



Natural Language Processing and Text Visualization

Natural Language Lab at SFU

<http://natlang.cs.sfu.ca/>

Text is tough (to visualize)*

- Very high dimensionality
- Topic models are popular because they reduce the dimensionality
- Language is compositional and ambiguous
- Reading is foveal, needs attention
- Language can be unordered and abstract
- Multiple pieces of information depending on viewpoint

* i247: Information Visualization and Presentation by Marti Hearst

Text is not pre-attentive

SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
GOVERNS PRECISE EXAMPLE MERCURY SNREVOG ESICERP ELPMAXE YRUCREM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC
SUBJECT PUNCHED QUICKLY OXIDIZED TCEJBUS DEHCNUP YLKCIUQ DEZIDIXO
CERTAIN QUICKLY PUNCHED METHODS NIATREC YLKCIUQ DEHCNUP SDOHTEM
SCIENCE ENGLISH RECORDS COLUMNS ECNEICS HSILGNE SDROCER SNMULOC

Text can be abstract

- Abstract concepts are difficult to visualize
 - The dog.
 - The dog cavorted.
 - The man walks the cavorting dog.
 - As the man walks the cavorting dog, thoughts arrive unbidden of the previous spring, so unlike this one, in which walking was marching and dogs were baleful sentinels outside unjust halls.
- Combinations of abstract concepts are even more difficult.

Text is about multiple topics

- Categories are not ordered
- Organizing by topics alone miss important distinctions
- Consider an article about:
 - NAFTA
 - The effects of NAFTA on truck manufacture
 - The effects of NAFTA on productivity of truck manufacture in the neighbouring cities of El Paso and Juarez

Search and Text Visualization

- Nominal data is hard to visualize
- Goals of search vs. text analysis
 - Only a tiny fraction of those people who want to use search want to analyze text.
 - For those analysts, there are many interesting ideas available.

Programming Languages

C, C++, Java, Python, ...

- unambiguous
- fixed
- designed
- learnable?
- known simple semantics

Natural Languages

French, English, Korean, Chinese, Tagalog, ...

- ambiguous
- evolving
- transmitted
- learnable
- complex semantics

Natural Language Processing (NLP)

- NLP is the application of a computational theory of human language
- Language is the predominant repository of human interaction and knowledge
- Goal of NLP: programs that “listen in”
- The AI Challenge: the Turing test
- Lots of speech and text data available

Natural Language: What is it?

- Answers from linguistics

Natural Language (NL) vs. Artificial Language

- NL is complex, displays recursive structure
- Learning of language is an inherent part of NL
- Language has idiosyncratic rules and a complex mapping to thought

Language has structure

- Finnish word structure
 - talossansakaanko ‘not in his house either?’
 - kynässänsäkäänkö ‘not in his pen either?’
- English phrase structure
 - It is likely that John went home.
 - That John went home is likely.
 - OK: Where is it likely that John went **t**?
 - Not OK: *Where is that John went **t** likely?

Language is recursive

- Combine the following two sentences:
 - The clown watches the ballerina
NP1 V1 NP2
 - The musician hits the clown
NP3 V2 NP4
- Many possible combinations of the two sentences:
 - The clown watches the ballerina and the musician hits the clown
- Use a modifier to combine them:
 - The clown who the musician hits watches the ballerina
NP1/4 NP3 V2 V1 NP2
 - The musician hits the clown who watches the ballerina
NP3 V2 NP4/1 V1 NP2

Language is recursive

- Finite resources but possibly infinite utterances (via recursion)
- **Sparse** language:
 - a sparse language is a set of strings where the number of strings of length n is bounded by a polynomial function of n
 - Regular and context-free languages are **dense** as shown by Chomsky, Flajolet, Incitti

Language is Parsed

- Google's Computer Might Betters Translation Tool
 - New York Times March 8, 2010
- Number of Lothian patients made ill by drinking rockets
 - Edinburgh Evening News, March 4, 2010
- Violinist linked to JAL crash blossoms
 - *<http://languagelog.ldc.upenn.edu/nll/?p=1693>*

Language is ambiguous

- Lung cancer in women mushrooms
 - Mushrooms is noun or a verb?
- Ban on nude dancing on governor's desk
 - Similar to “if-then-else” ambiguity
- Island Monks Fly in Satellite to Watch Pope Funeral
 - “fly in” vs. “fly [_{OBJ} in Satellite]” hidden segmentation
- British Left Waffles on Falkland Islands
 - Is it British/Noun Left/Verb or British Left/NP Waffles/Verb?

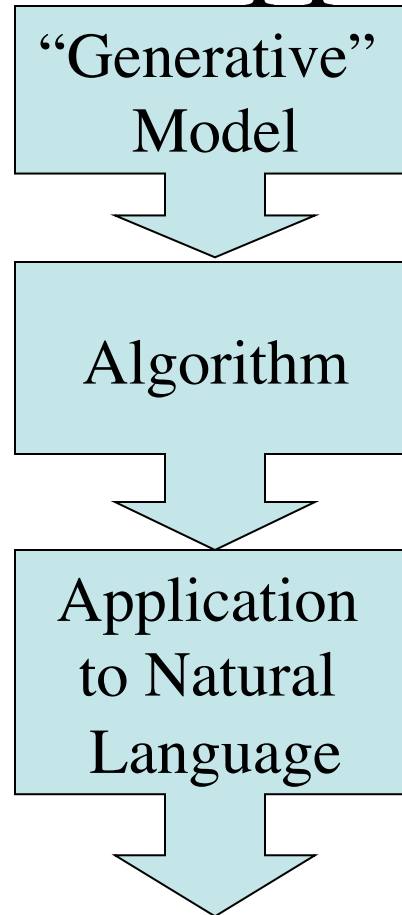
- **Phonetics** acoustic and perceptual elements
- **Phonology** inventory of basic sounds (phonemes) and basic rules for combination
 - e.g. vowel harmony. **Anupu** is pronunciation of **Anoop** in Classic Period Mayan
- **Morphology** how morphemes combine to form words, relationship of phonemes to meaning
 - e.g. **delight-ed** vs. **de-light-ed**
- **Syntax** sentence (utterance) formation, word order and the formation of constituents from word groupings
 - e.g. **The clown who the musician hits watches the ballerina**
- **Semantics** how do word meanings recursively compose to form sentence meanings (from syntax to logical formulas)
 - e.g. **Everyone is not here** => what does this mean? **Nobody** / Not everyone is here.
- **Pragmatics** meaning that is not part of compositional meaning,
 - e.g. **This professor dresses even worse than Anoop!**



Terminology: Grammar

- Grammar can be prescriptive or descriptive
- *Descriptive grammar* is a **model** of the form and meaning of a speaker of a language
- Grammar books for learning a language are *prescriptive grammars*, usually style manuals or rules for how to write clearly
- Except for some NLP apps like grammar checking or teaching, we are usually interested in creating models of language

General Approach



Phonology / Morphology / Syntax / Semantics / Pragmatics

Some definitions

- **Classification:** assigning to the input one out of a finite number of classes, e.g.: Document -> spam, formalization -> Noun
- **Sequence learning/Tagging:** assigning a sequence of classes, e.g.: I/ Pron can/Modal open/Verb a/Det can/Noun
- **Parsing:** assigning a complex structure, e.g.: formalization -> (Noun (Verb (Adj formal) -ize) -ation)
- **Grammar development:** human driven creation of a model for some linguistic data
- **Transduction:** transforming one linguistic form to another, e.g. summarization, translation, tokenization
- **Tracking/Co-reference:** after detecting an entity (say a person) tracking that entity in subsequent text; co-reference of a pronoun to its antecedent; “lexical chains” of similar concept
- **Clustering:** unsupervised grouping of data using similarity, constructing “phylogenetic” trees

NLP: Lots of Applications

- Doc classification
- Doc clustering
- Spam detection
- Information extraction
- Summarization
- Machine translation
- Cross Language IR
- Multiple language summarization
- Language generation
- Plagiarism or author detection
- Error correction, language restoration
- Language teaching
- Question answering
- Knowledge acquisition (dictionaries, thesaurus, semantic lexicons)
- Speech recognition
- Text to Speech
- Speaker Identification
- (multi-modal) Dialog systems
- Deciphering ancient scripts

Information Extraction

<DOC><SO> WALL STREET JOURNAL (J),
PAGE B5 </SO>

<TXT><p>

<PERSON-1>

New York Times Co. named Russell T. Lewis,

<ORGANIZATION-1>

general manager of its
major newspaper,

responsible for all business-side activities.

He was executive vice president and

<PERSON-2>

general manager. He succeeds Lance R. Primis,
who in September was named president
and chief operating officer of the parent.

</p></TXT></DOC>

<SUCCESSION-1>

ORGANIZATION : <ORGANIZATION-2>

POST : "president"

WHO_IS_IN : <PERSON-1>

WHO_IS_OUT : <PERSON-2>

tion

AL (J),

<TXT> <p>

<PERSON-1>

New York Times Co. named Russell T. Lewis,

<ORGANIZATION-1>

general manager of its
major newspaper,

responsible for all business-side activities.

He was executive vice president and

<PERSON-2>

general manager. He succeeds Lance R. Primis,
who in September was named president
and chief operating officer of the parent.

</p> </TXT> </DOC>

LocalizationID

PSID

1) Select "valid" if the passage contains strong evidence of an experimentally determined localization.

[PubMed Entrez](#)

PMID

[PubMed Centreal](#)

PMCID

The cytoplasmic membrane proteins ExbB and ExbD support TonB-dependent active transport of iron siderophores and vitamin B12 across the essentially unenergized outer membrane of Escherichia coli.

