Text and Document Visualization

BAIS 6140 – Information Visualization

L. Miguel Encarnação

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?

Example questions, tasks & goals

- Questions
  - 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?
- Tasks & Goals
  - Identify “hidden” messages or stories in this document collection.
  - Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
  - Find connections between documents.

Related topic – Information Retrieval

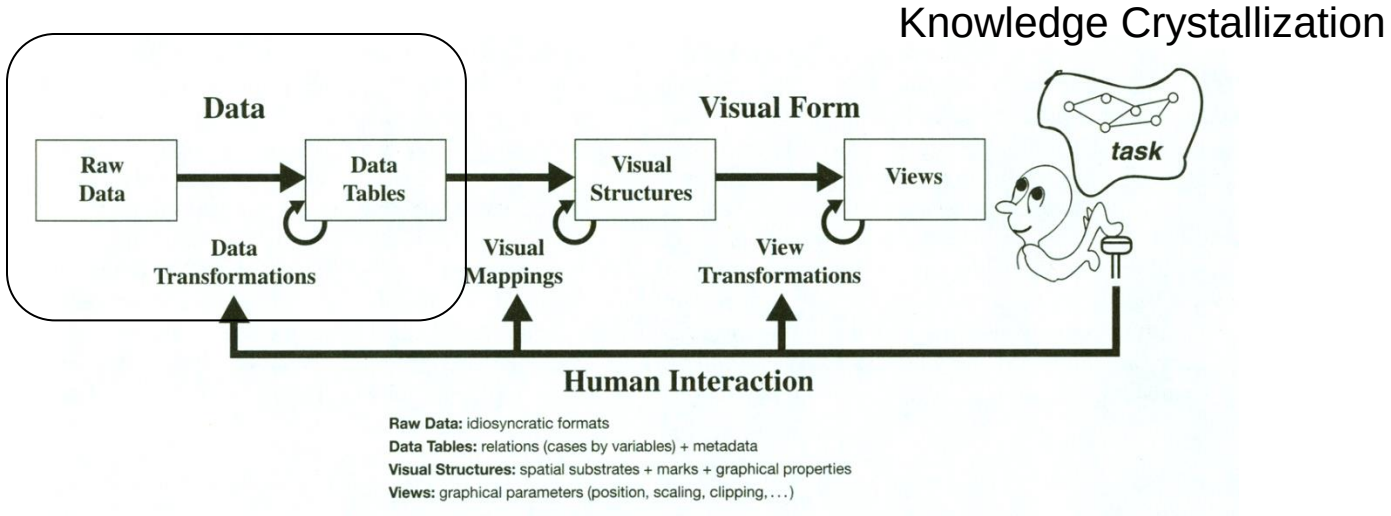
- Information Retrieval
  - Active search process that brings back particular/specific items (will discuss that some today, but not always focus)
  - InfoVis and HCI can help
- InfoVis seems to help most when
  - Perhaps not sure precisely what you're looking for
  - More of a browsing task

Related topics – Sensemaking

- Sensemaking
  - Gaining a better understanding of the facts at hand in order to take some next steps
- 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



Challenge: Text is unstructured

Structured

	Name	Age	Gender
Case 1	Johnny	18	M
Case 2	Lisa	22	F
Case 3	M	19	N

Variables or Data Sets?

Text

	Text 1	Text 2
Case ?	The customer likes the picture of a sewing woman.	The boy loves the girl like crazy.

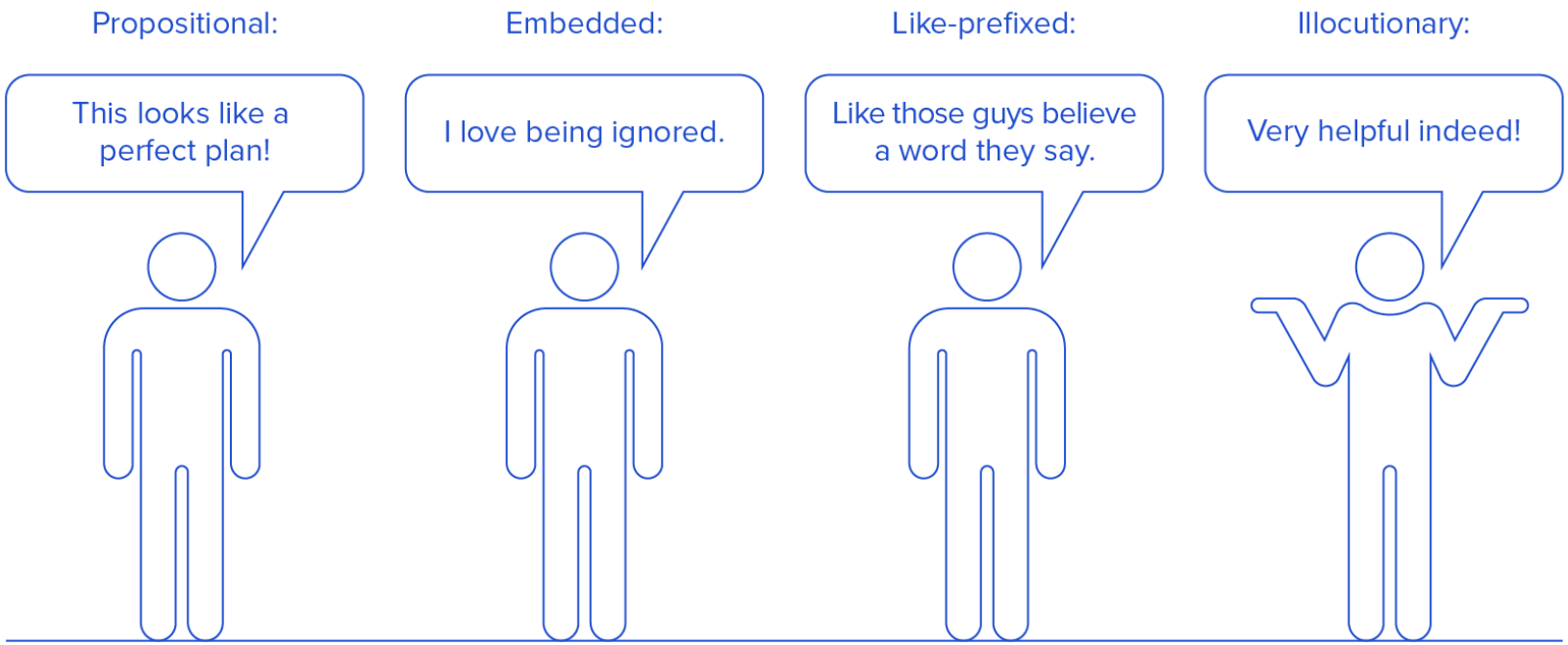
Text is unstructured

	Text 1	Text 2	Entity of Analysis
Case ?	The customer likes the picture of a sewing woman.	The boy loves the girl like crazy.	Keywords
	The customer likes the picture of a sewing woman.	The boy loves the girl like crazy.	Analogies
	The customer likes the picture of a sewing woman.	The boy loves the girl like crazy.	Phrases
	The customer likes the picture of a sewing woman.	The boy loves the girl like crazy.	Sentiments
	The customer likes the picture of a sewing woman on Facebook.	The boy loves the girl on Facebook like crazy.	Semantics (Meaning)

More semantic challenges

Example	Challenge
This guy is crazy!	Lexical Ambiguity (word)
Visiting relatives can be boring.	Syntactic Ambiguity (sentence)
You really outdid yourself!	Irony & Sarcasm
Te luciste!	Language
You are doing a good job.	Culture / Context

More Sarcasm



https://www.toptal.com/deep-learning/4-sentiment-analysis-accuracy-traps

Today's Agenda

- Micro-level: More emphasis on individual words, actual document contents

- Macro-level: Emphasis on large document collections, themes and concepts across collection, how documents relate

Documents

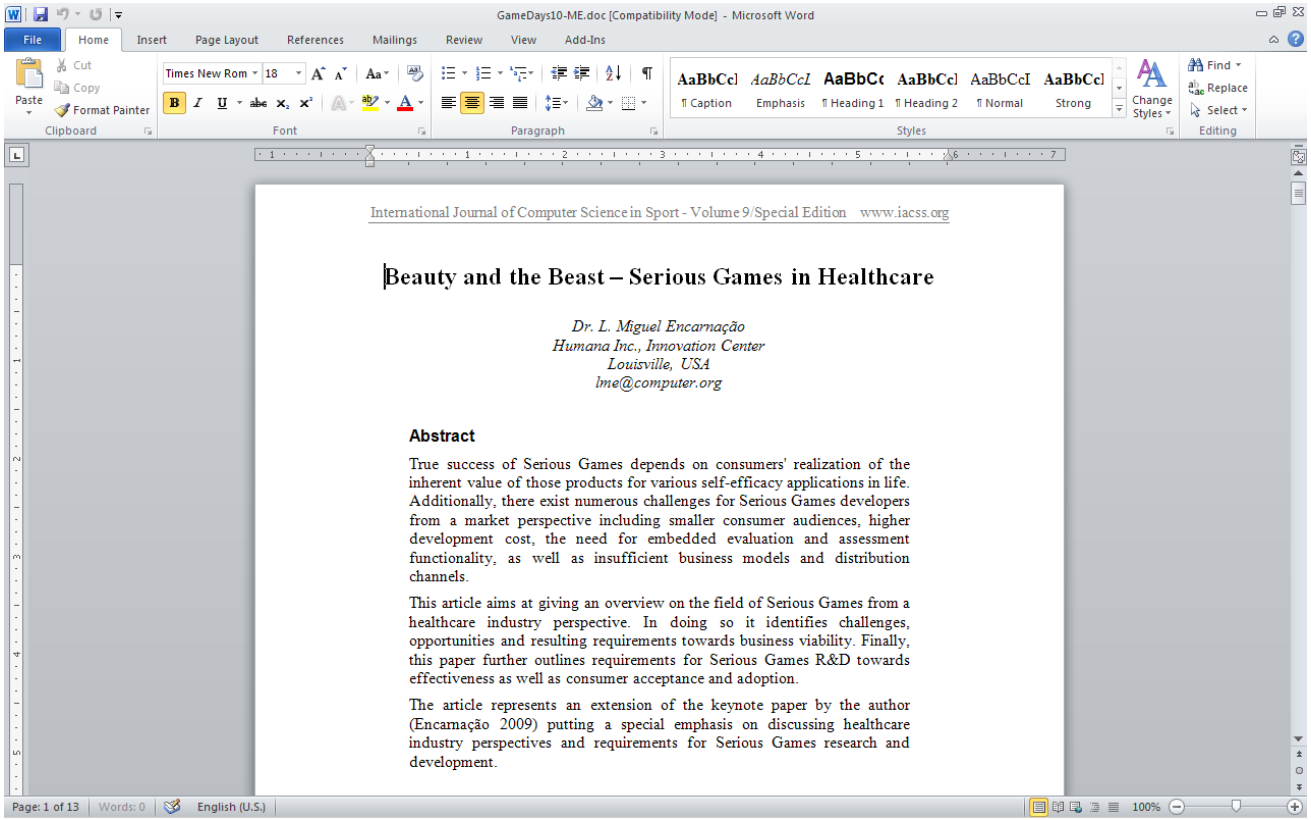
- Collections of text
- Structure imposes certain author-defined role/meaning on text in overall document
  - Title
  - Abstract
  - Pre-amble
  - Chapter
  - Section
  - Paragraph

Document collections

- Sets of documents with or without contextual relationship
  - Library of Congress
  - Editions of a book
  - Genre
  - Works of a particular author
  - Program code
  - Wikipedia
  - …

