Text and Document Visualization 1

CS 7450 - Information Visualization
October 19, 2015
John Stasko

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?

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.

Related Topic - IR

- Information Retrieval
  - Active search process that brings back particular/specific items (will discuss that some today, but not always focus)
  - I think InfoVis and HCI can help some…
- InfoVis, conversely, seems to be most useful when
  - Perhaps not sure precisely what you're looking for
  - More of a browsing task than a search one

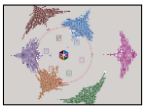
Related Topic - Sensemaking

- Sensemaking
  - Gaining a better understanding of the facts at hand in order to take some next steps
  - (Better definitions in VA lecture)
- InfoVis can help make a large document collection more understandable more rapidly

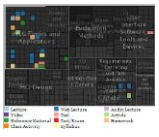
Challenge

- 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

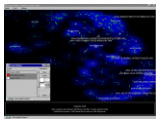
This Week's Agenda



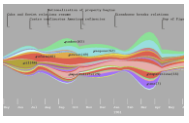
Visualization for IR
Helping search



Visualizing text
Showing words, phrases, and sentences



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



Information Retrieval

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

Generalize More

- How about the "holy grail" of a visual search engine?
  - Hot idea for a while
- My personal view: It's a mistake in the general case. Text is just better for this.

Search Visualization

http://www.kartoo.com

Defunct

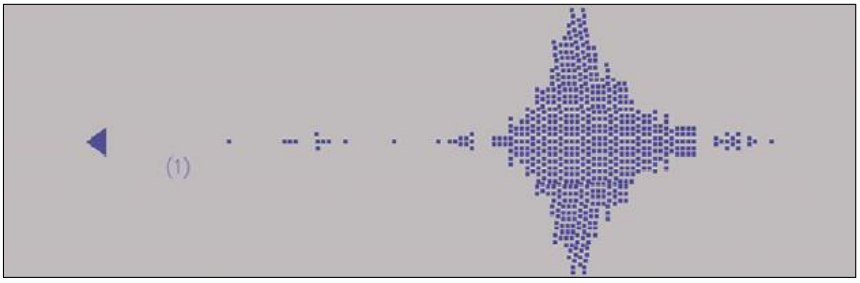
Sparkler

- Abstract result documents more
- 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

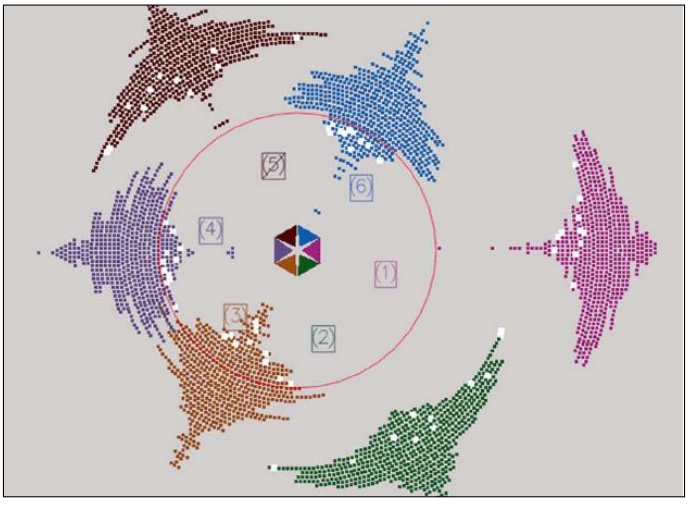
- Triangle – query
- Square – document
- Distance between query and documents represents their relevance



Visualizing Multiple Queries

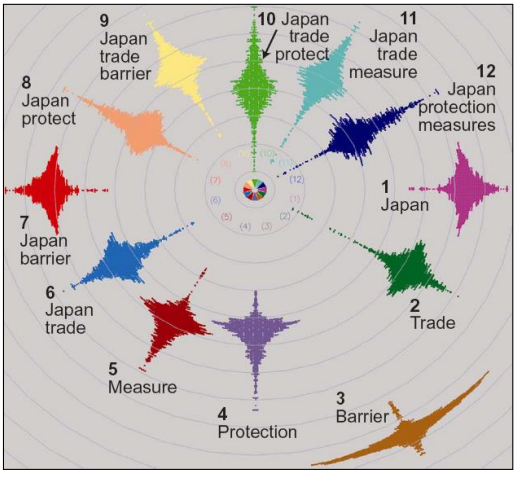
Six queries here

Bullseye allows viewer to select quality results

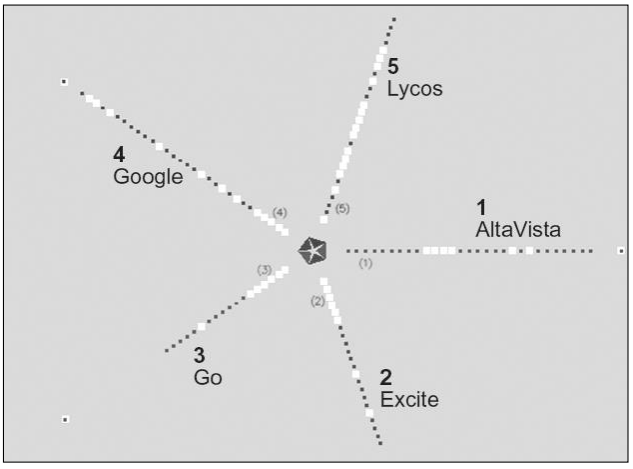


Test Example

- Text Retrieval Conference (TREC-3) 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