Valid ☐

Invalid ☐

Maybe ☐

Reviewer

Comments

2) If the passage is valid then select whether the protein, organism, and location names are also valid. (If you want to defer your decision then select neither valid nor invalid)

Protein:

☐

Valid

Invalid

☐

Organism:

☐

Valid

Invalid

☐

Location:

☐

Valid

Invalid

☐

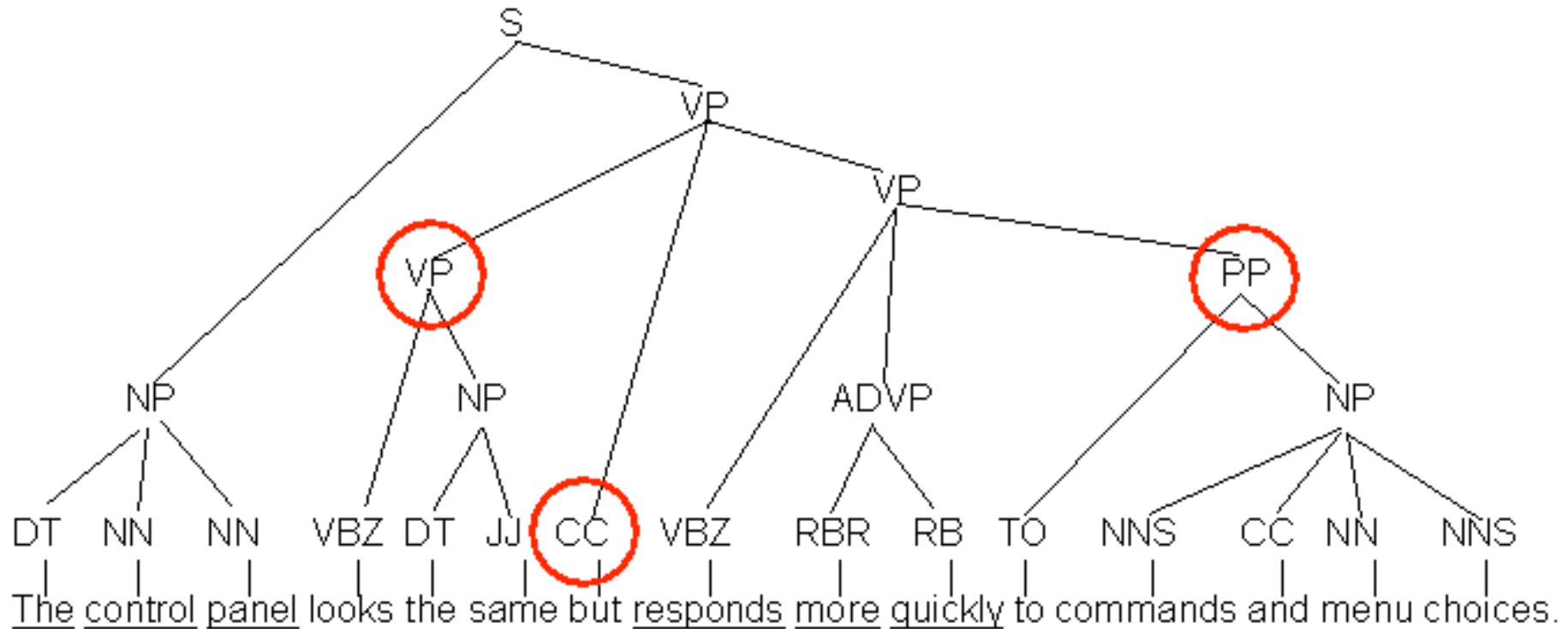
SQuASH: SFU QA Summarization System

Input: 25 news articles, Complex question **Output:** 250-word summary

Q. Describe developments in the movement for the independence of Quebec from Canada.

A. Canadian Prime Minister Jean Chretien has ruled out the possibility that the French-speaking province of Quebec could declare independence even if a majority of Quebecers vote for secession in the next referendum. Premier Lucien Bouchard and the separatist Parti Quebecois have retained control of Canada's largest province, possibly setting the stage for a bitter confrontation with the rest of Canada, and possibly another referendum on independence. Canada's Supreme Court managed to rule on the question of Quebec secession without infuriating leaders on either side of a bitter ideological divide. ...

Sentence Compression



Paraphrasing

- open borders imply increasing racial fragmentation in *european countries* .
- open borders imply increasing racial fragmentation in *the countries of europe* .
- open borders imply increasing racial fragmentation in *european states* .
- open borders imply increasing racial fragmentation in *europe* .
- open borders imply increasing racial fragmentation in *european nations* .
- open borders imply increasing racial fragmentation in *the european countries* .

Why is paraphrasing useful?

Sentiment detection

Annotate tweets using labels from http://en.wikipedia.org/wiki/List_of_emoticons

10 Happiest Tweets

- @WRiTExMiND no doubt! <--guess who I got tht from? Bwahaha anyway doe I like surprising people it's kinda my thing so ur welcome! And hi :)
- @skvillain yeh wiz is dope, got his own lil wave poppin! I'm fuccin wid big sean too he signed to kanye label g.o.o.d music
- And @pumahbeatz opened for @MarshaAmbrosius & blazed! So proud of him! Go bro! & Marsha was absolutely amazing! Awesome night all around. =)
- Awesome! RT @robscoms: Great 24 hours with nephews. Watched Tron, homemade mac & cheese for dinner, Wii, pancakes & Despicable Me this am!
- Good Morning 2 U Too RT @mzmonique718: Morningggg twitt birds!...up and getting ready for church...have a good day and LETS GO GIANTS!
- Goodmorning #cleveland, have a blessed day stay focused and be productive and thank god for life
- AMEN!!!>>>RT @DrSanlare: Daddy looks soooo good!!! God is amazing! To GOD be the glory and victory #TeamJesus Glad I serve an awesome God
- AGREED!! RT @ILoveElizCruz: Amen to dat... We're some awesome people! RT @itsVonnell_Mars: @ILoveElizCruz gotta love my sign lol
- #word thanks! :) RT @Steph0e: @IBtunes HAppy Birthday love!!! =) still a fan of ya movement... yay you get another year to be dope!!! YES!!
- Happy bday isaannRT @isan_coy: Selamat ulang tahun yaaa RT @Phitz_bow: Selamat siang RT @isan_coy: Slammat pagiiii

Sentiment detection

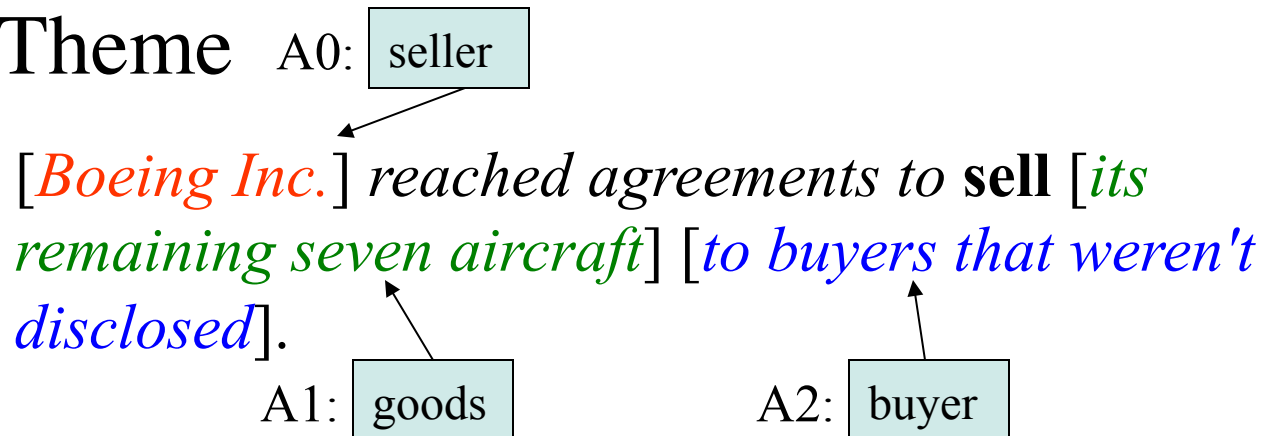
Annotate tweets using labels from http://en.wikipedia.org/wiki/List_of_emoticons

10 Saddest Tweets

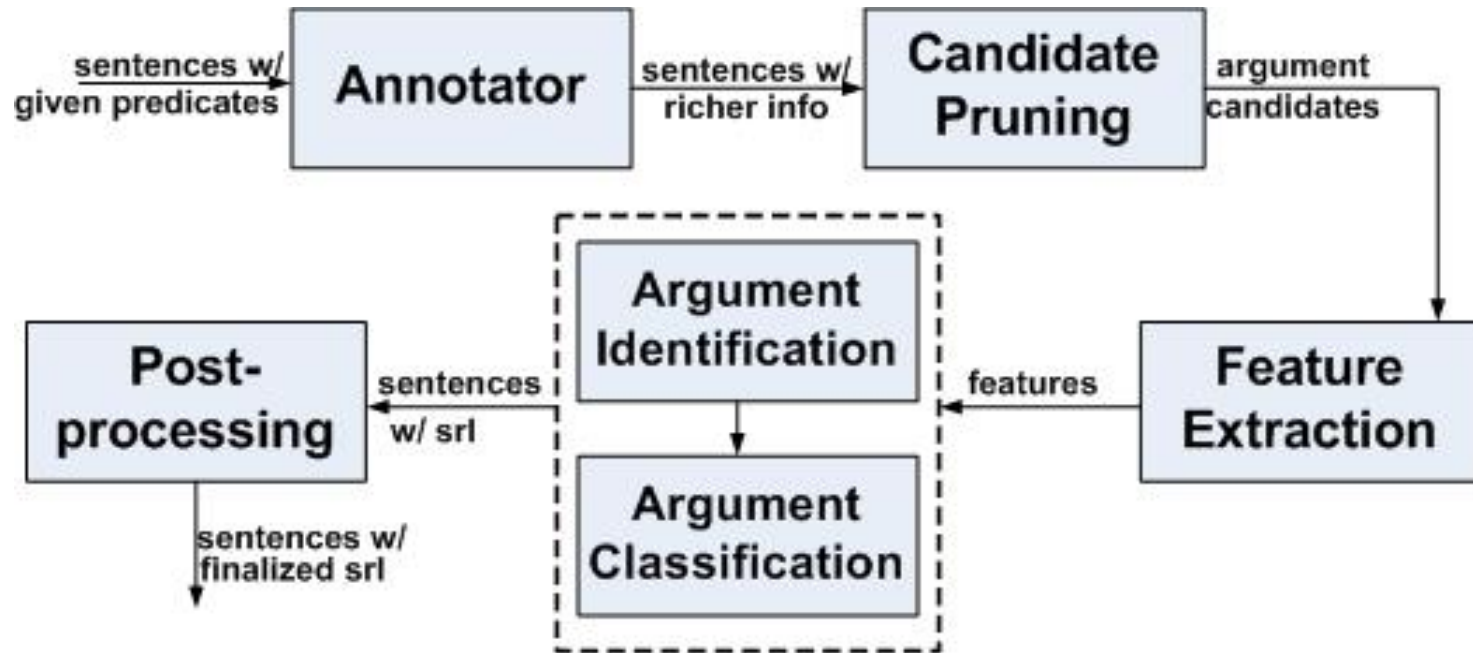
- Migraine, sore throat, cough & stomach pains. Why me God?
- Ik moet werken omg !! Ik lig nog in bed en ben zo moe .. Moet alleen opstaan en tis koud buitn :(
- I Feel Horrible ' My Voice Is Gone Nd I'm Coughing Every 5 Minutes ' I Hate Feeling Like This :-/
- SMFH !!! Stomach Hurting ; Aggy ; Upset ; Tired ;; Madd Mixxy Shyt Yo !
- Worrying about my dad got me feeling sick I hate this!! I wish I could solve all these problems but I am only 1 person & can do so much..
- Malam2 menggigil+ga bs napas+sakit kepala....badan remuk redam *I miss my husband's hug....#nangismanja#
- Waking up with a sore throat = no bueno. Hoping someone didn't get me ill and it's just from sleeping. D:
- Aaaa ini tenggorokan gak enak, idung gatal bgt bawaannya pengen bersin terus. Calon2 mau sakit nih -___-
- I'm scared of being alone, I can't see to breathe when I am lost in this dream, I need you to hold me?
- Why the hell is suzie so afraid of evelyn! Smfh no bitch is gonna hav me scared I dnt see it being possible its not!