One Text Visualization

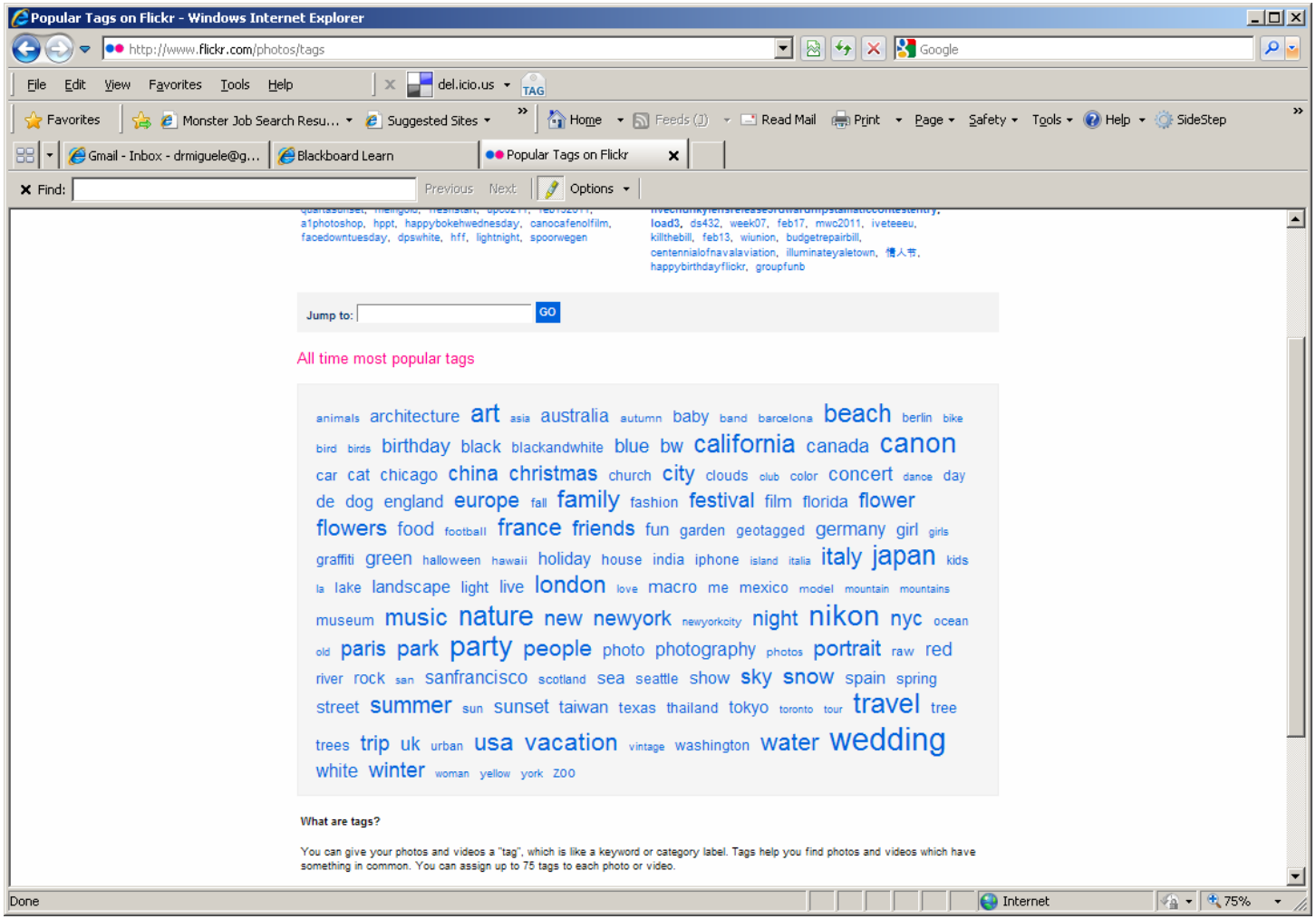
- Uses:
  - Layout
  - Font
  - Style
  - Color
  - …