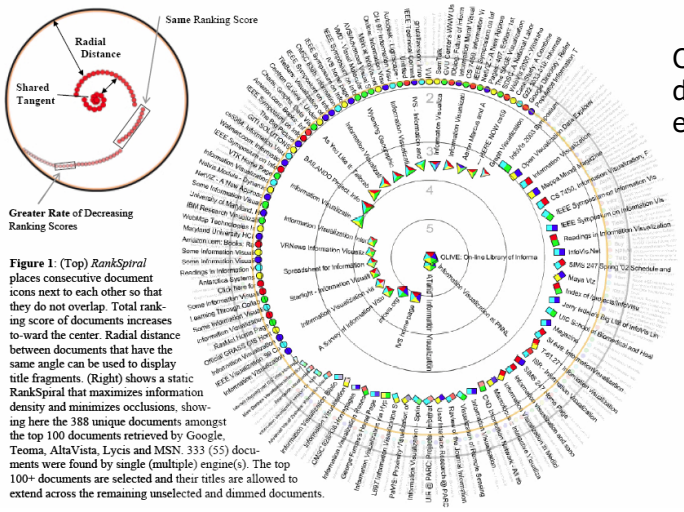
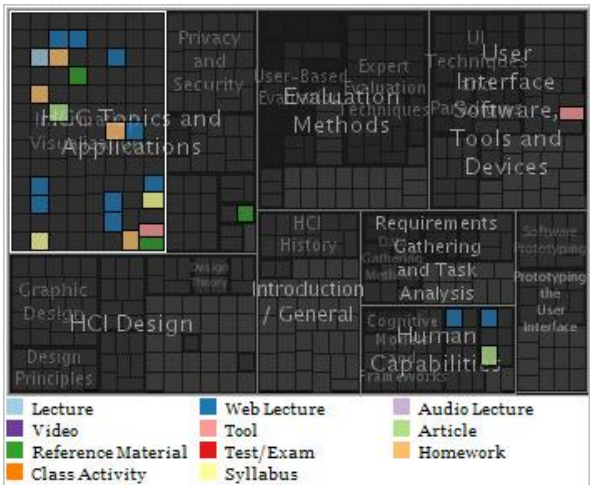


Figure 1: (Top) *RankSpiral* places consecutive document icons next to each other so that they do not overlap. Total ranking score of documents increases to-ward the center. Radial distance between documents that have the same angle can be used to display title fragments. (Right) shows a static RankSpiral 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