Semantic Role Labeling (SRL)

- For a given verb (predicate), SRL aims to identify and label all its arguments with semantic roles, such as Agent, Patient, and Theme

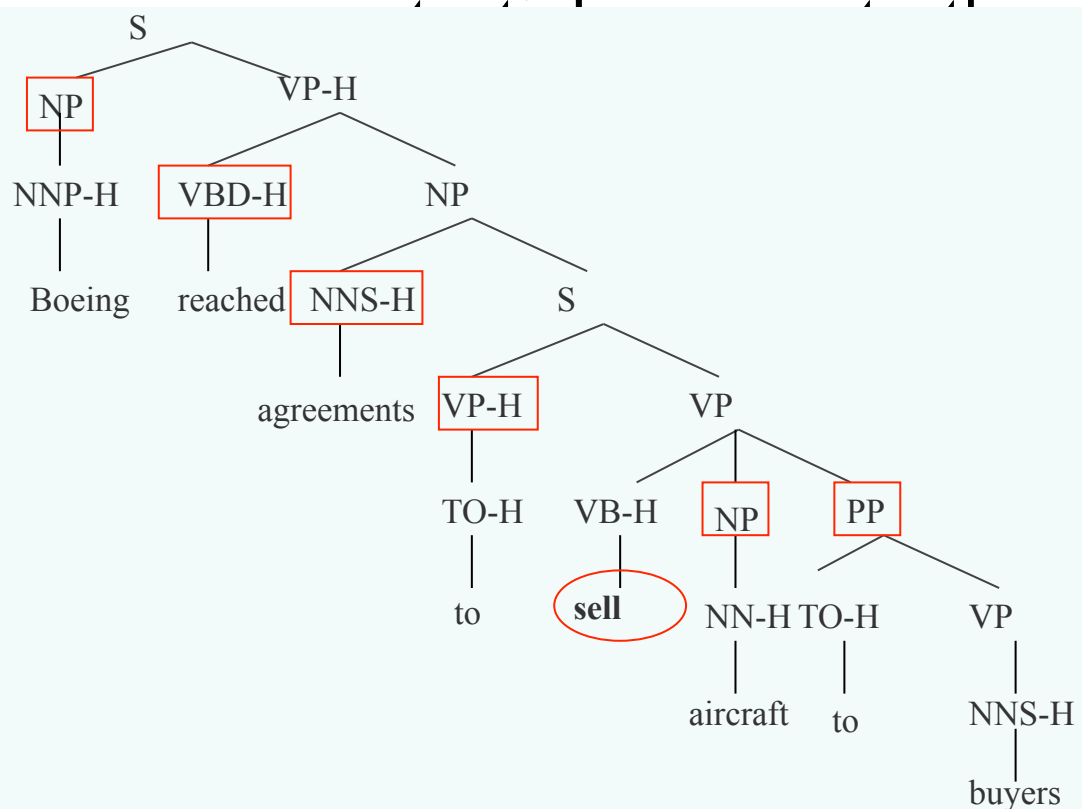


Architecture of a SRL system



Architecture of a SRL system

- On a given parse tree, run the pruning component: some candidates are pruned, others are labeled
- Run a binary classifier on some spans labeled
- Run binary classifier on A0, A1 vs not
- Combine output for each ARG with highest confidence node: A0, A1,



Accuracy of Semantic Role Labeling

	SRL		
	Prec.	Rec.	F1
Overall	81.90	78.81	80.32
A0	88.37	88.91	88.64
A1	81.50	81.27	81.38
A2	73.44	68.74	71.01
A3	75.00	55.49	63.79
A4	74.74	69.61	72.08
A5	100.00	80.00	88.89
AM-*	78.19	69.98	73.86
R-AM-*	73.91	61.44	67.10

Predicates & Entities

Sen. Mitchell added that the agreement requires that the Contras not initiate any military action .

speaker: Sen. Mitchell

say: added

utterance: that the agreement requires that the Contras not initiate any military action

Common role labels used to automatically cluster predicates!

The State Department said there was a `` possibility " that some Nicaraguan rebels were selling their U.S.-supplied arms to Salvadoran guerrillas .

speaker: The State Department

say: said

utterance: there was a `` possibility " that some Nicaraguan rebels were selling their U.S.-supplied arms to Salvadoran guerrillas

User can specify clusters based on one argument label (speaker) or multiple labels (thing_put + where_put)

The State Department said there was a `` possibility " that some Nicaraguan rebels were selling their U.S.-supplied arms to Salvadoran guerrillas , but insisted it wasn't an organized effort .

<say> said </say>

<sayer> The State Department </sayer>

<utterance> there was a `` possibility " that some Nicaraguan rebels were selling their U.S.-supplied arms to Salvadoran guerrillas </utterance>

<sell> selling </sell>

<seller> some Nicaraguan rebels </seller>

<thing_sold> their U.S.-supplied arms </thing_sold>

<buyer> to Salvadoran guerrillas </buyer>

<provide> supply </provide>

<provider> U.S. </provider>

<thing_provided> arms </thing_provided>

<benefactive> to Salvadoran guerrillas </benefactive>

<insist> insisted </insist>

<insister> The State Department </insister>

<thing_insisted> it was n't an organized effort </thing_insisted>

Word Segmentation (in Chinese)

北京大学生体育馆

- 北京 (Beijing) 大学生 (university students) 体育馆 (gym)

The gym for university students in Beijing.

- 北京大学 (Peking University) 生 (give birth to) 体育馆 (gym)

Peking University gave birth to the gym?

Statistical Machine Translation

SMT uses parallel corpora to automatically learn a translation

SOURCE: 目前 , 某些 西方 国家 已经 宣布 终止 对 津巴布韦 的 经济援助 .

H1: at present , some western nations have already announced their
termination of economic aid to zimbabwe .

H2: at present , certain western countries have already suspended their economic
aids to zimbabwe .

H3: so far , some western countries have declared ending economic aid to zimbabwe .

H4: some western countries have already halted economic aid to zinbarbwe at present .

SYSTEM: at present , some western countries have announced the* end* of the*
financial* assistance* to zimbabwe .

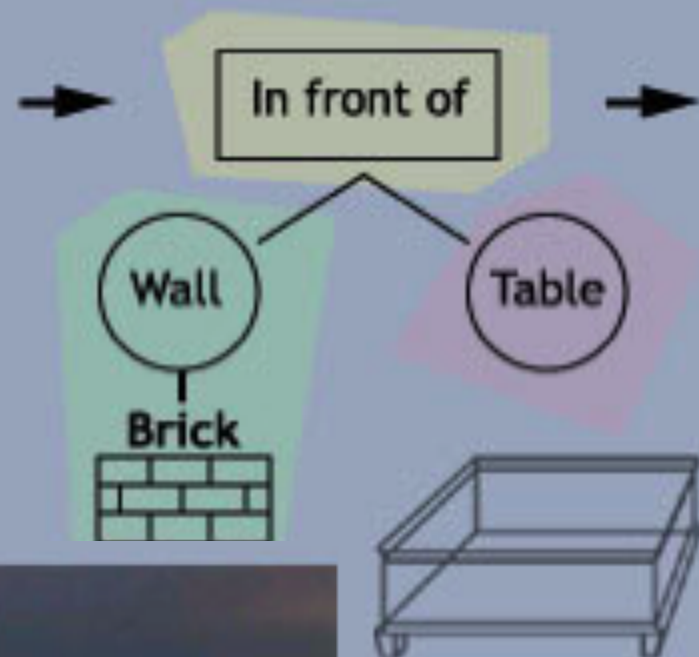
Open Source Machine Translation! www.statmt.org

Holy Grail: Understanding Language

- Can we *generate* language from our knowledge of language?
- Can we convert a natural language utterance into a *model* (or some other fancy logic thing)
- Can we map it into a *database*?
- Can we map it into a *mental picture* (or a *real* one?)
- Demo: WordsEye (from Richard Sproat's group at AT&T)

Text to semantic model to image

The vase is on the Richard Sproat coffee table. The table is in front of the brick wall. The Van Gogh picture is on the wall. The Matisse sofa is next to the table. Mary is sitting on the sofa. She is playing the violin. She is wearing a straw hat.





The Devil is
in the details

Text Mining Support

- TAKMI, by Nasukawa and Nagano, '01
- The system integrates:
 - Analysis tasks (customer service help)
 - Content analysis
 - Information Visualization

Table 2 Analysis of association among [liquid]s and [problem]s

	Damage	Fail	Sticking	Dead	Bad	Freeze
Water	94 (11.1%)	27 (3.19%)	5 (0.59%)	21 (2.48%)	17 (2.01%)	16 (1.89%)
Coffee	31 (6.87%)	12 (2.66%)	13 (2.88%)	7 (1.55%)	6 (1.33%)	5 (1.11%)
Juice	3 (2.94%)	1 (0.98%)	7 (6.86%)	4 (3.92%)	5 (4.9%)	0 (0.0%)
Soda	7 (7.37%)	2 (2.11%)	12 (12.63%)	4 (4.21%)	1 (1.05%)	1 (1.05%)
Tea	3 (7.5%)	1 (2.5%)	1 (2.5%)	0 (0.0%)	0 (0.0%)	1 (2.5%)
Beer	2 (5.88%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (11.76%)