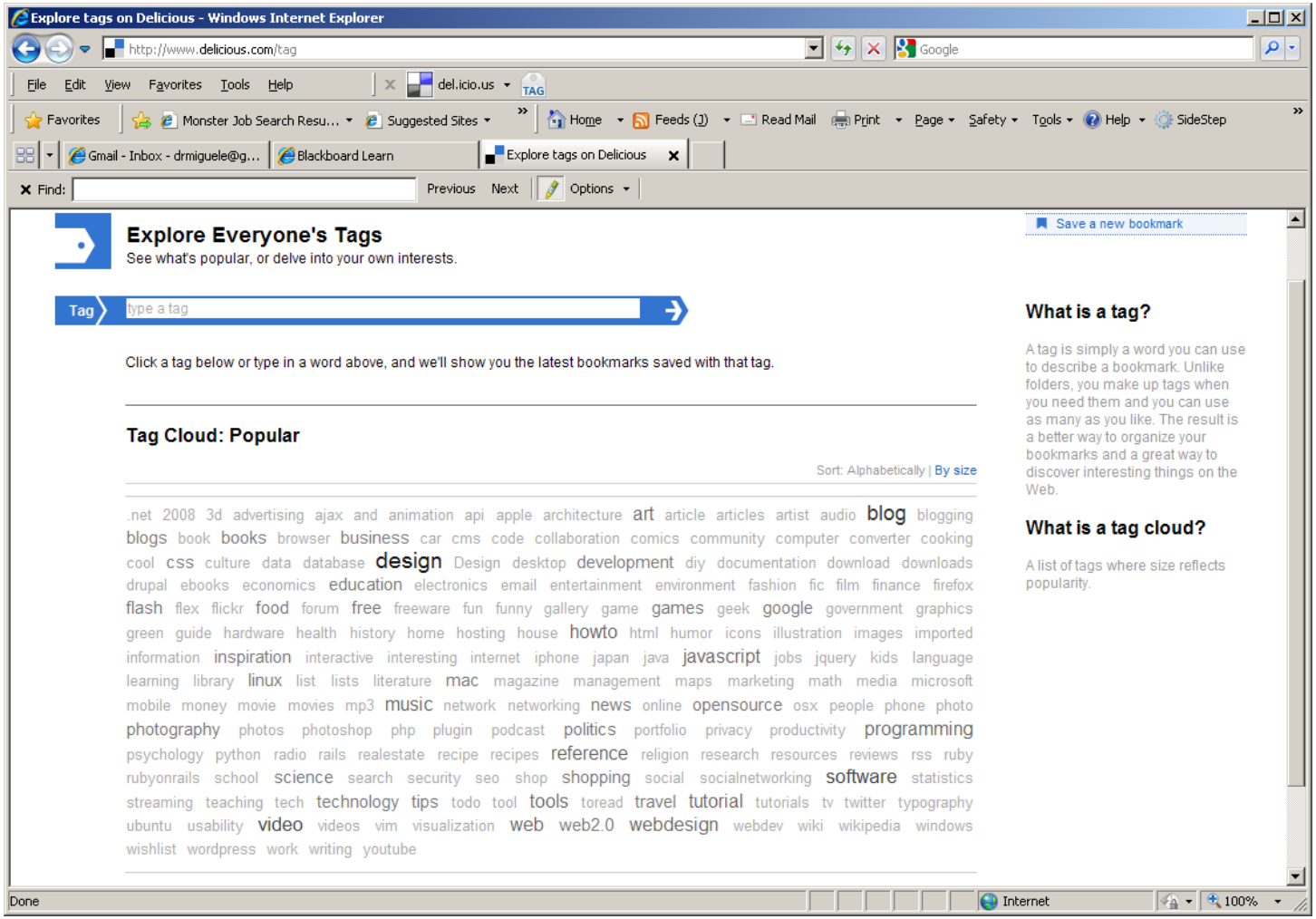
Tag/Word Clouds

- 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

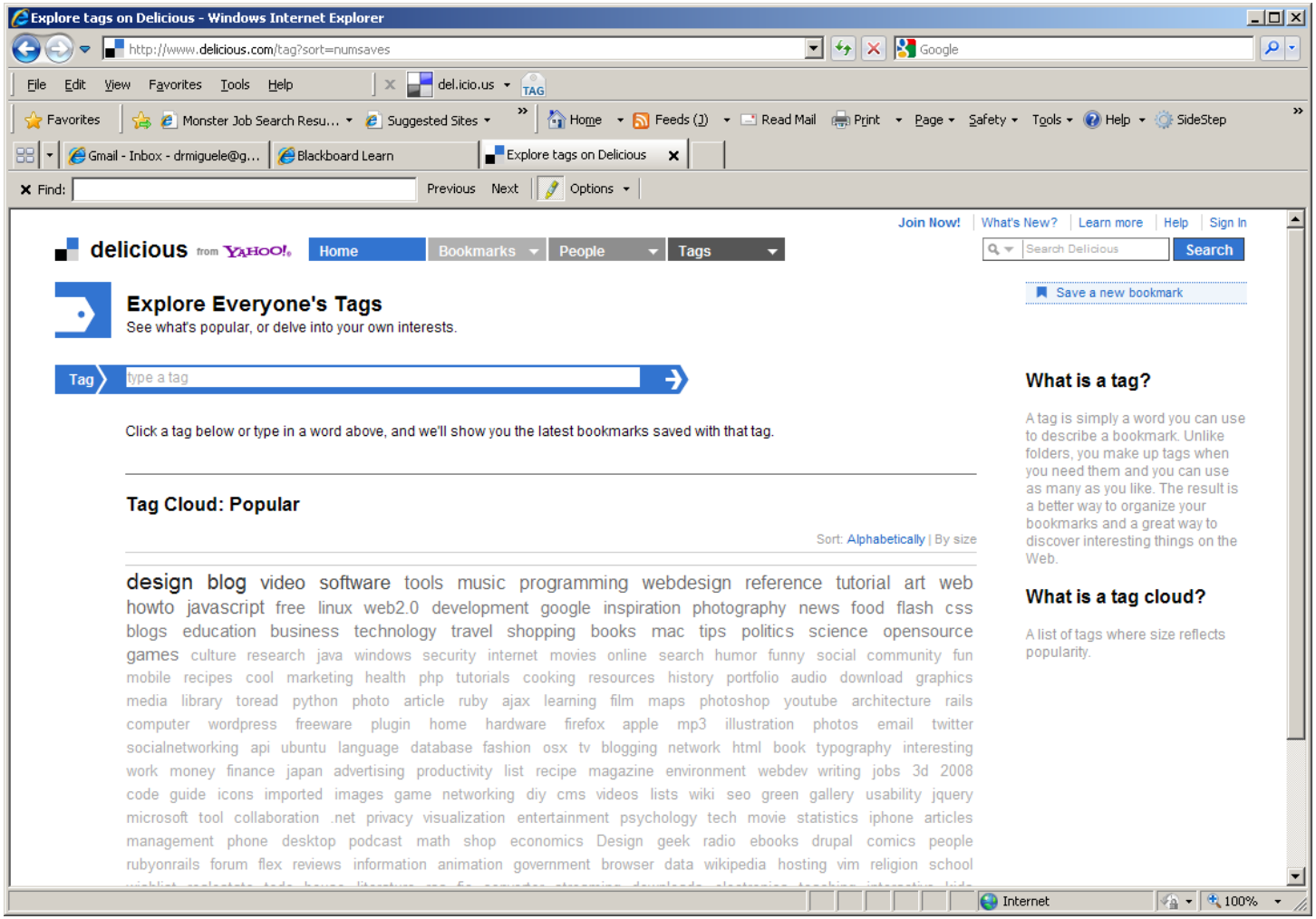
Flickr tag cloud



Delicious tag cloud

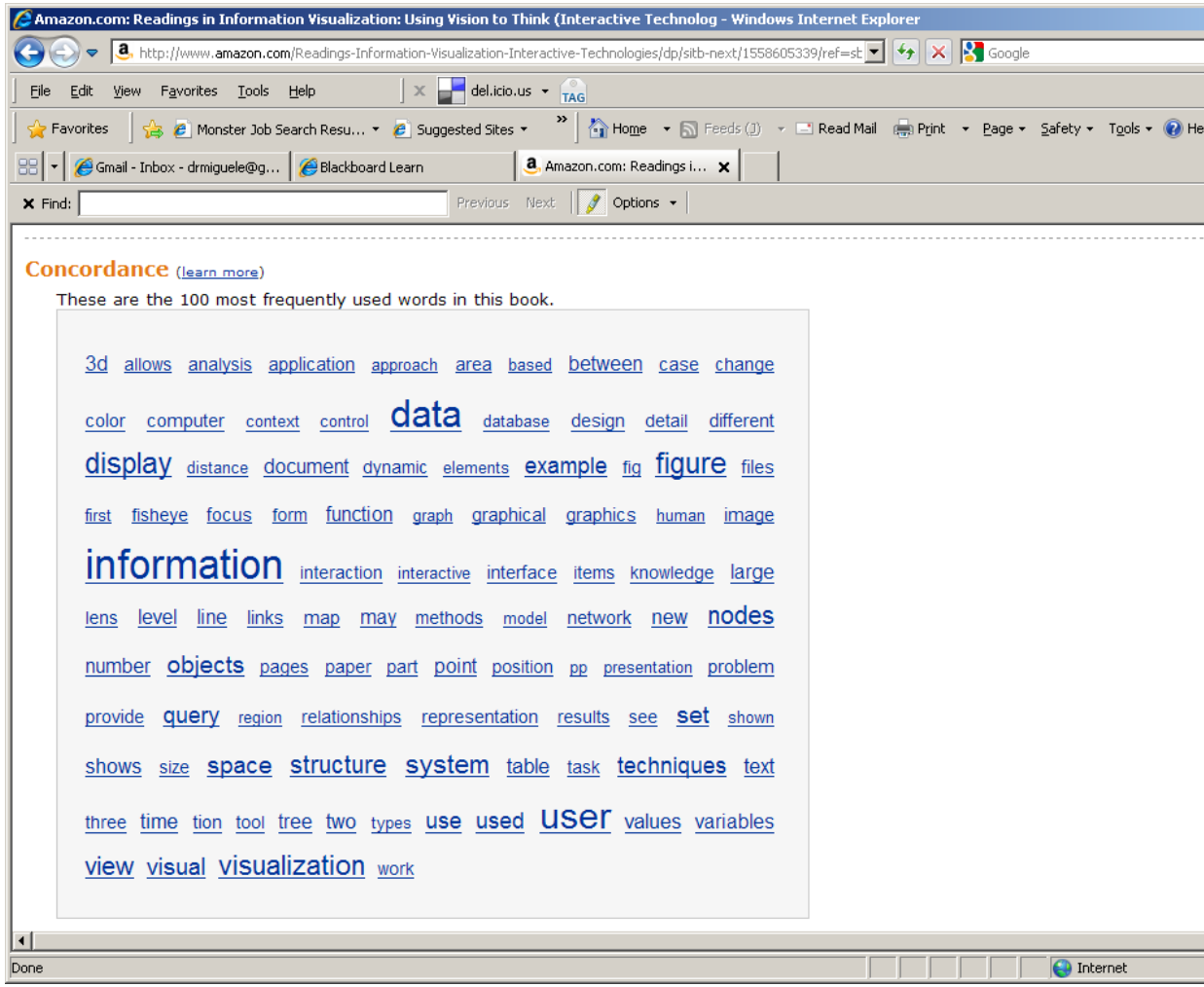


Alternate order

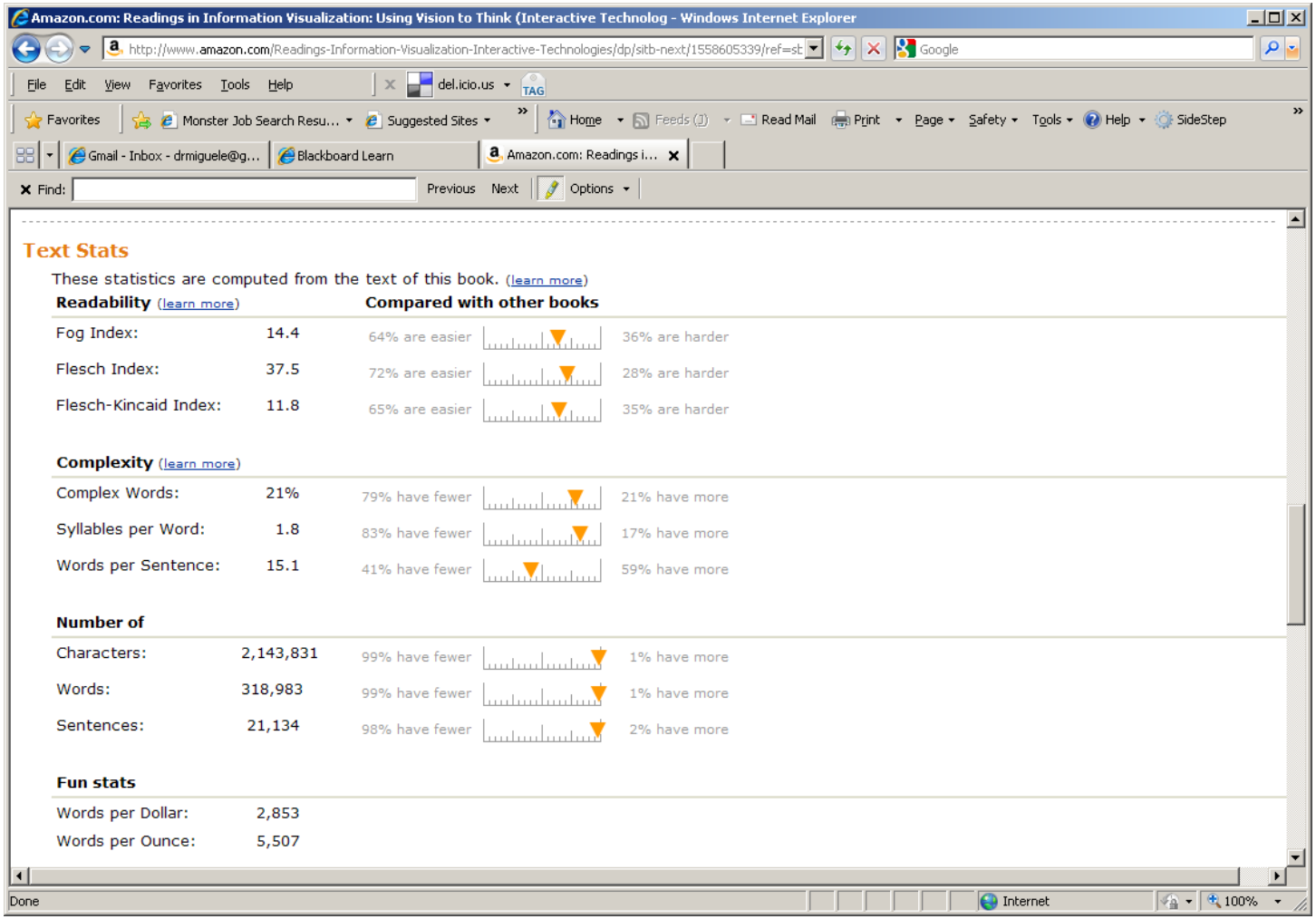


Amazon product concordance

- Maybe now a "word cloud"



Sidenote – alternate text data



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

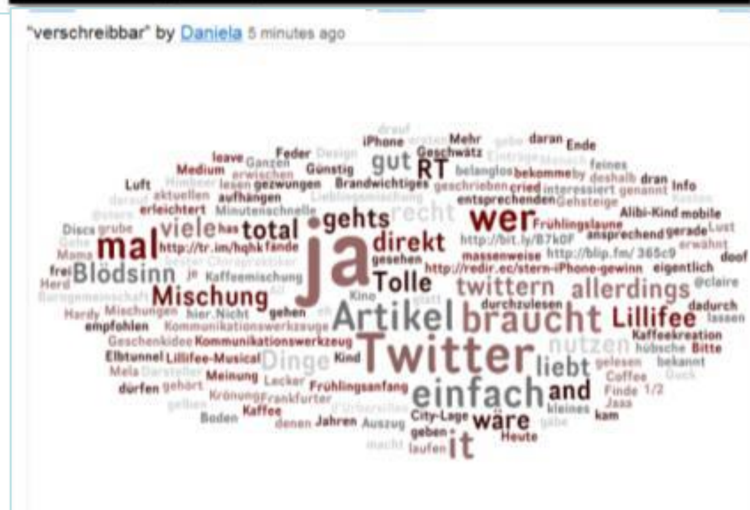
Hearst & Rosner, HICSS'08

Wordle

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

can do volunteering
can do volunteering from Scope, Leonard Cheshire and Russell Commission. Available at www.scope.org.uk —

"Generals Douglas McArthurs Speech" by Bob the Builder 31 minutes ago



"verschreibbar" by Daniela 5 minutes ago

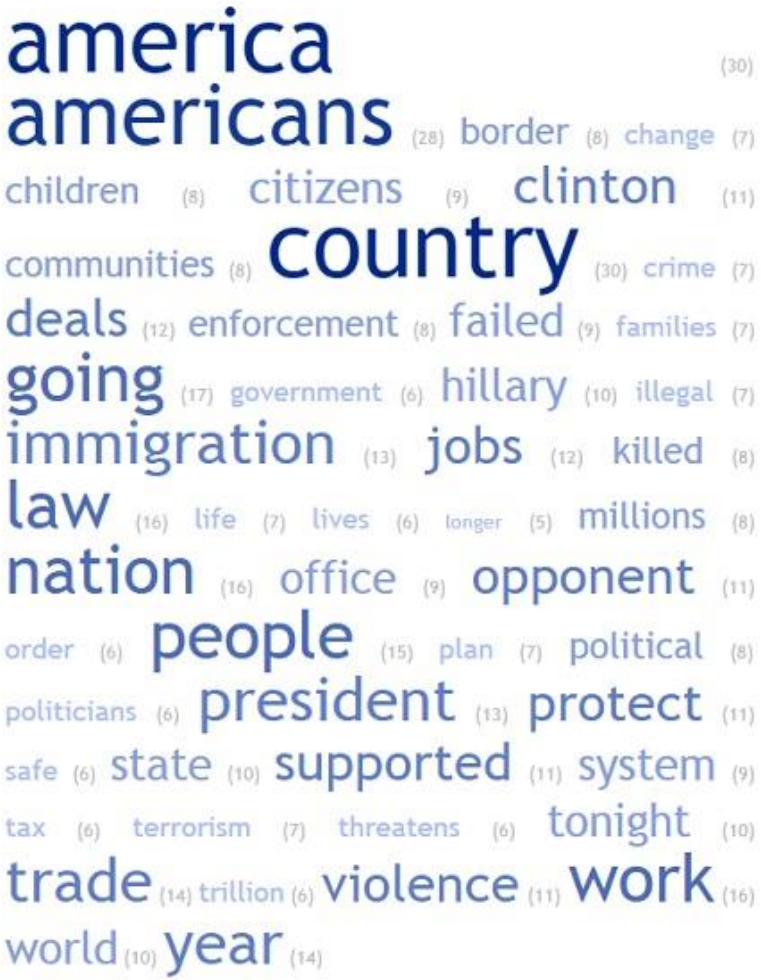
Wordle – Our syllabus

Tagclouds (Wordle)

- Tightly packed words, sometimes vertical or diagonal
- Word size is linearly correlated with frequency (typically square root in cloud)
- Multiple color palettes
- Layout algorithm not published