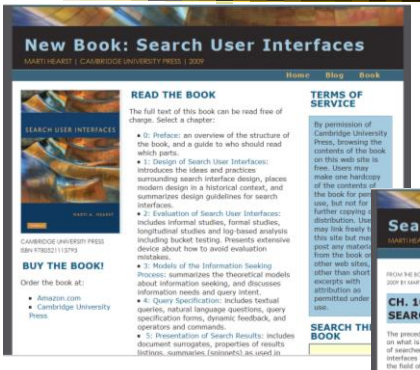
ResultMaps



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

Clarkson, Desai & Foley *TVCG* (InfoVis) '09

To Learn More



Marti Hearst's Book

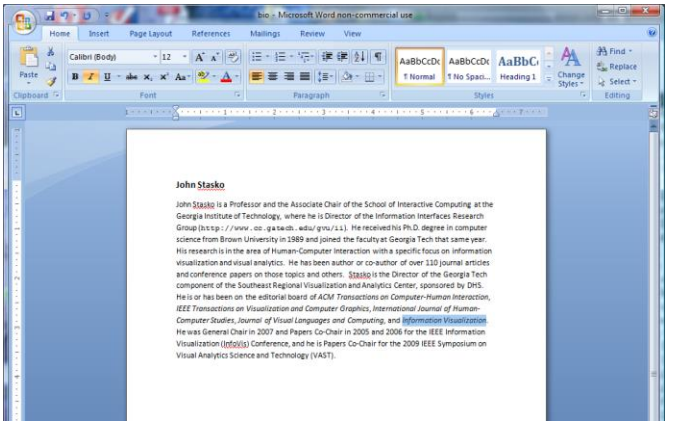
Chapter 10

http://searchuserinterfaces.com/book/

Transition 1

- OK, let's move up 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

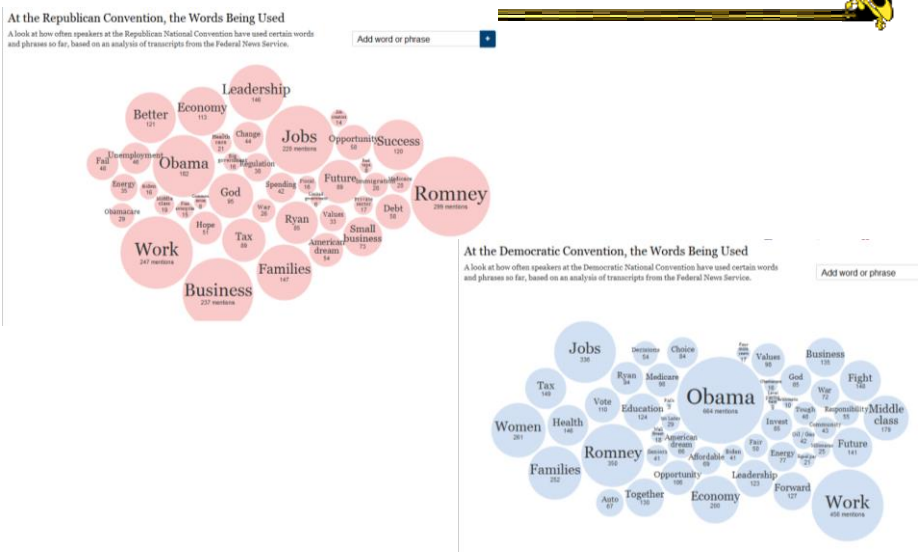
One Text Visualization



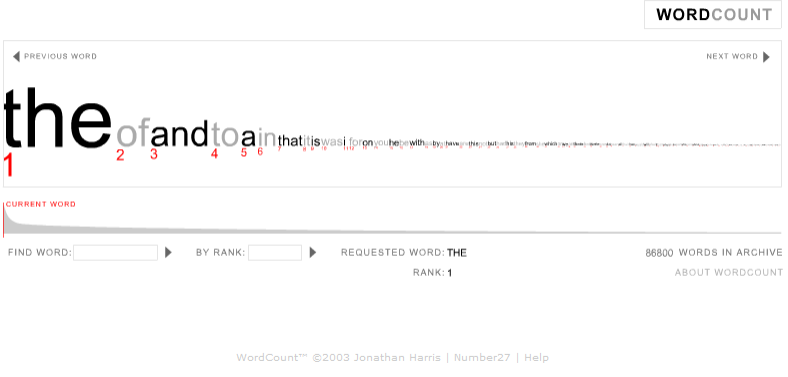
Uses:
Layout
Font
Style
Color
…

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

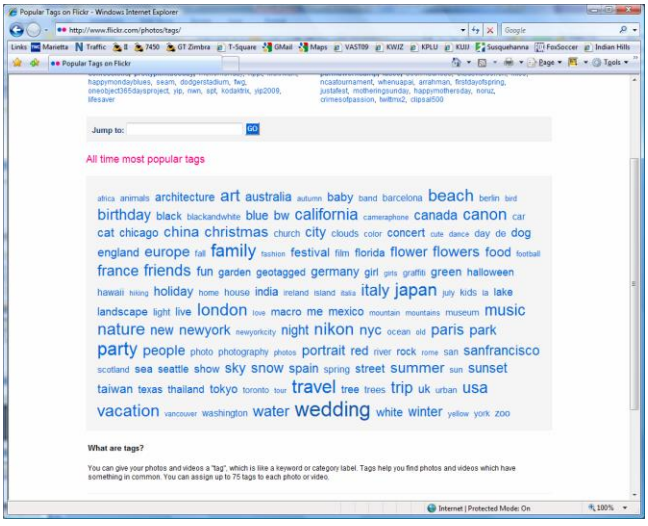
- Currently very "hot" in research community
- Have proven to be very popular on web
- Idea is to show word/concept importance through visual means
  - Tags: User-specified metadata (descriptors) about something
  - Sometimes generalized to just reflect word frequencies

History

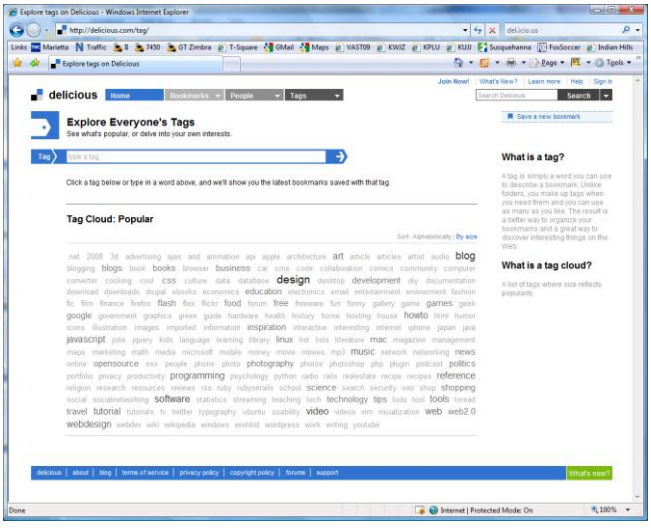
- 90-year old Soviet Constructivism
- Milgram's '76 experiment to have people label landmarks in Paris
- Flanagan's '97 "Search referral Zeitgeist"
- Fortune's '01 Money Makes the World Go Round

Viégas & Wattenberg
interactions '08

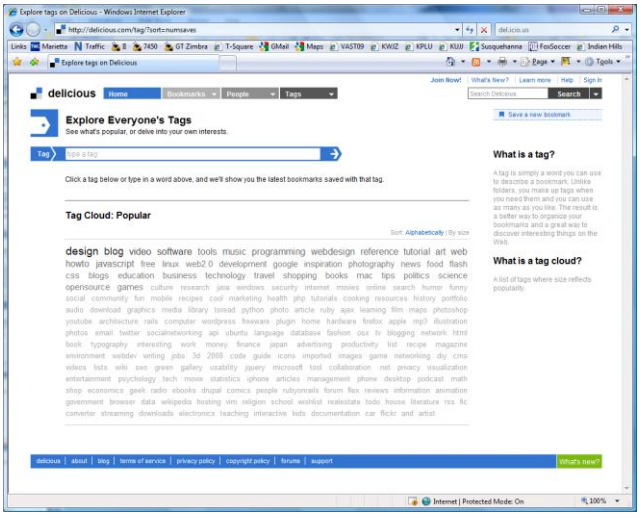
Flickr Tag Cloud



delicious Tag Cloud

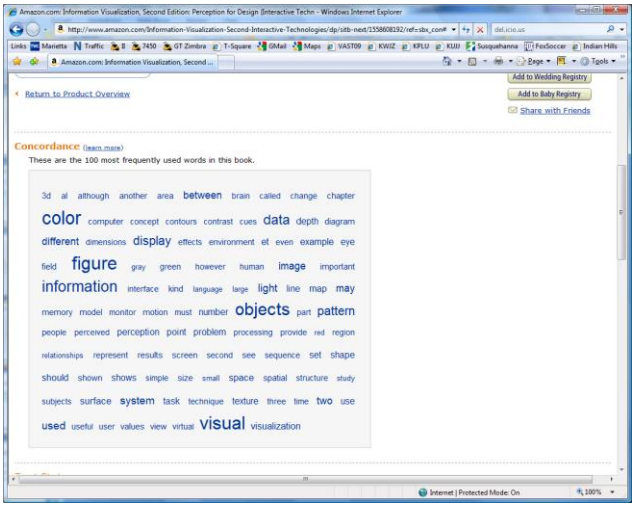


Alternate Order



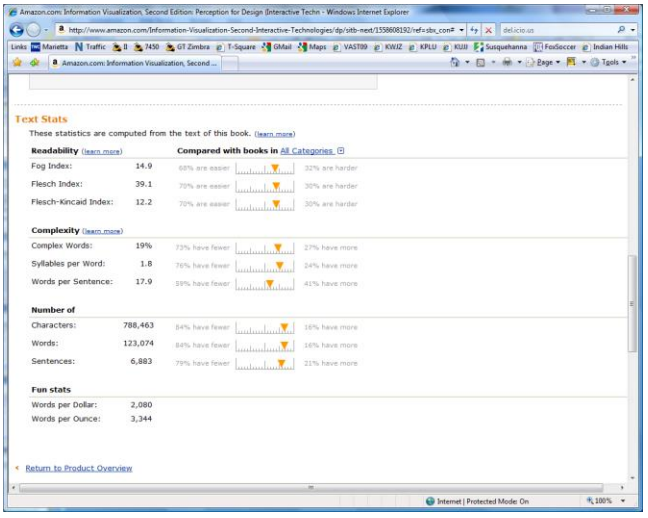
Amazon's Product Concordance

Maybe now a "word cloud"