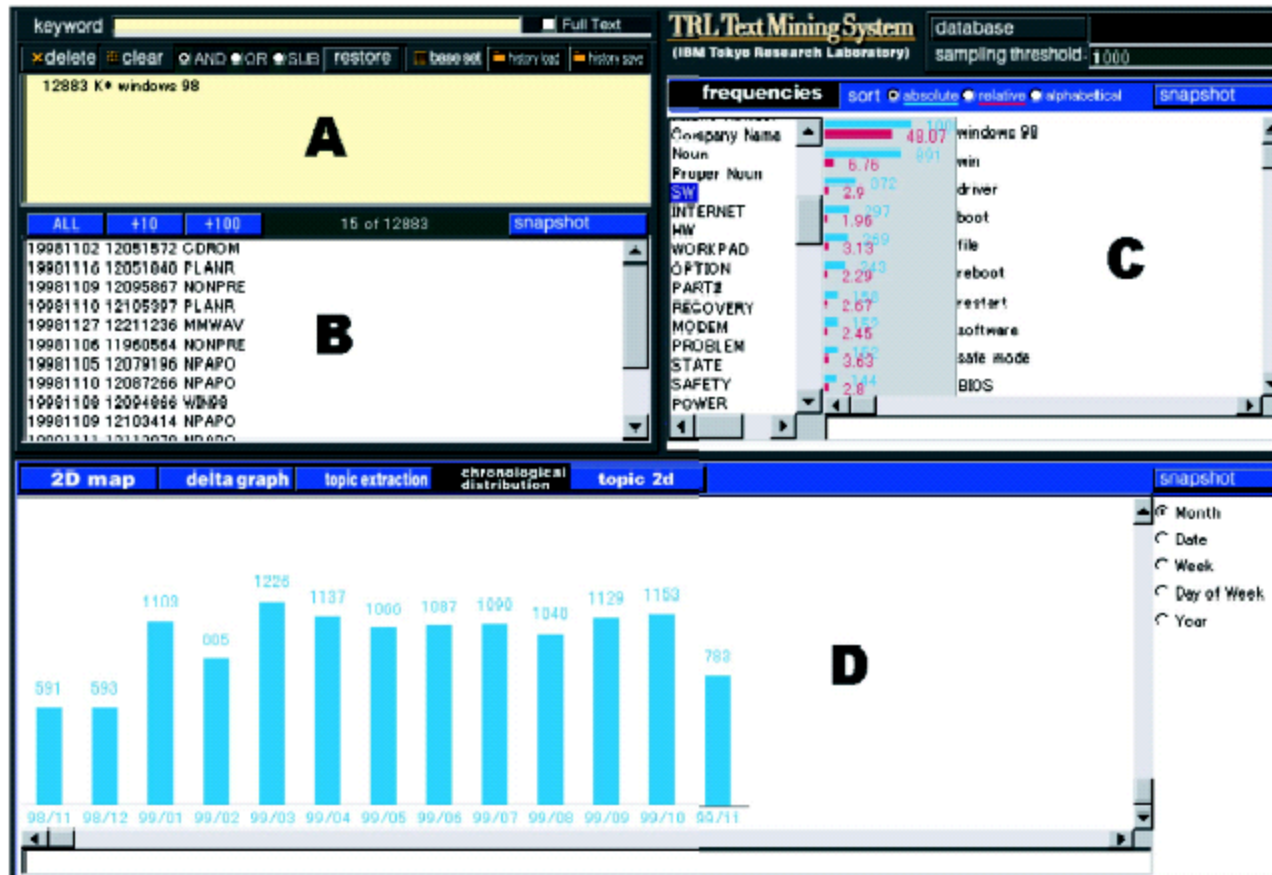
* i247: Information Visualization and Presentation by Marti Hearst

Text Mining

TAKMI, by Nasukawa and Nagano, 2001

- Documents containing “windows 98”

Figure 1 GUI of TAKMI



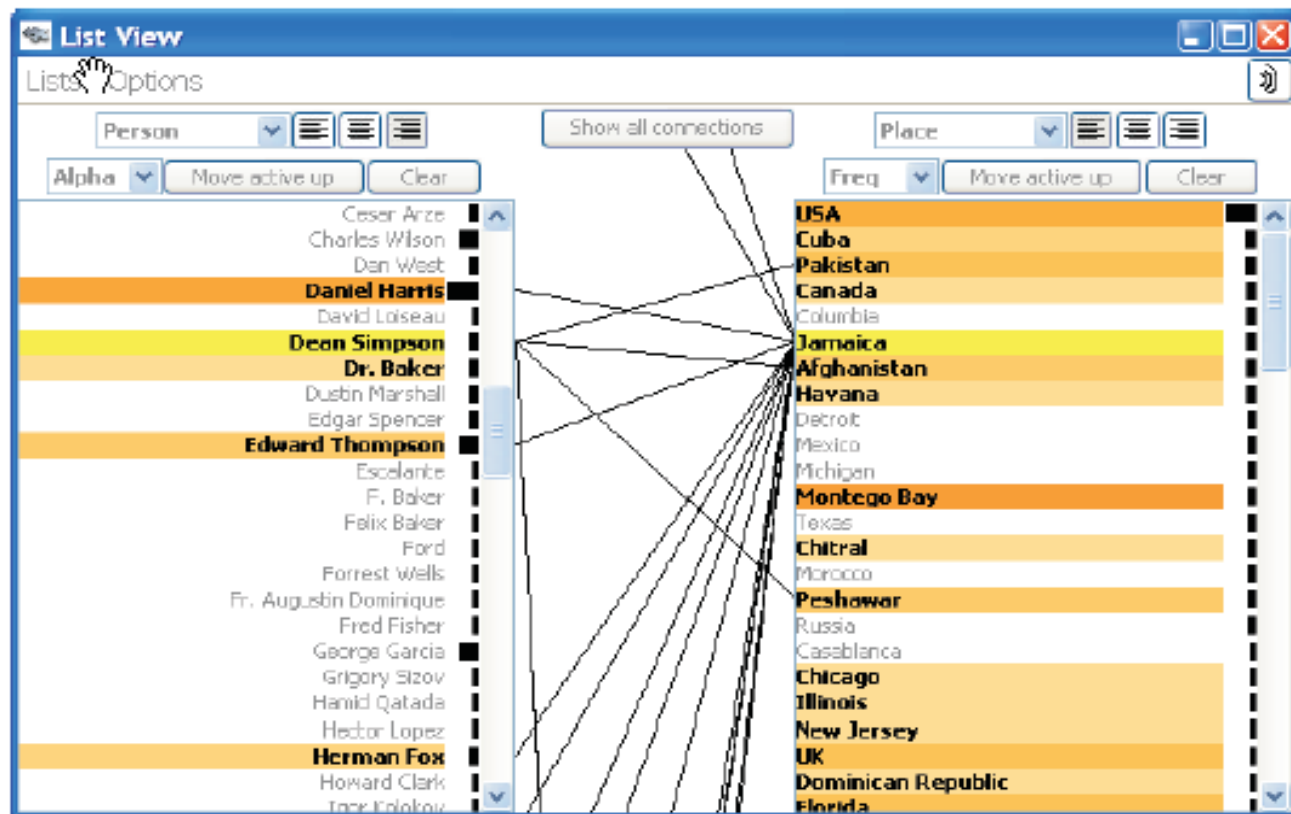
TAKMI, by Nasukawa and Nagano, 2001

- Patent documents containing “inkjet”, organized by entity and year

Figure 11 Topic extraction in [organization names] from 308 patent documents containing the word "inkjet"



Text Mining: Jigsaw by Stasko et al.



Text Mining: WebFountain

WebFountain - Search - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

Beta [advanced](#) [help](#)

Your [Search](#) of (**ibm webfountain**) returned the following results in 5.086 seconds.

[New Search](#) [Refine Document Sample](#)

Names vs. Domains

Built using Chart FX Development

- Sergey Brin
- Carson Pirie Scott
- Chris Sherman
- Darren Friedlander
- Bill Gates
- Gary Price
- Robert Carlson
- San Jose
- Andrew Tomkins
- Dan Gruhl

Names:

Built using Chart FX Development

Emails:

- 44 - webmaster@watson.ibm.com
- 22 - mailroomuk@zdnet.co.uk
- 19 - custserv@infoday.com
- 18 - howard_manus@businessweek.com

1 - 10 of 794 results.
2206 duplicates removed from a sample of 3,000
Approximate Enumeration Size: 12,000

[Next Page](#)

- [John Battelle's Searchblog: WebFountain, the Long Version](#) [\[View Cache\]](#)

<http://battellemedia.com/archives/000428.php>

Enter IBM . » IBM's WebFountain review from Chichi Michi . WebFountain is a classic IBM solution to the search problem . WebFountain the Long Version . IBM PR response is interesting . With WebFountain IBM has sliced the web into subjective structured datasets . John Battelle has been to IBM Almaden and was given the grand tour of the home of IBM Webfountain . [\[49 more\]](#)

Date: May 1, 2006 Rank: 0.9227114
- [John Battelle's Searchblog: Book Related Archives](#) [\[View Cache\]](#)

http://battellemedia.com/archives/cat_book_related.php

Enter IBM . WebFountain Technorati Visits . So Why WebFountain Why Now . WebFountain is a classic IBM solution to the search problem . WebFountain the Long Version . With WebFountain IBM has sliced the web into subjective structured datasets . IBM v Google The Chart . For more info on WebFountain Gary Price has created these links . [\[39 more\]](#)

Date: Feb 5, 2005 Rank: 0.88429064
- [How to build a WebFountain: An architecture for very large-s...](#) [\[View Cache\]](#)

<http://www.research.ibm.com/journal/sj/431/gruhl.html>

About IBM Privacy Terms of use Contact . The process of loading data into WebFountain is referred to as ingestion . The most challenging future problems for WebFountain lie in the mining space . WebFountain is a platform for very large scale text analytics applications . [\[30 more\]](#)

Date: Oct 4, 2004 Rank: 0.86236644
- [IBM sets out to make sense of the Web | CNET News.com](#) [\[View Cache\]](#)

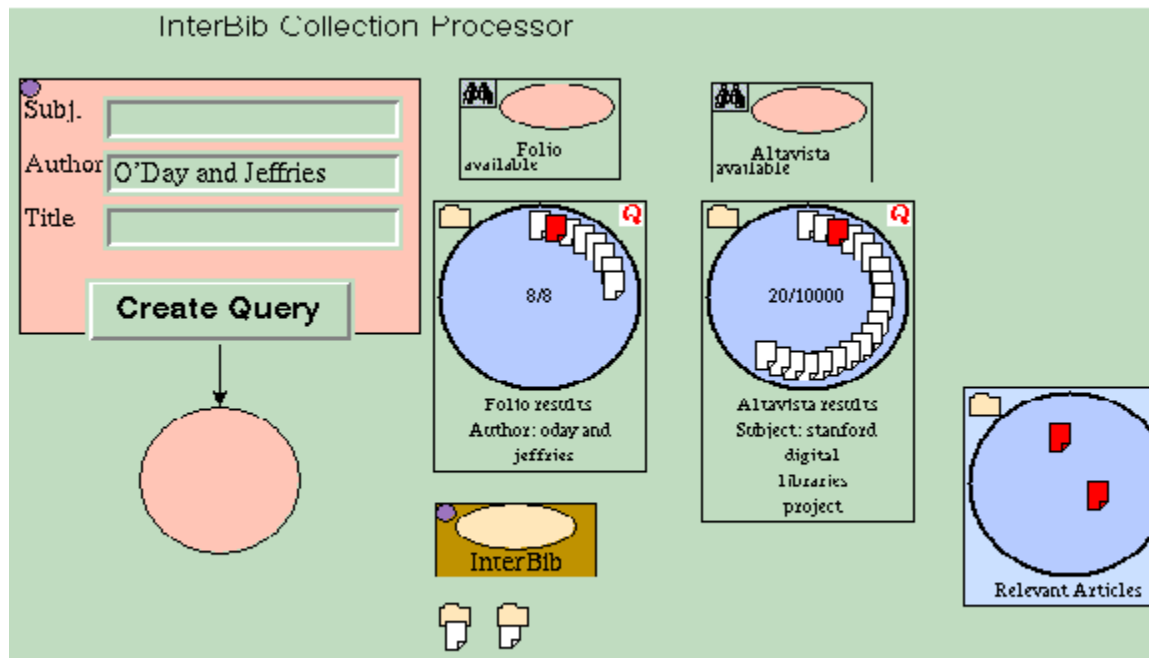
<http://news.com.com/2100-1032-5153627.html>

Salesforce.com sold on IBM Sybase technologies . With that insight WebFountain was born . IBM thinks Fast for bioscience searches . Trial over safety at IBM in jurors hands . By contrast IBM's WebFountain wants to help find meaning in the glut of online data . [\[29 more\]](#)

Date: Feb 5, 2004 Rank: 0.8489776
- [Feature Article: Cover Story](#) [\[View Cache\]](#)

Visualization Support for SenseMaking

- DLITE by Cousins et al. '97



Visualization in Sensemaking

TRIST (The **R**apid **I**nformation **S**canning **T**ool) is the work space for Information Retrieval and **I**nformation **T**riage.