Viegas et al. *TVCG'09*

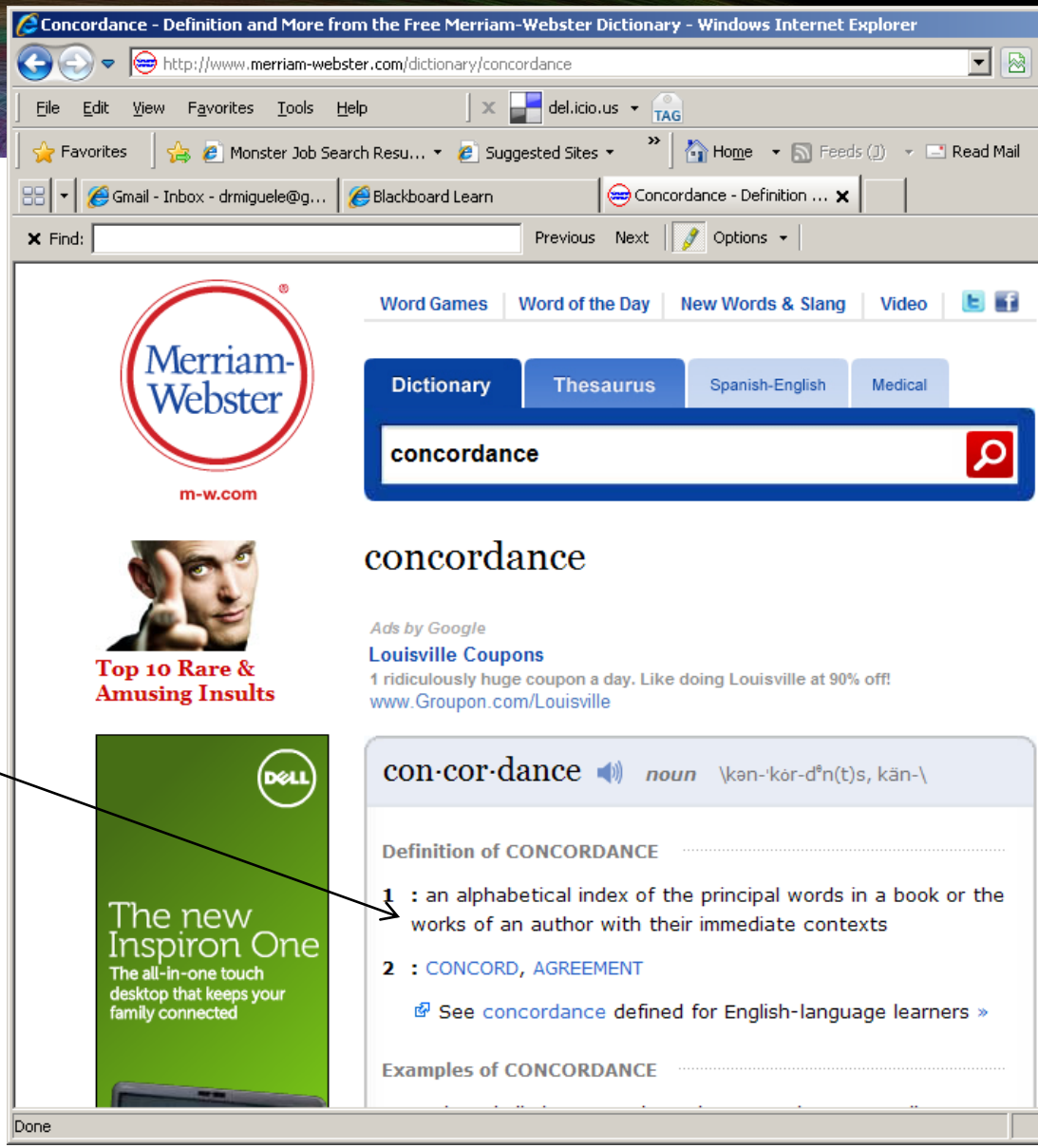
Tagcrowd – Trump's acceptance speech



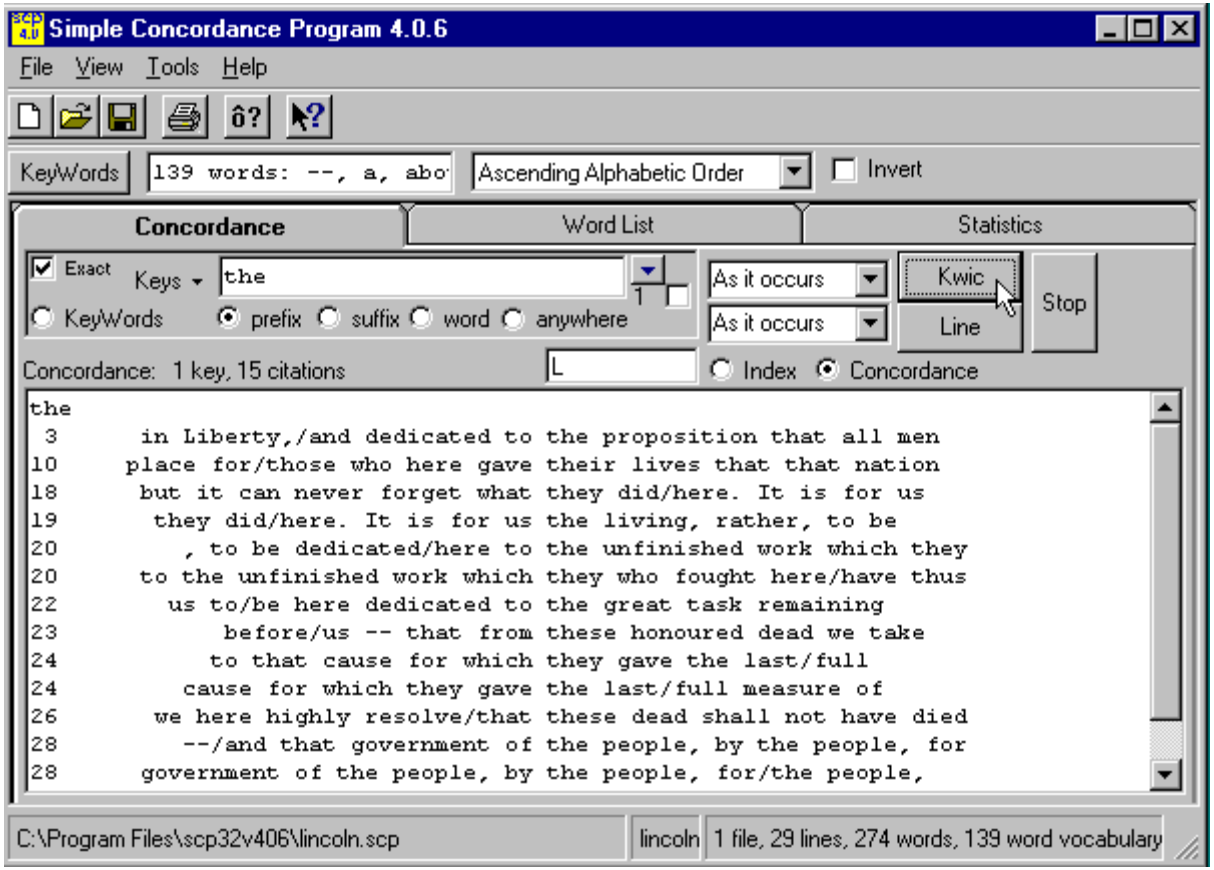
https://constitutioncenter.org/blog/word-cloud-analysis-of-donald-trumps-acceptance-speech/#tagcloud