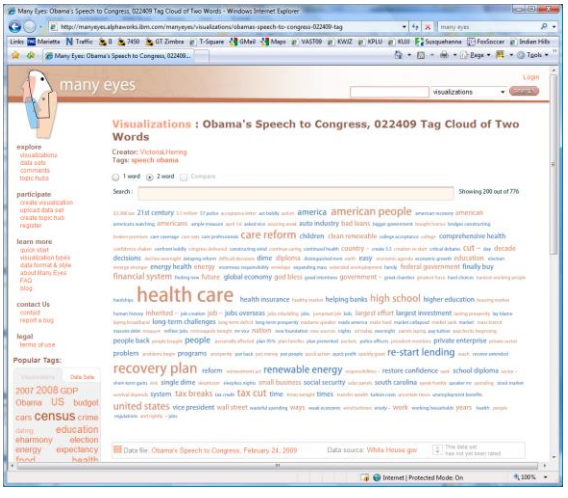
Sidenote

There are other types of info about a document on Amazon



Many Eyes Tag Cloud

Here, pairs of words are shown



Problems

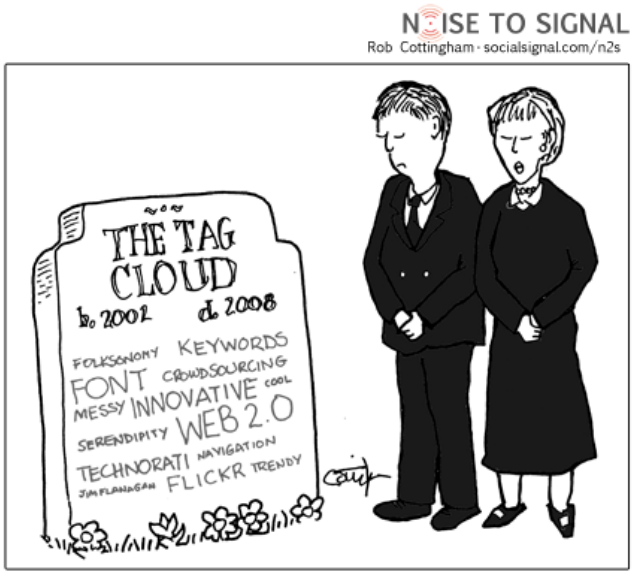
- Actually not a great visualization. Why?
  - Hard to find a particular word
  - Long words get increased visual emphasis
  - Font sizes are hard to compare
  - Alphabetical ordering not ideal for many tasks
- Studies have even shown they underperform

Gruen et al
CHI '06

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

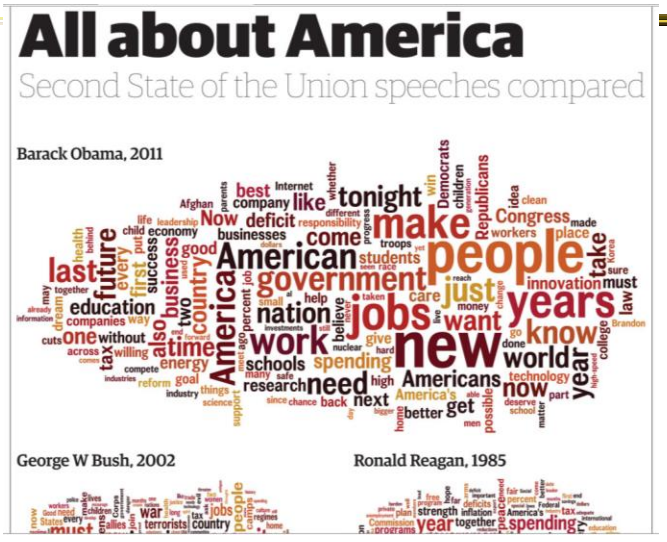
Fun Uses

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

2-day Survey in Jan. 09

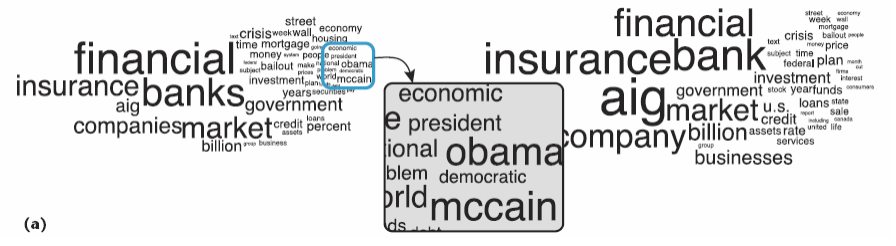
- 2/3 respondents were women
- Interest came from design, visual appeal, beauty
- Why preferred over word clouds:
  - Emotional impact
  - Attention-keeping visuals
  - Organic, non-linear
- Fair percentage didn't know what size signified

SoTU Wordles



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

A Little More Order



Order the words more by frequency

Cui et al
IEEE CG&A '10

Wordle Characteristics

- Layout, words are automatic
- If you had some control, what would you like to change or alter?