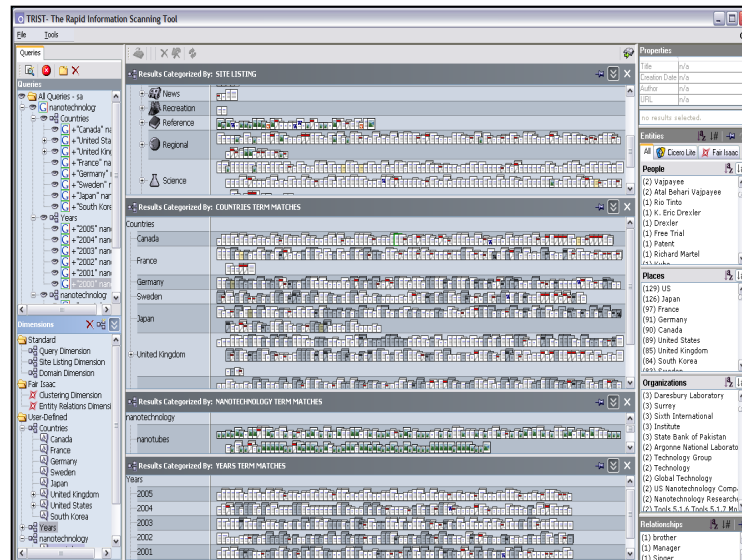
TRIST, Jonkers et al 05 User Defined and Automatic Categorization

Launch Queries

Query History

Dimensions

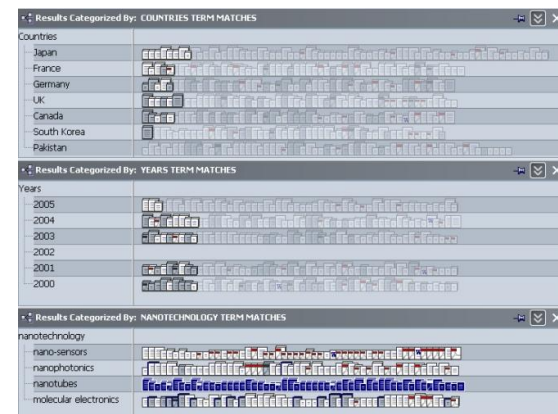
Annotated Document Browser



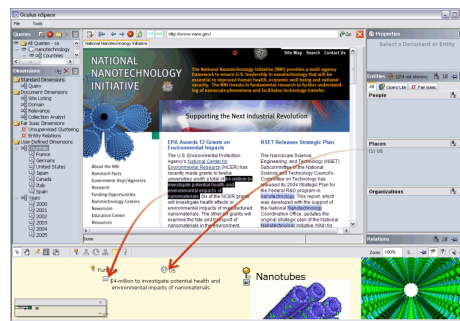
Comparative Analysis of Answers and Content

Rapid Scanning with Context

Entities

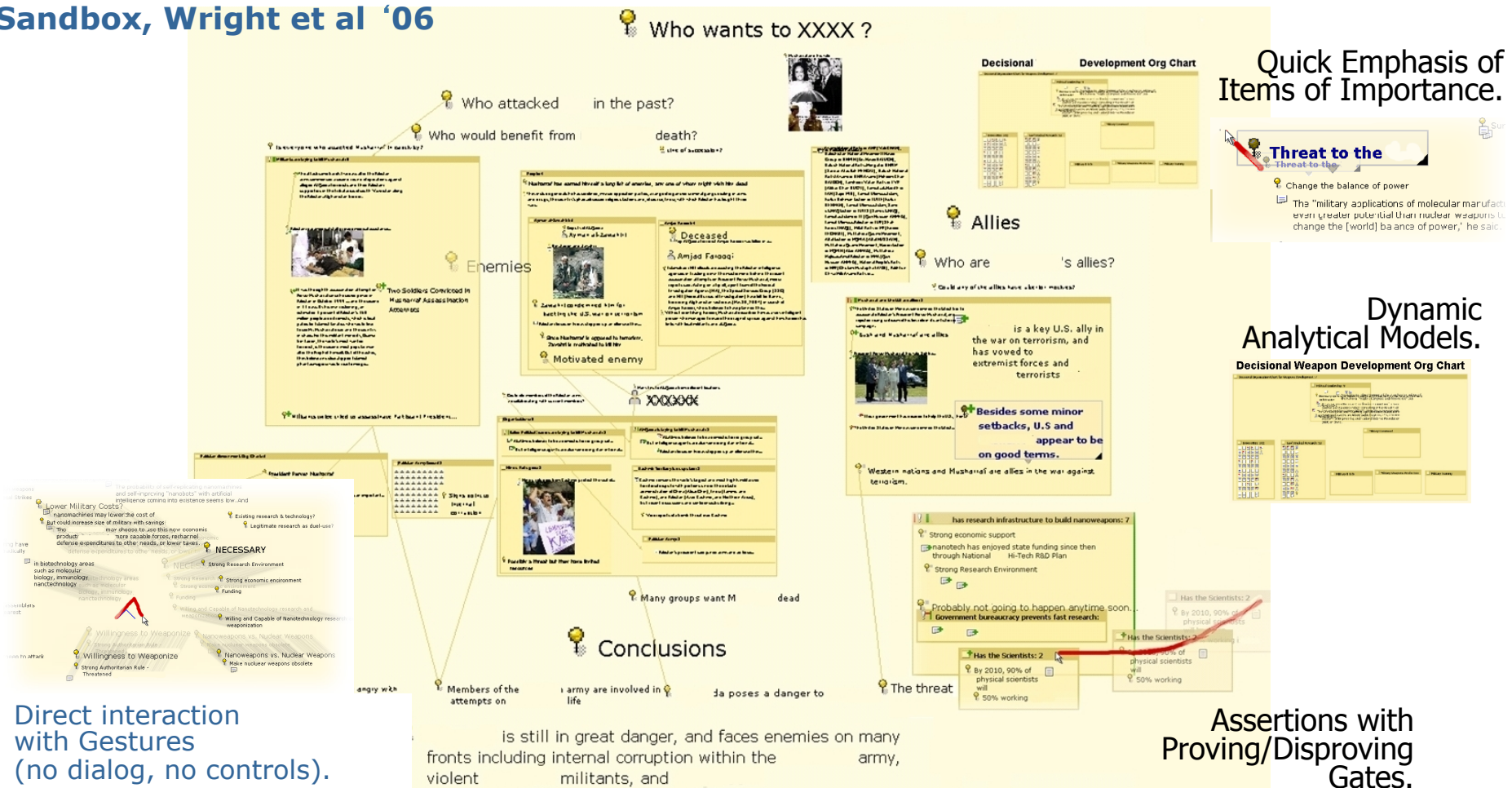


Linked Multi-Dimensional Views Speed Scanning



Visualization for Sensemaking

Sandbox, Wright et al '06



Concordances & Word Frequencies

Concordance - Larkin.Concordance

File Text Search Edit Headwords Contexts View Tools Help

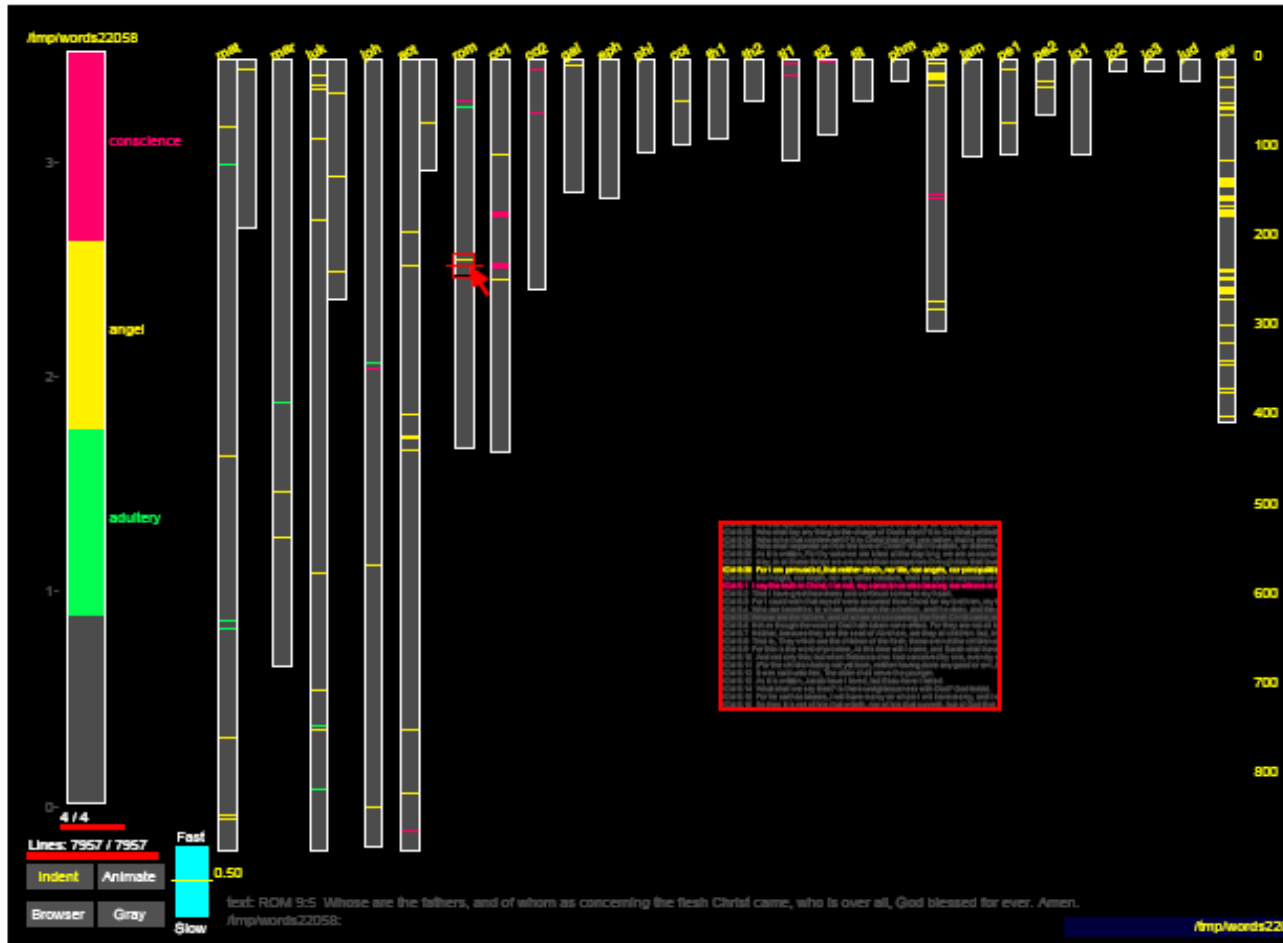
Headword No. Context... Word ...Context Reference

