will hand out a text document.
- Underline the entities (think about using only those that are "important")
- Then, sketch how you might visualize this data.
- Then, let's discuss.
  - What's hard about picking entities?
  - Do the entities give you a good idea of what the document is about?
  - Can you think of how you might visualize the entities?
- After, write each of your names and GT IDs on the back and hand to a TA.
  - counts as quiz for today

Next Time

- More about collections of documents and showing other characteristics of documents
  - Analysis metrics
  - Entities
  - Concepts & themes
  -

References

- Marti Hearst's i247 slides
- All referred to papers