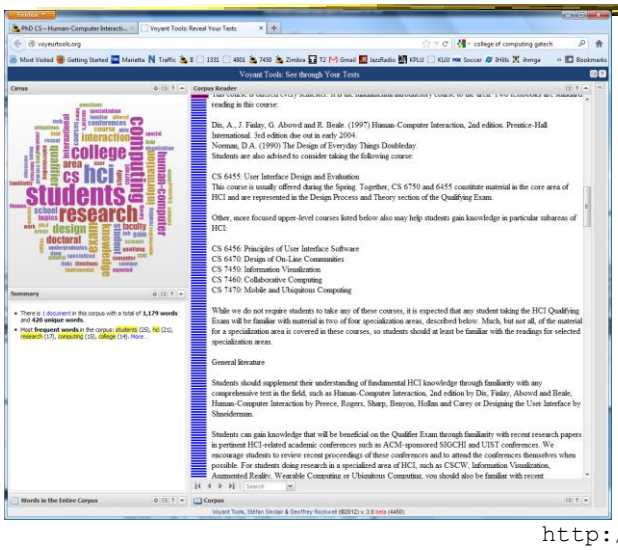
Mani-Wordle

- Start with nice default algorithm
- Give user more control over design
  - Alter color (within a palette)
  - Pin words, redo the rest
  - Move and rotate words
  - Smooth animation and collision detection for tracking changes

Koh et al
TVCG (InfoVis) '10

Video

Text Analysis on Web

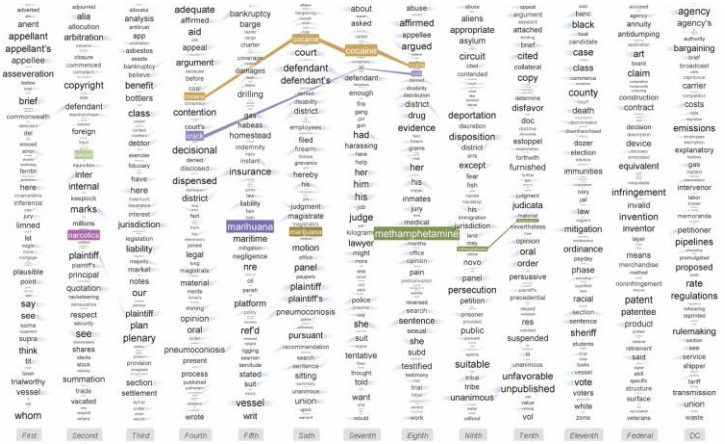


http://voyant-tools.org/

Multiple Documents?

- How to show word frequencies across multiple related documents?

Parallel Tag Clouds



Different circuit courts

Video

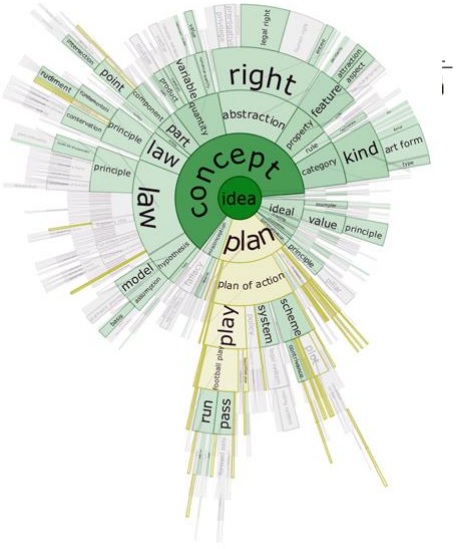
Collins et al
VAST `09

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?

DocuBurst

Uses WordNet, sets of synonyms grouped together

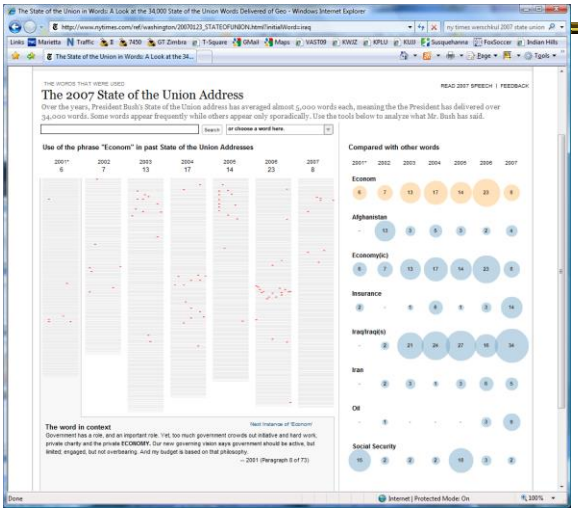


Size – # of leaves in subtree
Hue – diff synsets of word
Shade – frequency of use