Concordance

Definition

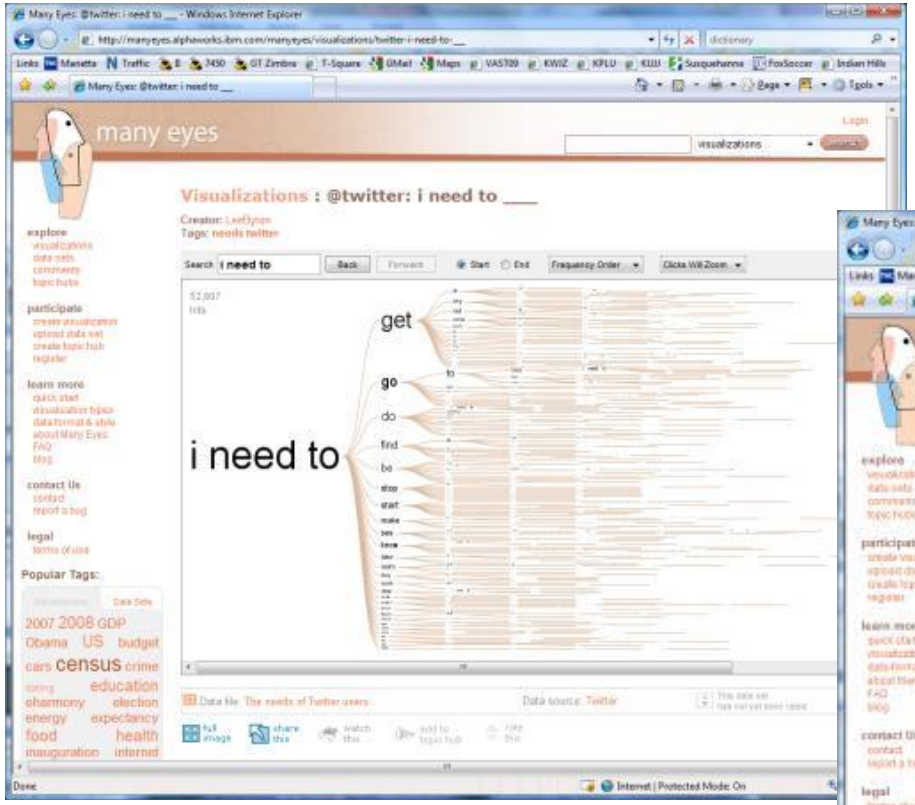


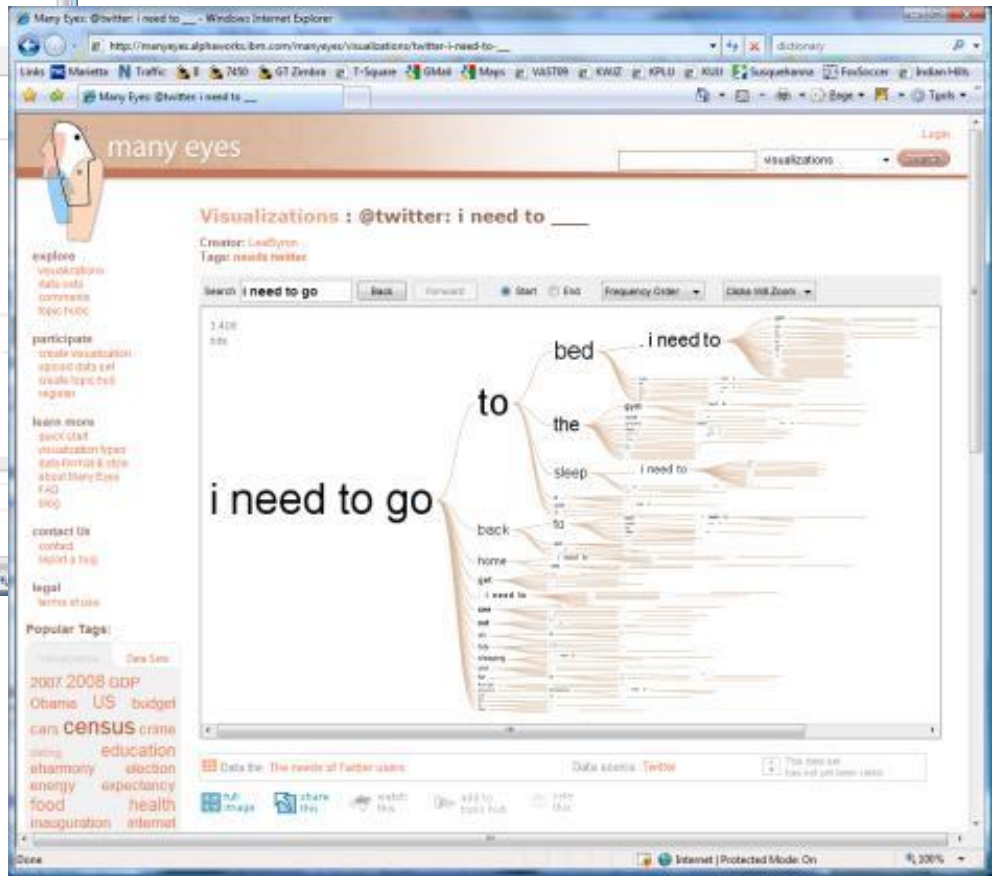
Concordance in Text



http://www.textworld.com/scp/

ManyEyes Word Tree



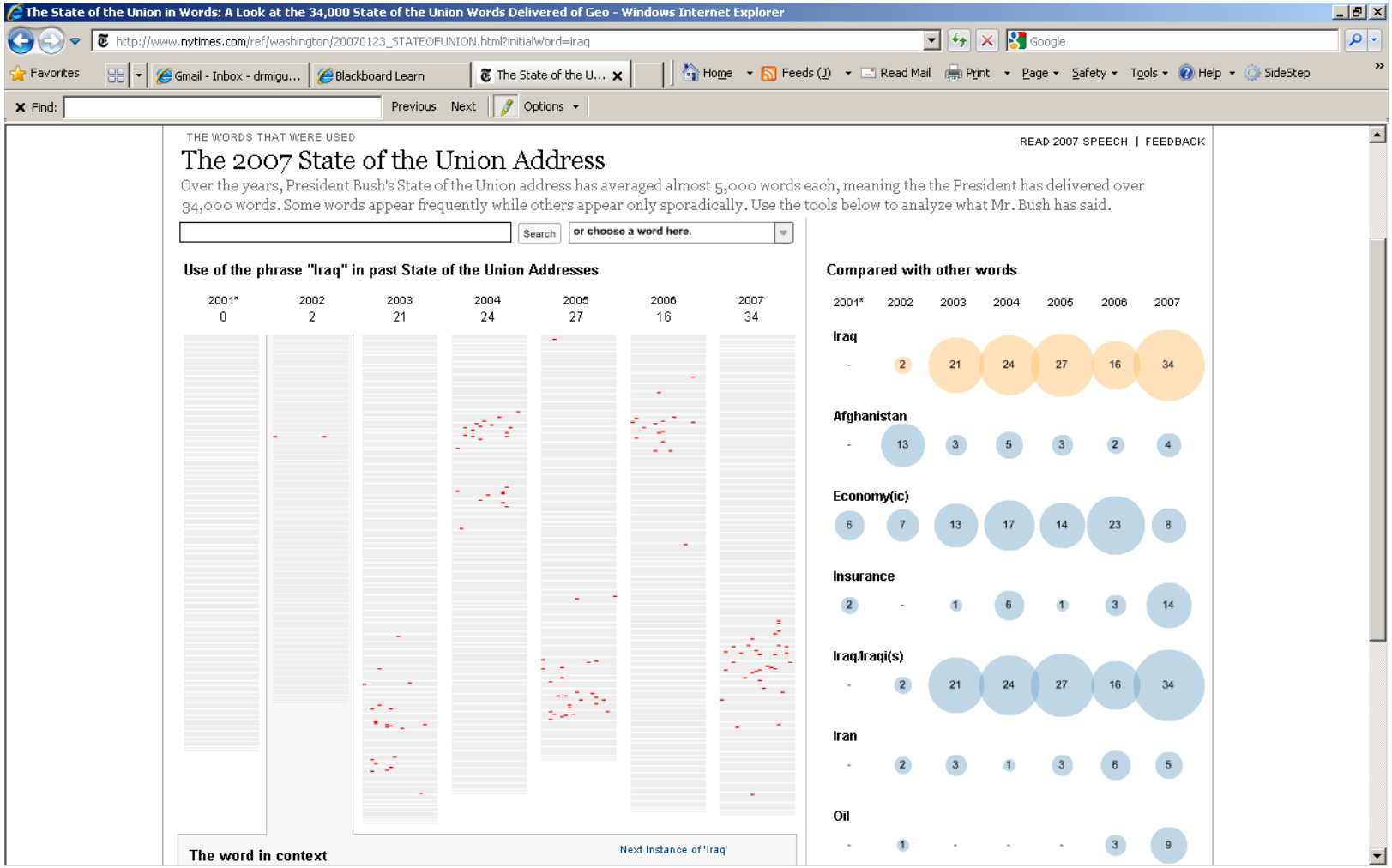


Word Tree

- 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

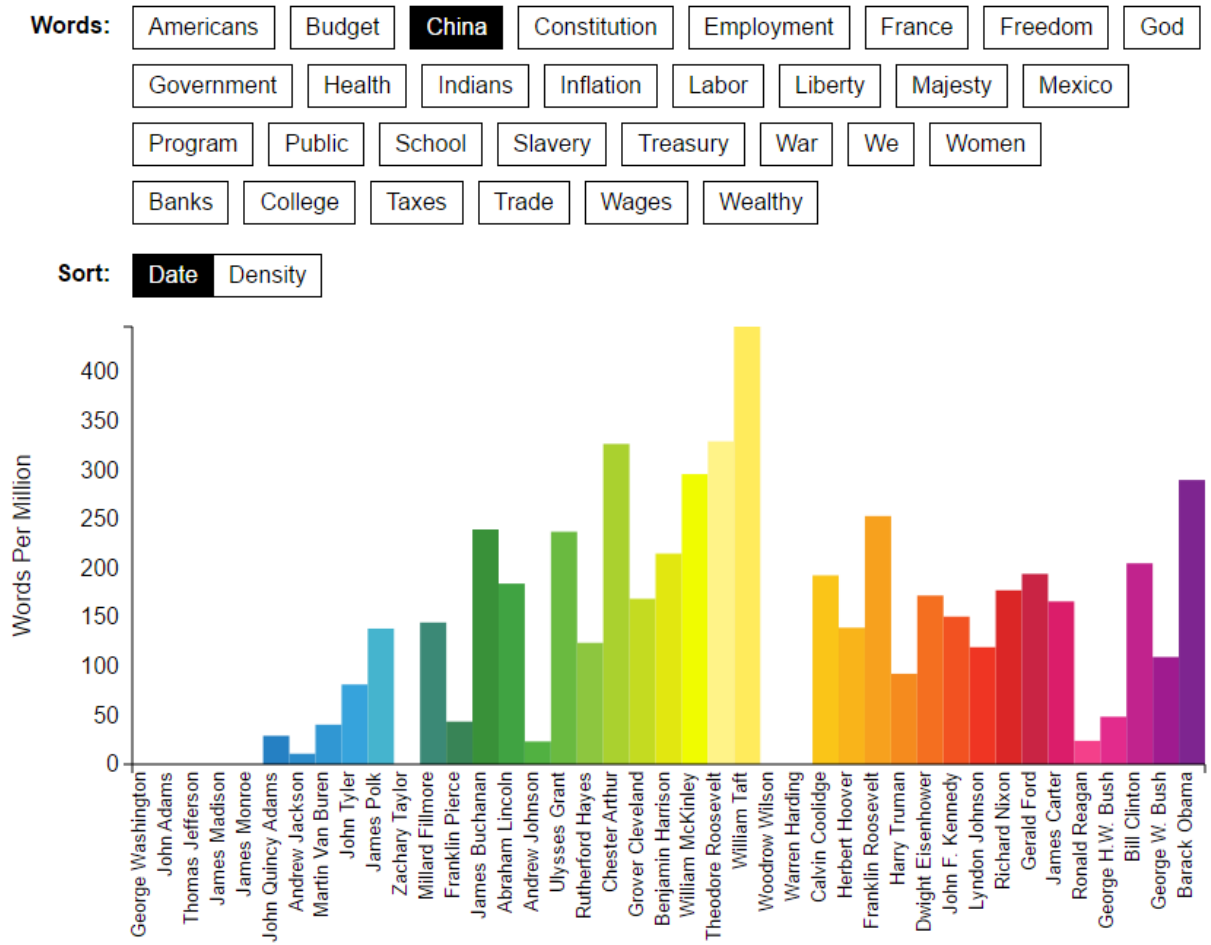
Wattenberg & Viégas, *TVCG'08*

Another Word View



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

Overview & Details on Demand



https://www.theatlantic.com/politics/archive/2015/01/the-language-of-the-state-of-the-union/384575/

Layering & separation

The New York Times

Politics

Search All NYTimes.com Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

Published: January 25, 2011

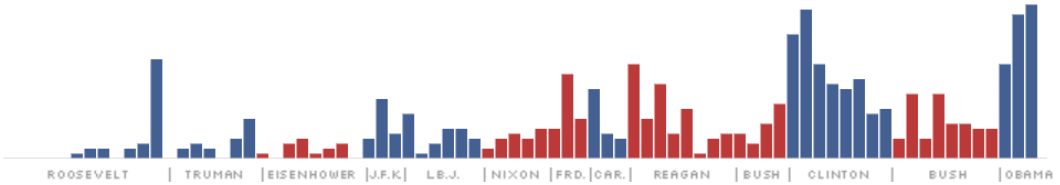
Patterns of Speech: 75 Years of the State of the Union Addresses

In 2010, President Obama was the first modern president to use the words "bubble," "supermajority" and "obesity" in a State of the Union speech. But other words have a longer history. Below, a historical look at the number of times presidents have used selected words in their State of the Union addresses (or analogous speeches) from 1934 to 2011.

'jobs'

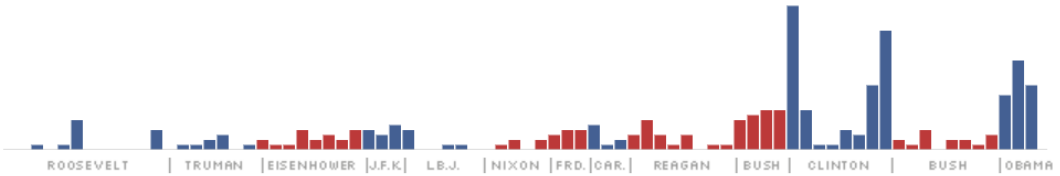
With unemployment above 9 percent, jobs were a focus of President Obama's speech. Historically, jobs get mentioned in the speech in rough correlation to the economic cycle, with spikes around 1975, 1981, 1991 and 2002.

Words included: JOB, JOBS



'invest'

Historically, Democrats use this word more than Republicans, and they mean "public" investment, or new government programs. Bill Clinton used it a lot at the beginning and end of his term, first to propose new programs, and last to take credit for successful ones. Mr. Obama mentioned the word 13 times, proposing new investments in information technology, clean energy and science research.



http://archive.nytimes.com/www.nytimes.com/interactive/2011/01/25/us/politics/state-of-the-union-words-used.html?hp

Multiple Words

- How about sequences or pairs of words?
- Are there good ways to present them?