HEAR	15	That my own	heart	drifts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart		Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed u
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spa
HEATH	1	This petrified	heart	has taken,	A Stone Churc
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra
HEAVE	1	Hands that the	heart	can govern	Heaviest of flo
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk
HEAVIER-THAN-...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away the

Words: 7318 Tokens: 37070 At word: 2990 Deleted lines: 1 [24] Word sort: Asc alpha (string) Context sort: Asc occurrence order

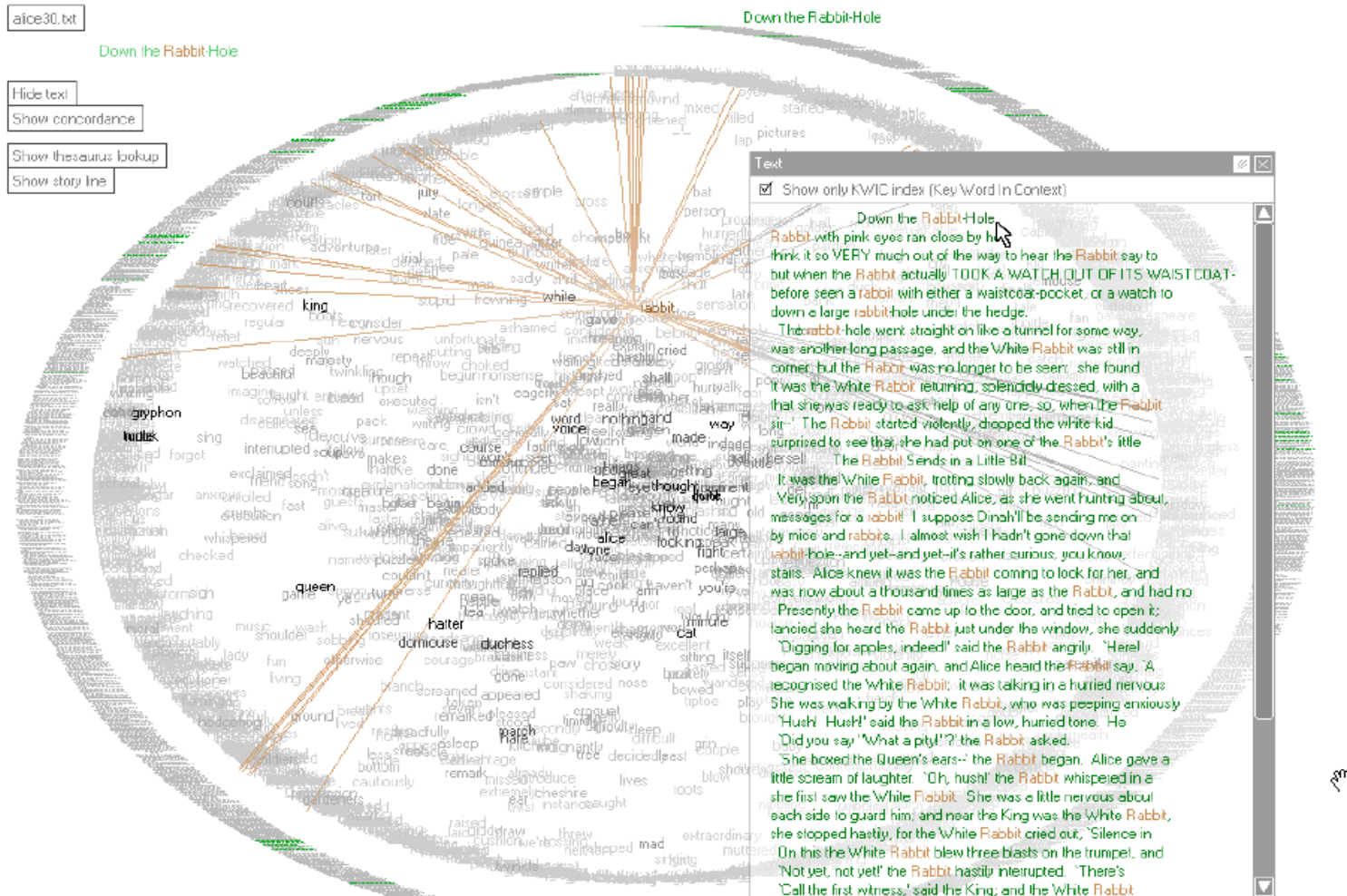
From www.concordancesoftware.co.uk

Concordances & Word Frequencies



SeeSoft by Eick et al.

Concordances & Word Frequencies



TextArc by Paley.

Concordances & Word Frequencies

Visualizations : Bubble Chart: Top 100 Words 19th Century Fiction Without stopwords

Can't see the visualization? Download the latest Java plugin [here](#). On Macs: best viewed in Safari.

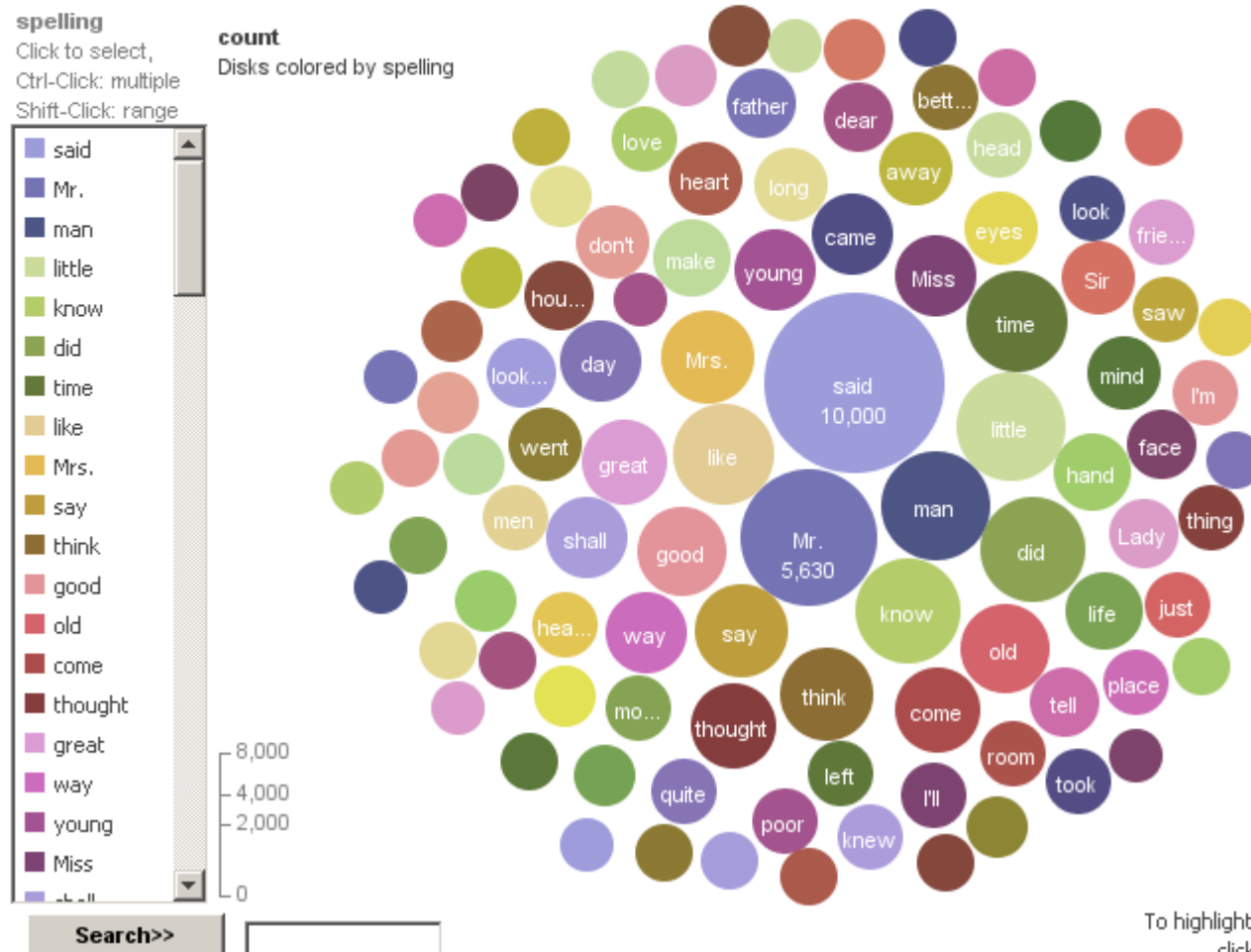
Created by: Amit Created on: Monday February 11, 1:58 AM

spelling

- Click to select,
- Ctrl-Click: multiple
- Shift-Click: range

count

Disks colored by spelling



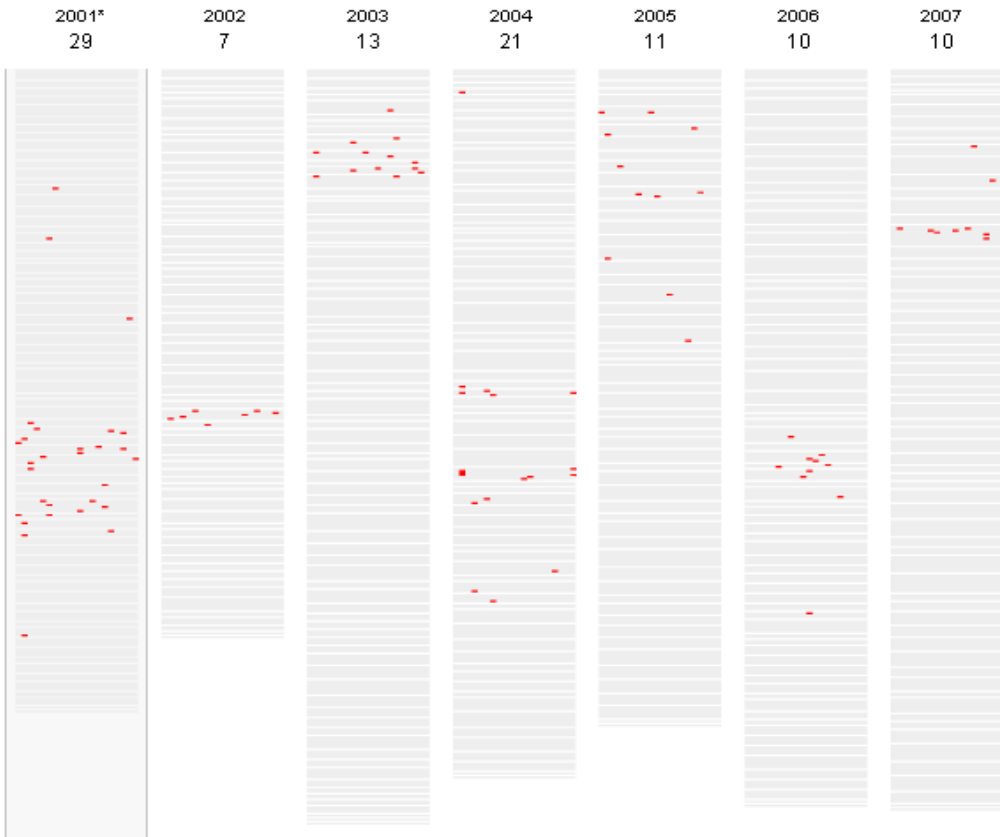
To highlight or find totals
click or ctrl-click.