Collins et al
EuroVis '09

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

Overview & Timeline

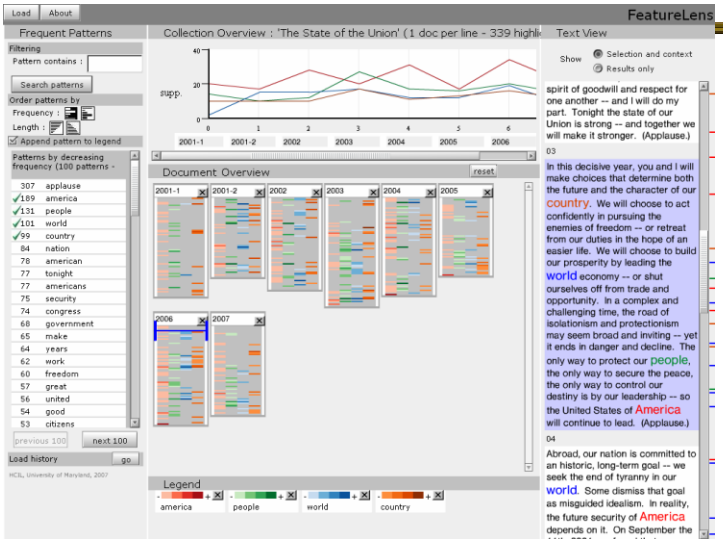


State of the Union Addresses

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

FeatureLens

Video

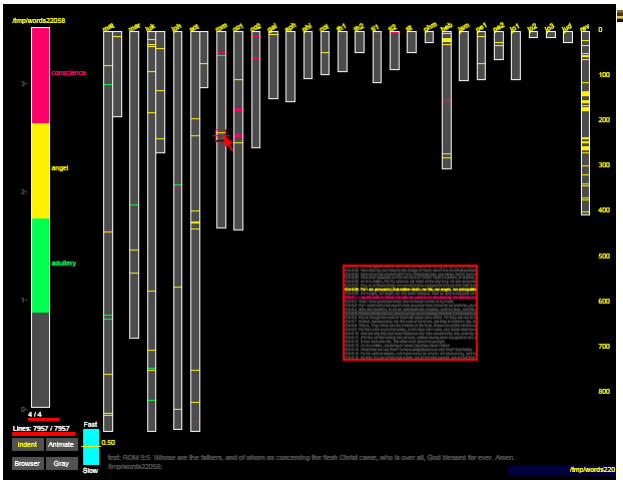


Show patterns of words or n-grams

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

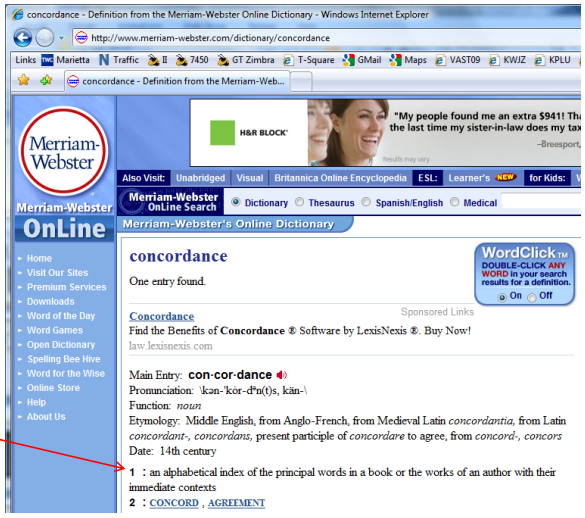
New Testament

Eick
Journal Comput. & Graph. Stats '94

Beyond Individual Words

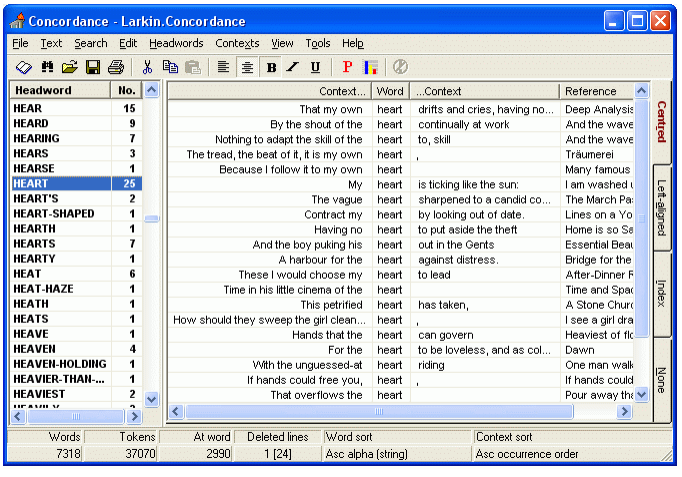
- Can we show combinations of words, phrases, and sentences?

Concordance



Definition

Concordance in Text



http://www.concordancesoftware.co.uk

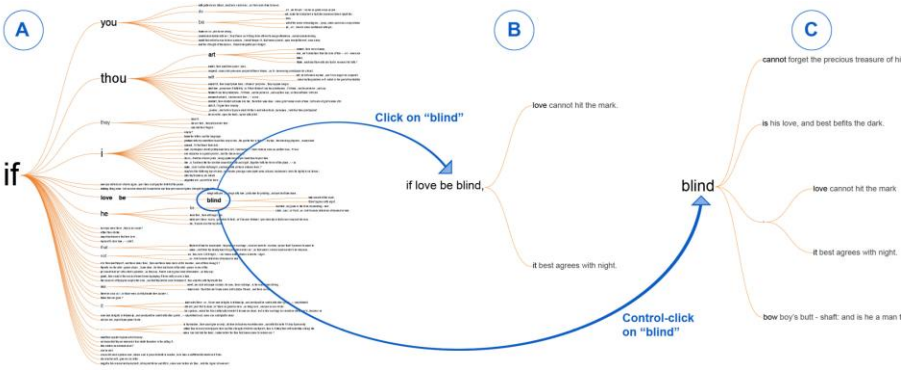
Word Tree

Word Tree

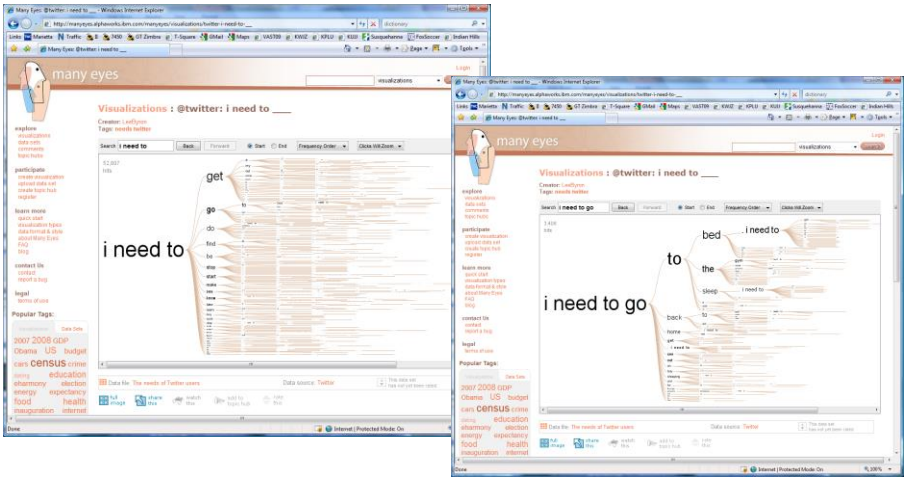
- Shows context of a word or words
  - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus
- Ordered alphabetically, by frequency, or by first appearance