Phrase Nets (in ManyEyes now)

- 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'09*

Examples

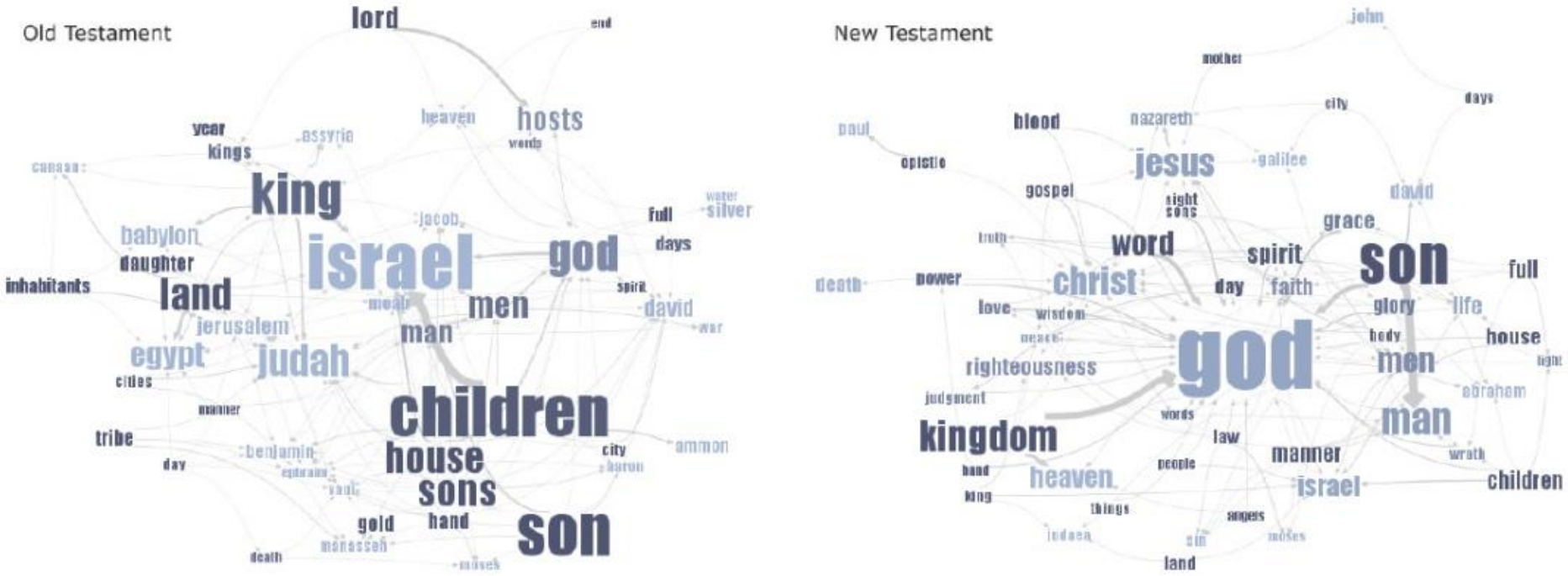


Fig 4. Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

Examples

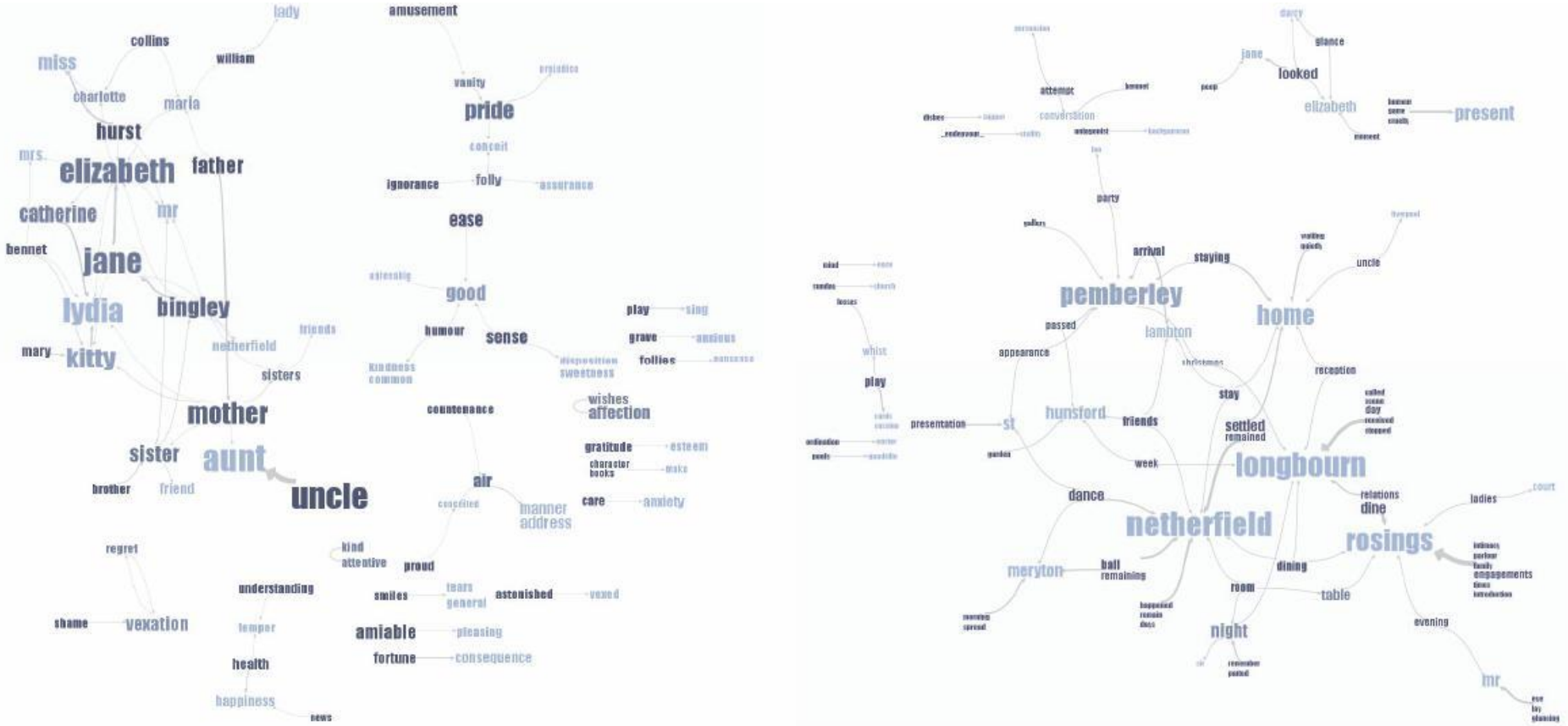
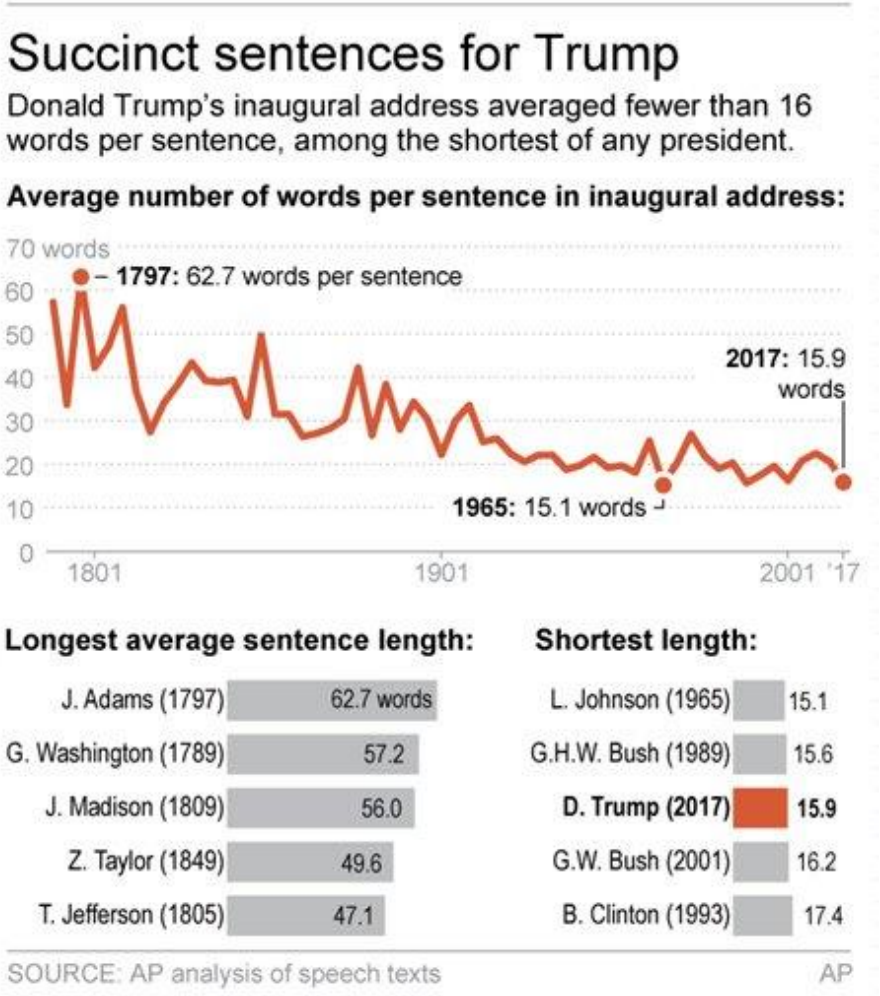


Fig 5. Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

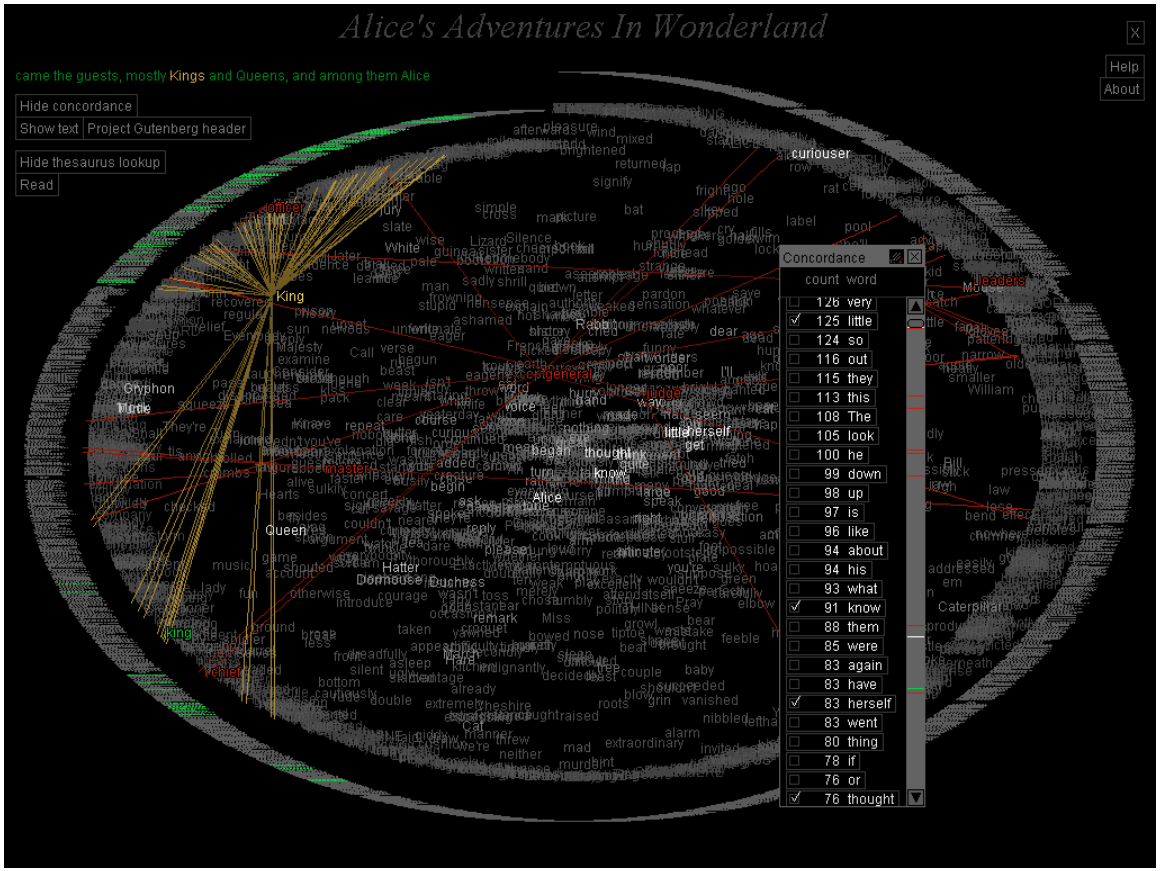
From words to sentences



Another Challenge

- Visualize an entire book
- What does that mean?
- How about showing word appearances?