Bubble Charts (implemented by Wattenberg)

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

Use of the phrase "Tax" in past State of the Union Addresses



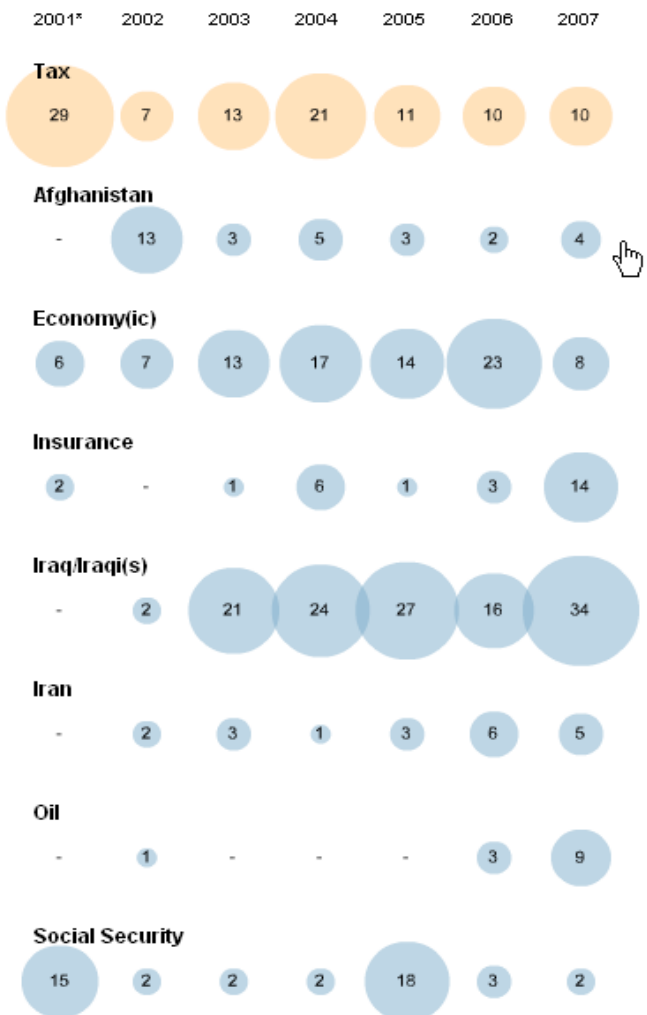
The word in context

I believe in local control of schools. We should not, and we will not, run public schools from Washington, D.C. Yet when the federal government spends **TAX** dollars, we must insist on results. Children should be tested on basic reading and math skills every year between grades three and eight. Measuring is the only way to know whether all our children are learning. And I want to know, because I refuse to leave any child behind in America.

-- 2001 (Paragraph 14 of 73)

[Next Instance of 'Tax'](#)

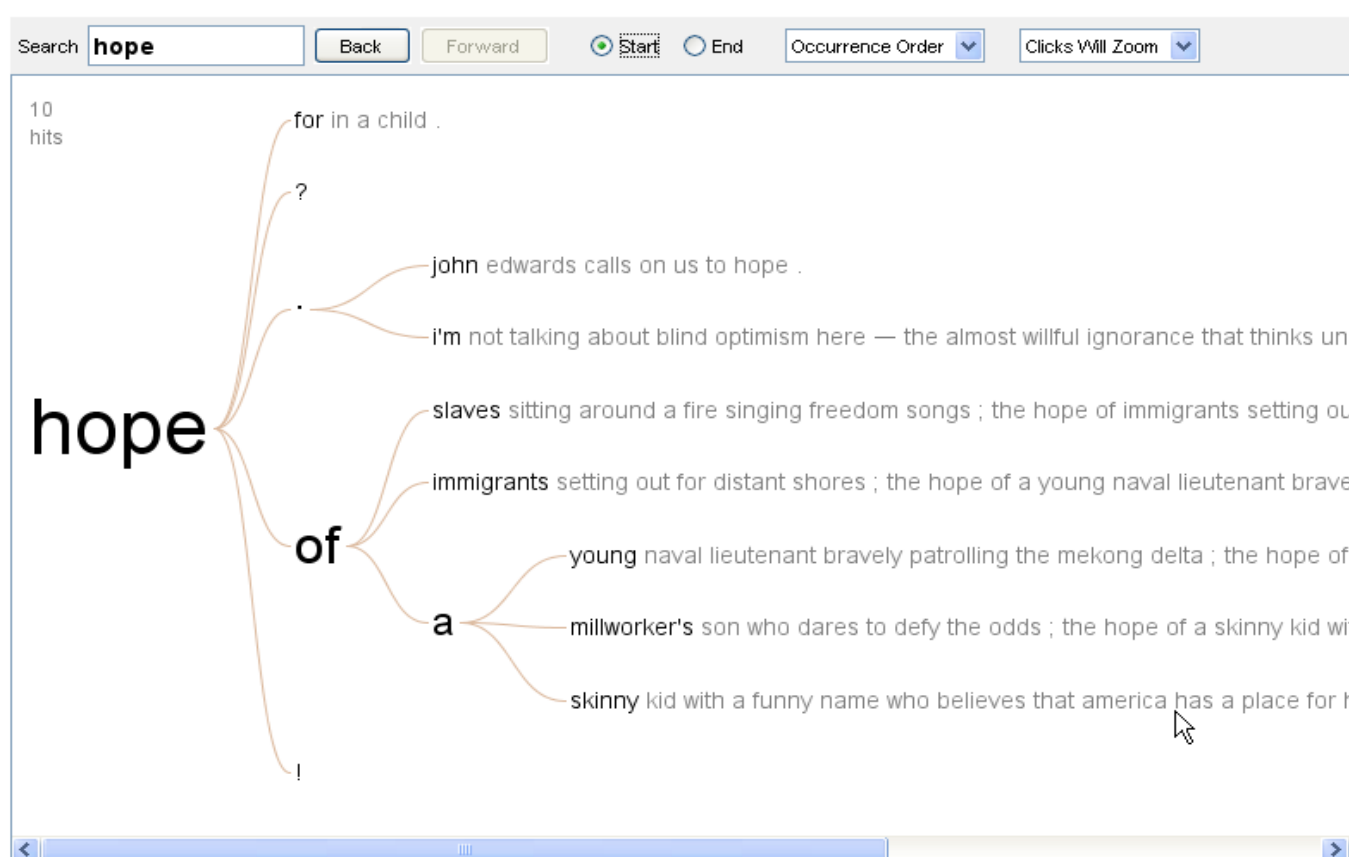
Compared with other words



Putting it together: Werschkul of the NYTimes

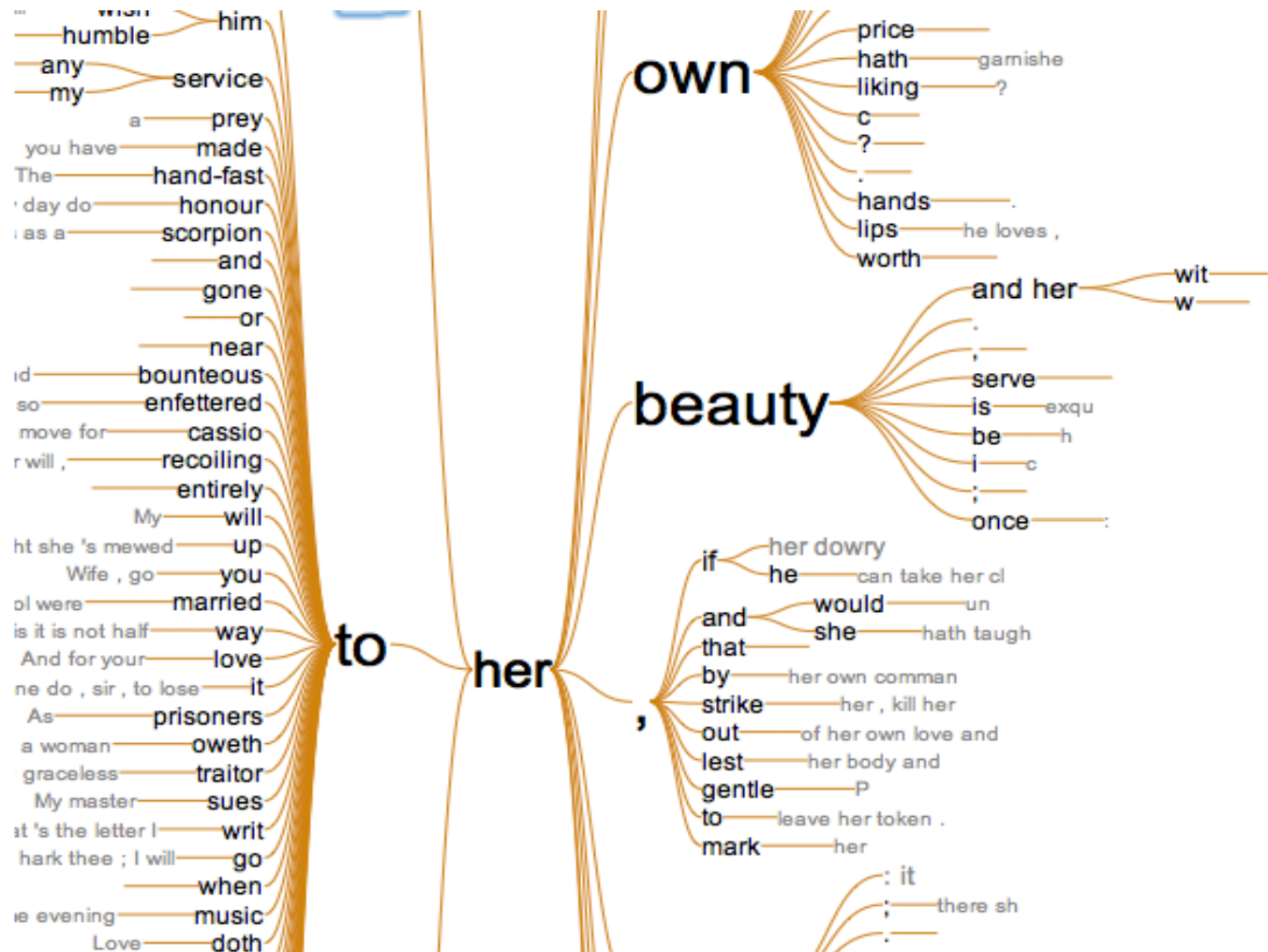
president, Mr. Bush did not deliver a formal State of the Union. His Feb. 27 speech to a joint session of Congress was a State of the Union, but without the title.