Wattenberg & Viégas
TVCG (InfoVis) '08

Interaction



Many Eyes' WordTree



In Many Eyes now

Phrase Nets

- 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

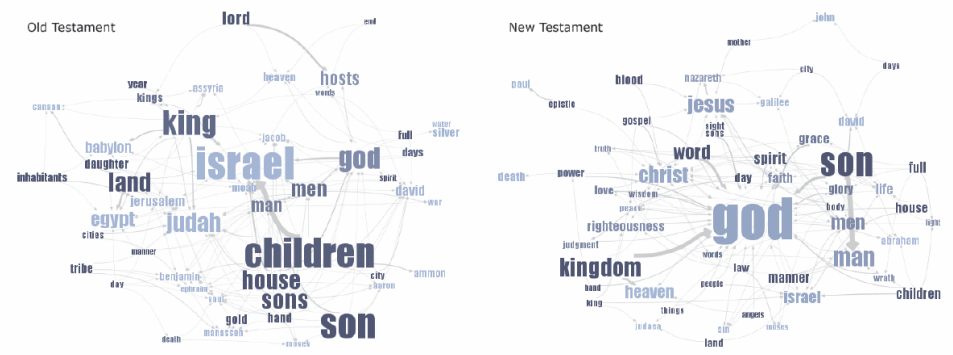


Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

Examples

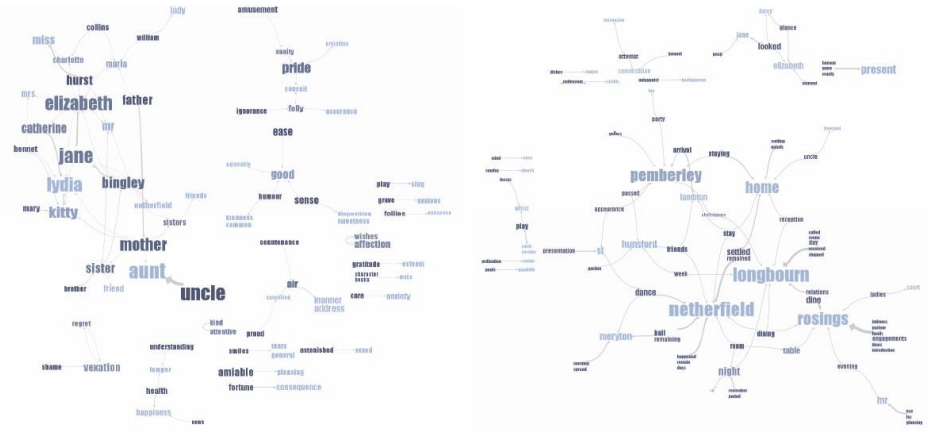


Fig 5. Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

User Interface

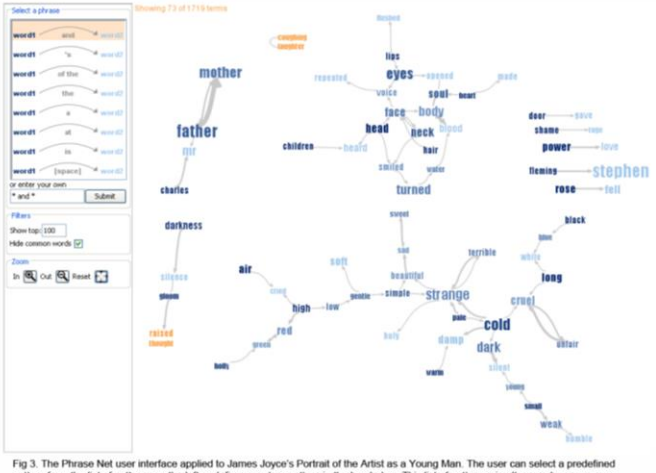


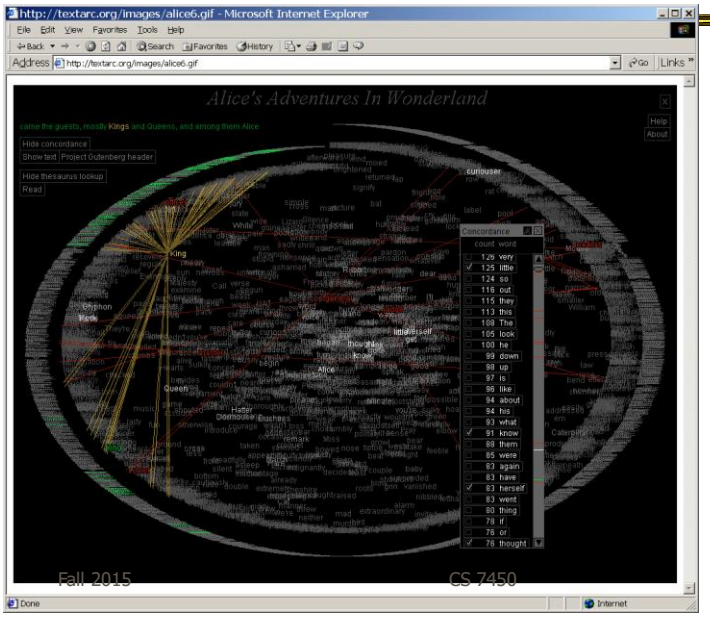
Fig 3. The Phrase Net user interface applied to James Joyce's Portrait of the Artist as a Young Man. The user can select a predefined pattern from the list of patterns on the left or define a custom pattern in the box below. This list of patterns simultaneously serves as a legend, a list of presets and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

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

Significant interaction

Brad Paley

Next Time

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

Project Design Document

- See project description & rubric
- My additions…
- Bring 2 copies
- Questions?