TextArc



- Sentences laid out in order of appearance
- Words near to where they appear
- Much interaction

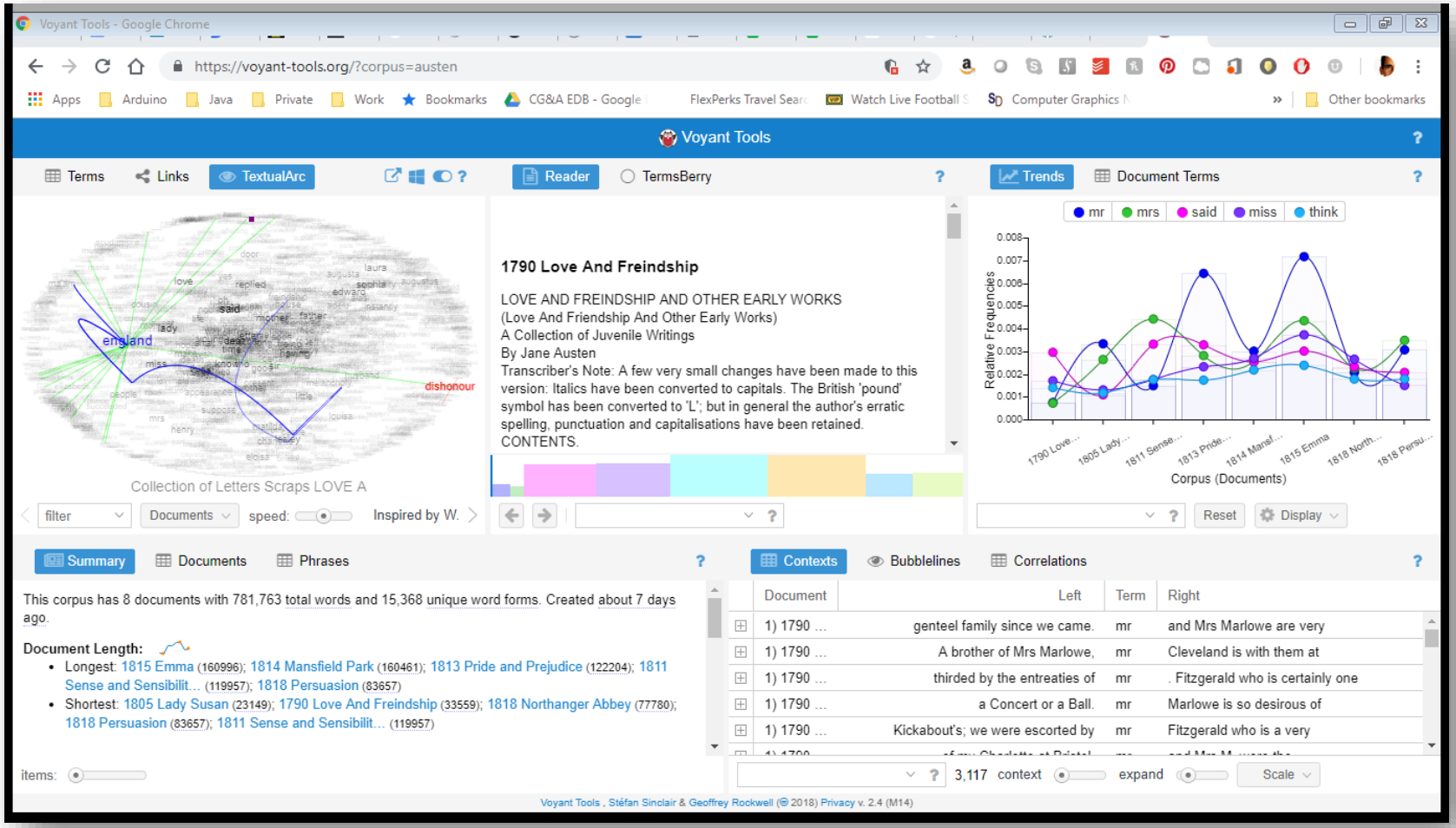
http://www.textarc.org

http://vallandingham.me/textarc/

Is it time?

- Definition of time in documents
  - Age?
    - publication date, editing date
  - Version?
  - Sequence of development?
    - StarWars I, II, III, …
  - Order/position/role in document?
    - Prologue, Rising action, Climax, Epilogue

Voyant Toolkit (https://voyant-tools.org)

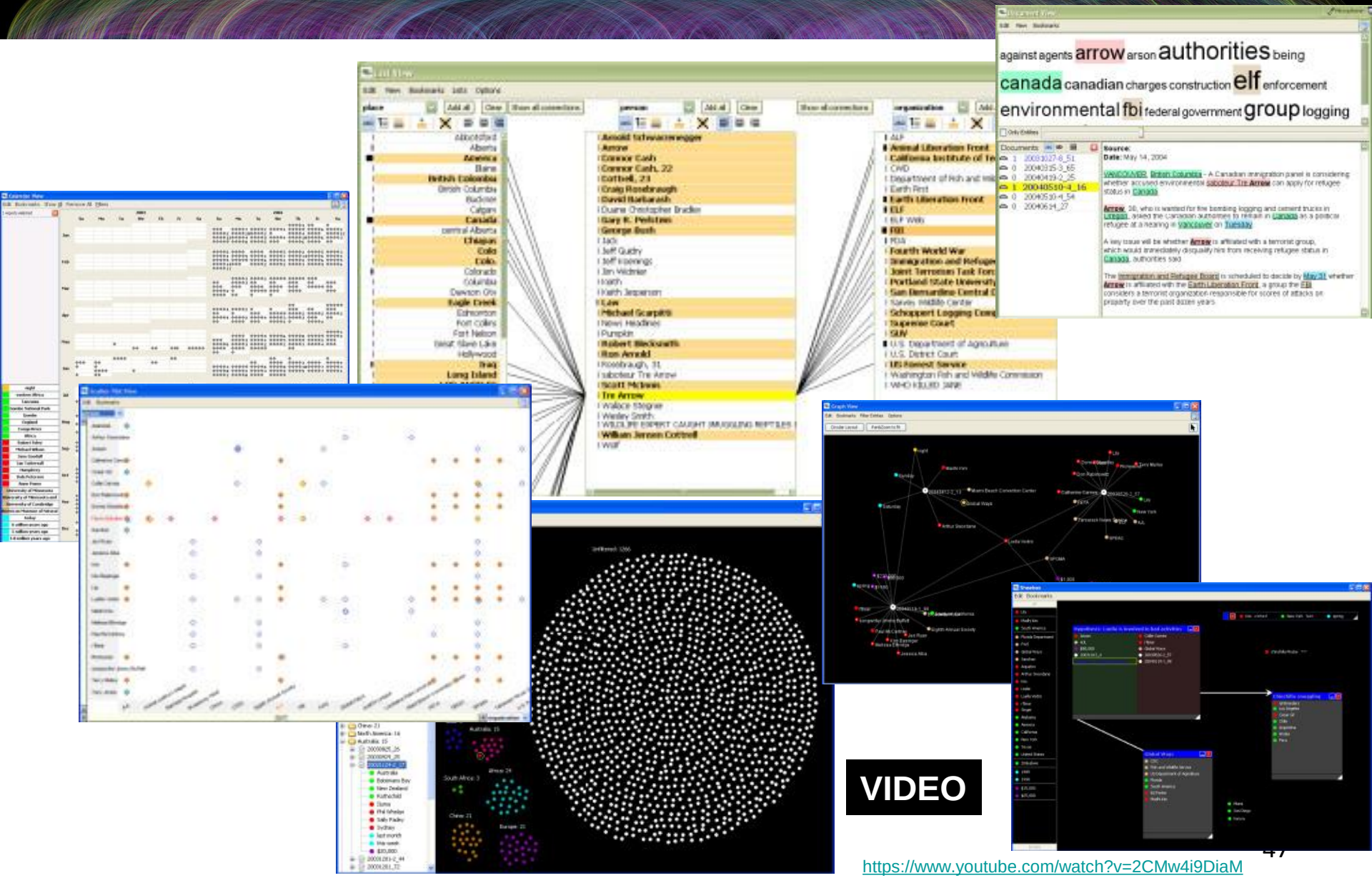


http://docs.voyant-tools.org/start/

Jigsaw (Stasko et al., *Information Visualization* '08)

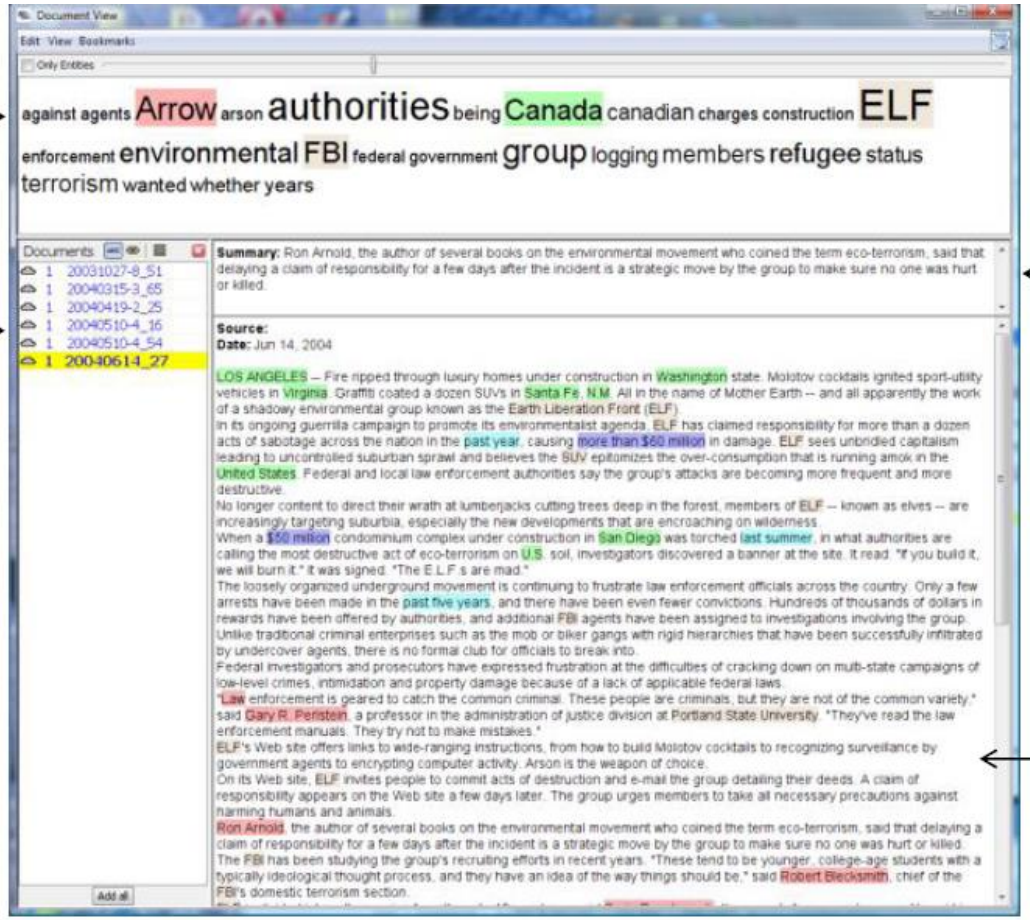
- Targeting sense-making scenarios
- Variety of visualizations ranging from word-specific, to entity connections, to document clusters
- Primary focus is on entity-document and entity-entity connection
- Search capability coupled with interactive exploration