Survey Results

- Hours outside of class
  - <1,2,3-5,4,4,4,5,5,5,5,5,5,5-20,6,6,6,6-8,8,8,8*,10,10,10,11
  - *for useless meetings
- When in class, I feel
  - Interested (22/25)
  - Engaged (19/25)
  - …
  - Confused (2/25)
  - Bored (1/25)

Survey Results

- Pace?
  - About right – strong winner
  - Little too fast (4)

Survey Results

- Expectations clear?
  - Yes (15)
  - More rubrics (3)
  - "I really understand the lectures given by instructor but I am so cofused by assignments and final project. Assignemnts instruction are extremely unclear and even undifiend in some cases. I really do not know what it takes to succeed in this course."
  - "Not at all. Assignments decription are really bad. I really have no idea what should I work on? what should I write? how to evaluate tools? It is not also clear how TA are doing gradings. I talked to them couple of times they gave some reasons but the reasons are really might change person to person. The same thing is happening for the final project of this course nothing is clear. If I could I would drop the course."
- Subjectivity

Survey Results

- Good?
  - "Class demos are awesome! More hands on sessions which involve all the students in the class in group activities will be awesome. And make random group assignments for these group activities. This helps us know the other people in the class. I only know 1/3rd of the students in this class."
  - "The homeworks have been very helpful - particularly homeworks 3 and 4. D3 is difficult, but I think it is a very valuable skill to have. I also think critiquing commercial tools was a very useful exercise to think critically about the visualization decisions made. I have also really loved seeing the short demo videos. That's interesting and inspiring."
  - "while it can be time consuming to read articles for each class, the material is very useful going into lecture. I am not certain how useful I found HW4 to be... while it was interesting to critique existing tools and systems, I would have liked to have more homework on how to design clear and effective data visualizations or perhaps spending more time figuring out what types of visualizations to use with certain data sets."

Survey Results

- Bad?
  - "I hope that the questions asked to the class could be a little more difficult or thought-provoking, maybe more open-ended if that makes sense? / it's always nice to bring some kind of controversy and discussion into the issues presented, although maybe that's not best for time management."
  - "This class needs more short group projects and short group activities in class with random group formation. This will keep things fresh and the students will take more interest in the course."
  - "Perhaps the class can be offered every sem resulting in smaller classes. Helps with having more in class activities like design critiques, discussing visualizations etc."
  - "practice with what types of visualizations to use to communicate data sets and intent."
  - "more assignments which make me build visualizations"
  - "Off the top of my head, maybe a bit more difficult assignments. But I understand that many in the class don't have programming experience and it might be too difficult for them."
  - "More examples of past course projects. What teams did well, what types of viz's they used correctly and incorrectly. Examples are always useful."

Survey Results

- Bad?
  - "The group project + homeworks + readings + final exam is too much for me and others I have talked to. If you have a big group project, I feel an exam in unnecessary."
  - "The course is heavy. A lot of things going on at the same time from readings, projects, hw's. It is difficult to get time to focus well on one area(eg the project). I feel there should just be an end project instead of a final exam like all of my other classes this semester."
  - "Make the whole class homework and project based and more programming related. The final exam is a hindrance that must be done away with in this day and age."
  - "My only criticism so far is the one reading assignment that was chapters 4-12 of the textbook. That was a bit too much at once that the information didn't "sink in" as much as it would have if it had been a bit more spread out (into two separate reading assignments, for instance). It was a bit overwhelming. Other than that, I have no criticisms."

Survey Results

- Bad?
  - "At the begining of the semester the instructor mentioned that this course is going to be lots of work. I said okay... I would challenge myself and learn many things. yes this course needs lots of time in a useless way. I ended up spending most of my time meeting TAs and my groupmates for this course and talking what would please the instructor and what would not. This has several reasons that I am going to explain it as bellow: First: Instructor needs to clarify his expectations for every homework or project in a very clear way. For example, in many courses system that we develop needs to have a certain functionalities. So that We try to develop that and we are clear that function must work otherwise we are not getting our grade. This is the first time that I took a class and definitation of everything is summarized in a single word "interesting". Interesting design, interesting dataset. Second: Instructor mentioned the life is unfair at the begining of the semester but he makes it even more unfair. Above 70% of the class are master HCI master students with the minimum knowledge of programming. So in our group we have three designers and two programmers. Designers all have different definitation of "interestinf design", and "interesting dataset". They everytime end up coming up with some imaginary designs and I need to explain them how difficult is to implement their design. What many professors in CS department are doing is that they will ask students to fill up a form explaining their skills and then the instrcutor and his TAs will distripute the groups in a way which is fair to everyone."

Survey Results

- What else could you do?
  - More reading
  - More practice
  - Meet with TAs/instructor more

Survey Results

- Additional comments?
  - "I do not like that there is a final exam for this type of class. This is more of a group project based class and it doesn't make much sense to have 0 tests/quizzes and then have a final at the end. Very difficult to know what to expect for an exam that is 20% of the grade."
  - "In a grad school class, I feel final exams are unnecessary when we are all trying to put a lot of work into a final project. It is archaic."
  - "The amount of papers, books and articles that this class expects the students to read on their own is staggering. Students have other difficult classes on their plate and cannot read all the papers and recommended readings. Efforts must be done to summarize those required readings so that students can read the summaries to save time."

Survey Results

- Additional comments?
  - "I hope that there will be some mechanism for reviewing group members at the end of the term project. Perhaps this was already the plan, but if not, I hope it will be possible. It becomes apparent early on which group members care about the class, the project, and their grades, and which ones can't be bothered to make an effort. I hope there will be some way to reflect on the contributions on other group members, so that those who have helped hold the group together and those who haven't done their fair share can be graded accordingly."

Upcoming

- Text and Documents 2
  - Reading
    Keim & Oelke '07
- Time Series Data
  - Reading
    Aigner et al '08

References

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