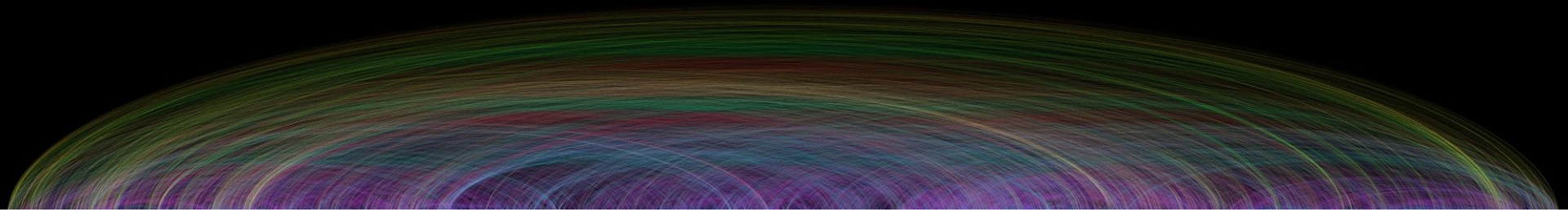
Jigsaw Views



https://www.youtube.com/watch?v=2CMw4i9DiaM

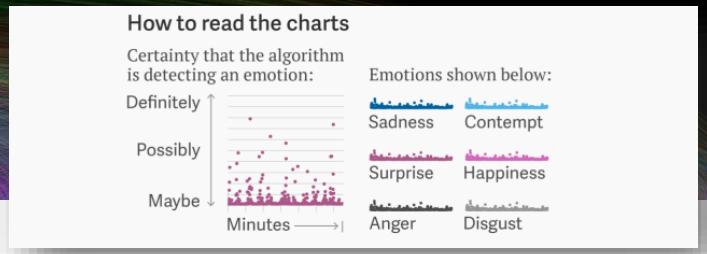
Document View





Sentiment Analysis

Sentiment Analysis

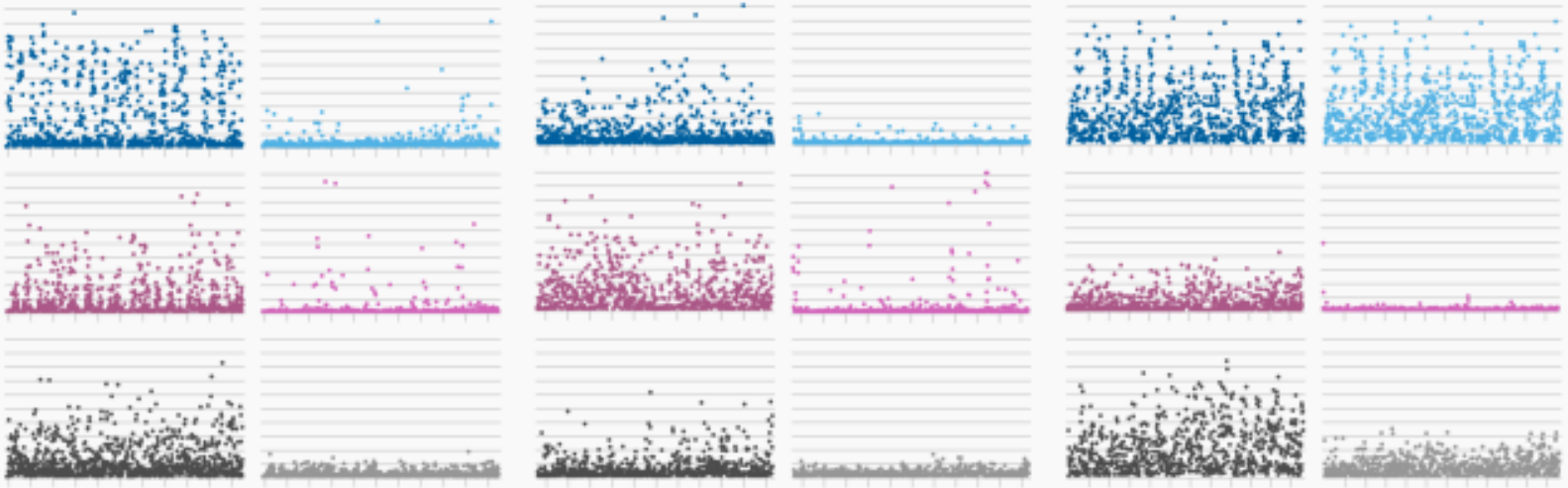


Emotions shown below:

Sadness Contempt Surprise Happiness Anger Disgust

Emotions detected by the computer

Debate with Clinton | Acceptance speech | Inauguration speech

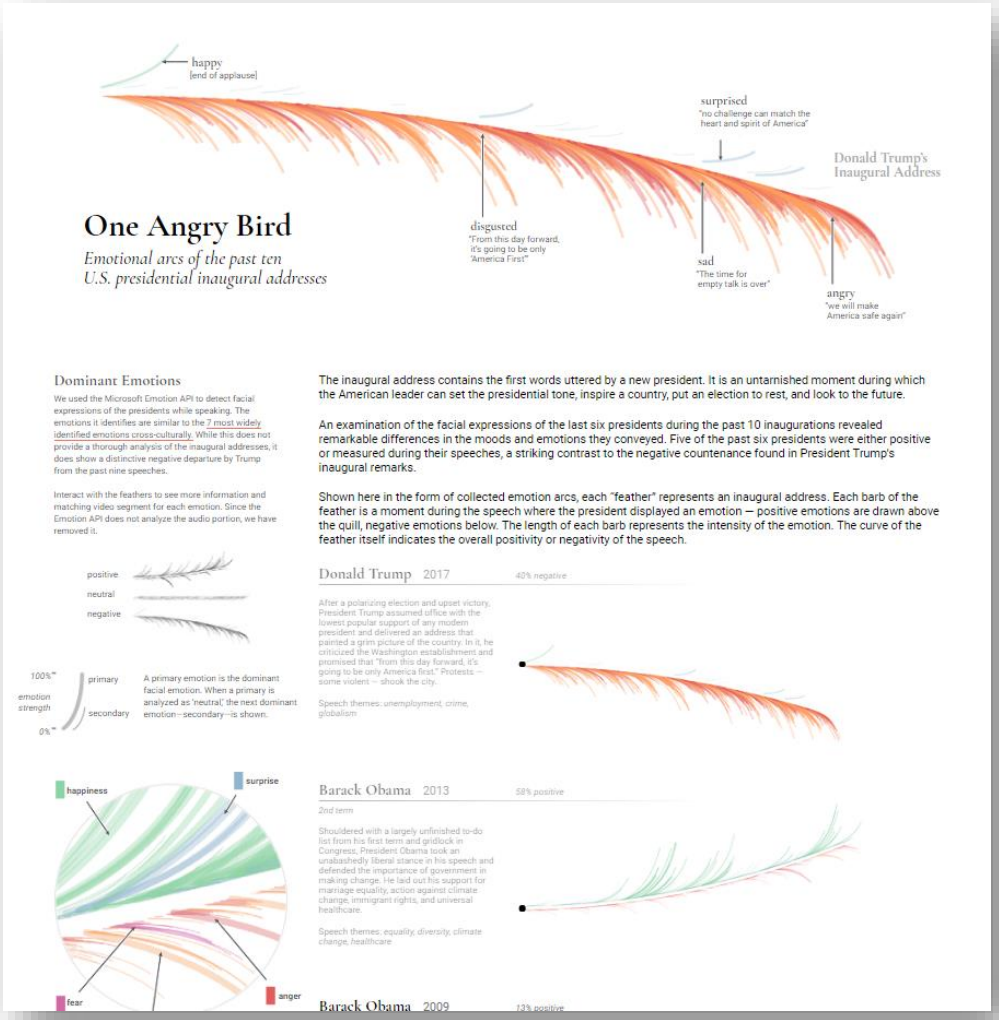


https://qz.com/896768/a-robots-analysis-of-trumps-inauguration-speech-reveals-what-he-was-feeling-while-taking-the-oath-of-office/

Quartz | qz.com | Sarah Slobin

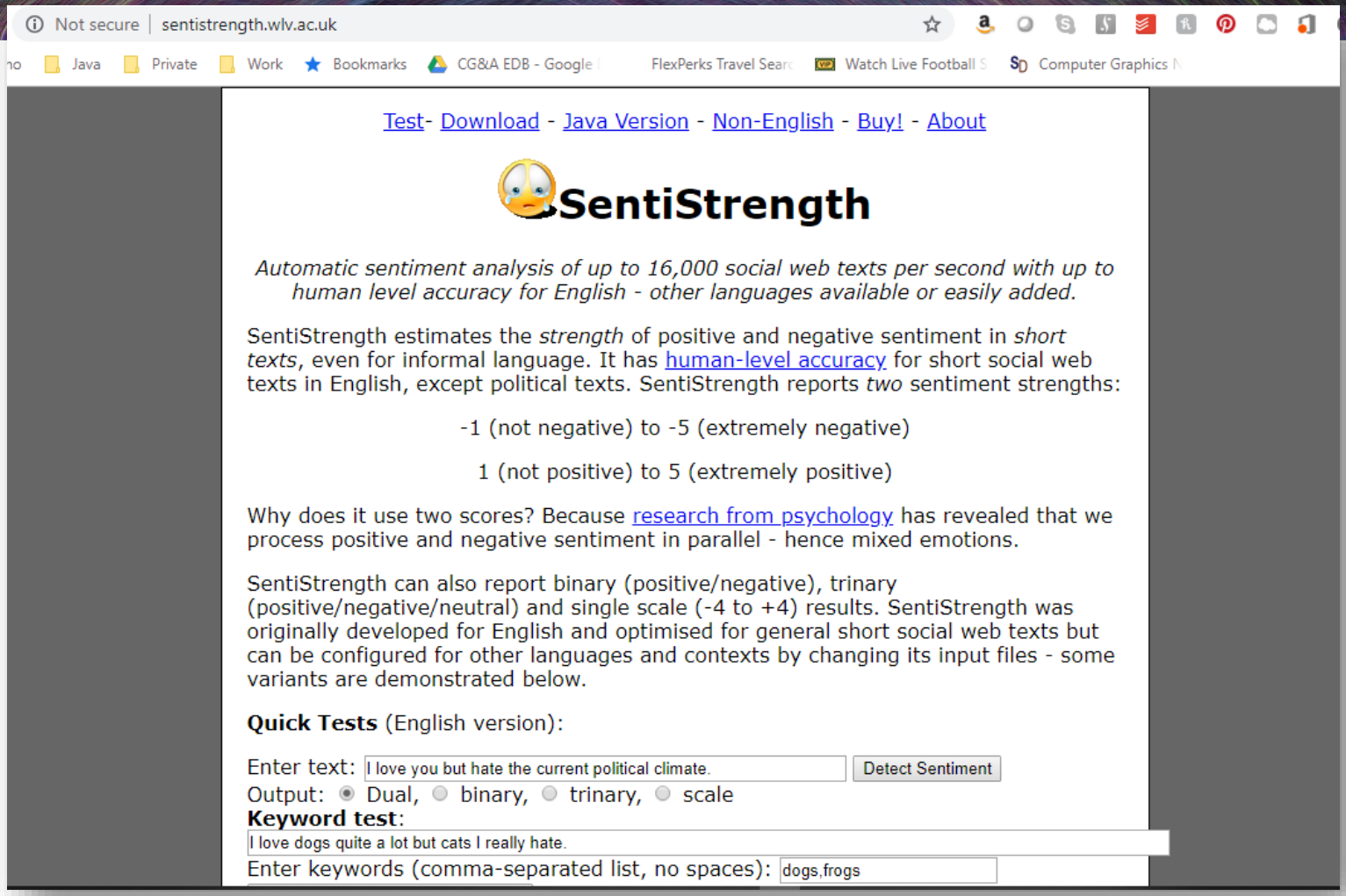
Data via Amir Imani, data scientist

Sentiment Analysis



http://emotions.periscopic.com/inauguration/

Nice look into how Sentiment Analysis works



See also: https://www.talkwalker.com/blog/best-sentiment-analysis-tools