Concordances & Word Frequencies



WordTree by Wattenberg

WordSeer (Hearst et al 2013)



Word

Search for relationships

search

✓ (any relation to)

described as

done by

done to

because

and

possessed by

in order to

with

to

from

of

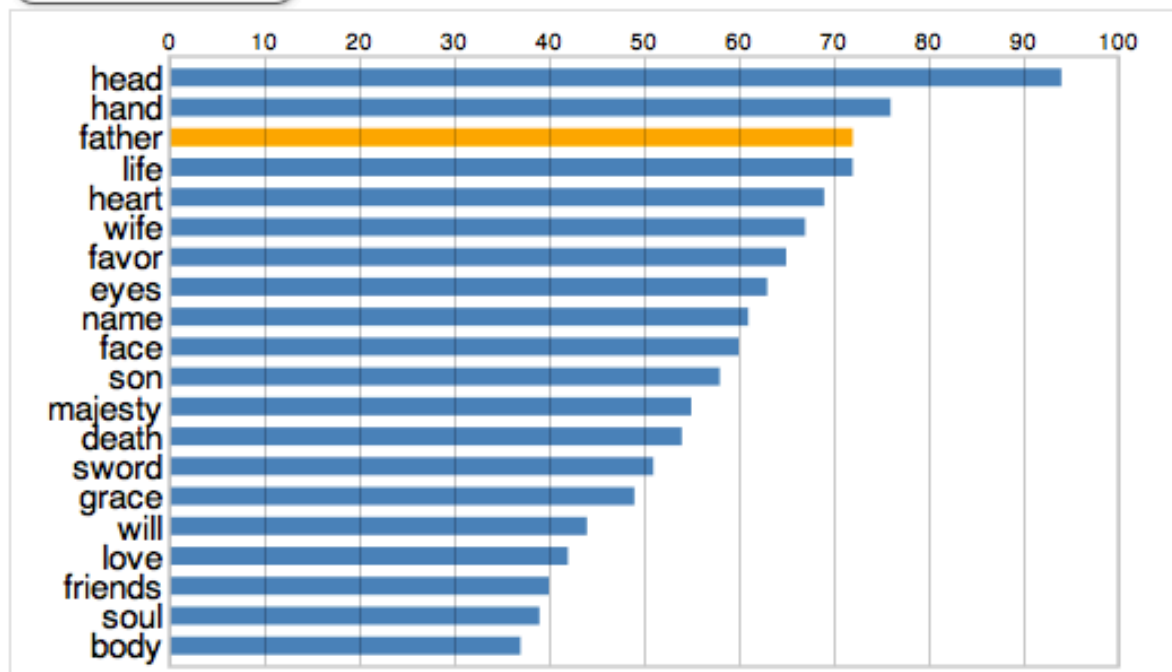
on

by




in

within collection (all documents) Go

Show a Random Sent

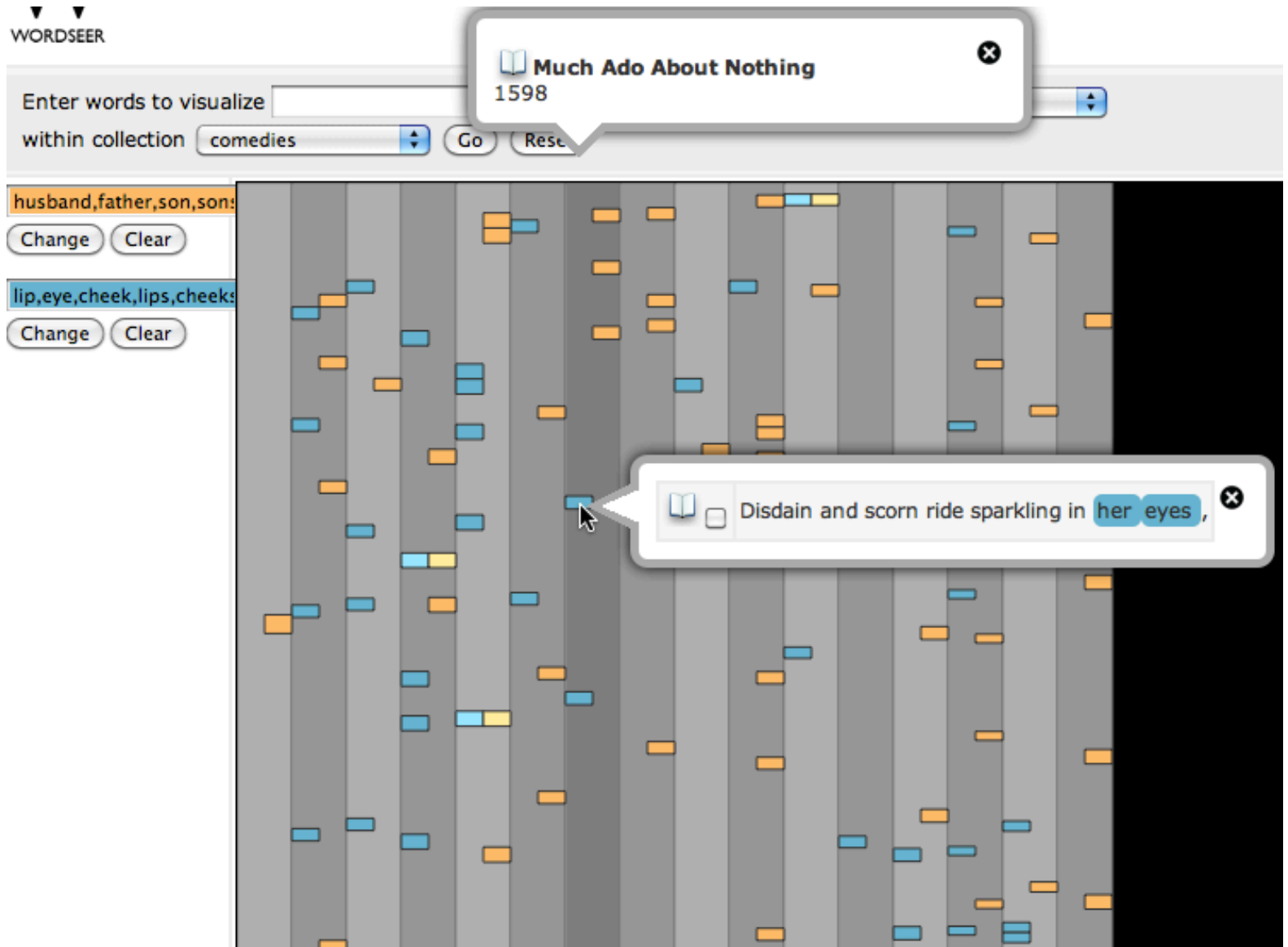


72 results

Select	Sentence	Type	Title	Author	Date	Publisher	Place Published
all <input type="checkbox"/>							
 <input type="checkbox"/>	But that I think his father loves him not	lines	The First Part of King Henry the Fourth	William Shakespeare	1597		
 <input type="checkbox"/>	So shall his father's wrongs be recompensed .	lines	The First Part of King Henry the Sixth	William Shakespeare	1591		
 <input type="checkbox"/>	And as his father here was	lines	The First Part of King	William	1591		

WordSeer (Hearst et al 2013)

Newspaper strip
vis



Definition

*Tag Cloud: A **visual** representation of social tags, organized into paragraph-style layout, usually in **alphabetical** order, where the **relative size** and **weight** of the font for each tag corresponds to the **relative frequency** of its use.*

On the positive side:

- Compact
- Draws the eye towards the most frequent (important?) tags
- You get three dimensions simultaneously!
 - alphabetical order
 - size indicating importance
 - the tags themselves

Weirdnesses

- Violates principles of perceptual design
 - Longer words grab more attention than shorter
 - Length of tag is conflated with its size
 - White space implies meaning when there is none intended
 - Ascenders and descenders can also effect focus
 - Eye moves around erratically, no flow or guides for visual focus
 - Proximity does not hold meaning
 - The paragraph-style layout makes it quite arbitrary which terms are above, below, and otherwise near which other terms
 - Position within paragraph has saliency effects
 - Visual comparisons difficult (see Tufte)

Weirdnesses

All time most popular tags

- Meaningful associations are lost
 - Where are the different country names in this tag clouds?

06 africa amsterdam animals architecture art asia august australia autumn baby barcelona
beach berlin birthday black blackandwhite blue boston bw california
cameraphone camping canada canon car cat cats chicago china christmas
church city clouds color concert d50 day dc dog england europe fall family
festival film florida flower flowers food france friends fun garden
geotagged germany girl graffiti green halloween hawaii hiking holiday home
honeymoon hongkong house india ireland island italy japan july kids la lake landscape
light live london losangeles macro march me mexico mountain mountains museum music
nature new newyork newyorkcity newzealand night nikon nyc ocean paris
park party people portrait red river roadtrip rock rome san sanfrancisco
scotland sea seattle show sky snow spain spring street summer sun sunset
sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa
vacation vancouver washington water wedding white winter yellow york zoo

What are tags?

You can give your photos a "tag", which is like a keyword or category label. Tags help you find photos which have something in common. You can assign up to 70 tags to each photo.

Weirdnesses

Which operating systems are mentioned?

This is a **tag cloud** - a list of tags where size reflects popularity.
sort: alphabetically | [by size](#)

.net ajax apple architecture **art** article articles audio bit200w07 **blog** **blogs** **books**
business code comics **community** computer cooking cool **css** culture database **design**
development diy download ebooks **education** entertainment environment fashion fic **finance**
firefox **flash** flickr fonts **food** forum **free** **freeware** **fun** **funny** game **games** **google**
graphics green gtd hardware health **history** home **howto** html **humor** illustration images
imported **inspiration** internet it japan **java** **javascript** jobs language library lifehacks
linux **mac** magazine maps marketing media mobile money movies mp3 **music** **news**
online **opensource** **osx** photo **photography** photos photoshop php plugin podcast
politics portfolio productivity **programming** python radio rails recipes **reference** religion
research resources rss **ruby** rubyonrails **science** **search** **security** seo **shopping** slash
social **software** sports tech **technology** **tips** **tools** toread **travel** **tutorial** tutorials tv
twitter typography ubuntu **video** videos **web** **web2.0** **webdesign** webdev wiki wikipedia
windows wishlist **wordpress** writing youtube

(red tags are tags you share with everyone else)

Alternative: “Semantic” Layout

- Improving Tag-Clouds as Visual Information Retrieval Interfaces, Hassan-Montero & Herrero-Solana, InSciT2006
- Tags grouped by “similarity, based on clustering techniques and co-occurrence analysis”

ajax apple art article audio blog blogging blogs books business code comics community computer cool
 css culture daily del.icio.us delicious design development diy firefox flash flickr free freeware fun
 funny games geek google graphics gtd hacks hardware history howto html humor images internet
 java javascript language lifehacks linux mac maps media movies mp3 music news opensource
 osx photo photography photos php politics productivity programming python rails reference
 research rss ruby science search security shopping social software tech technology tips tool tools
 toread travel tutorial tutorials usability video web web2.0 webdesign webdev wiki windows writing xml

Figure 1: Traditional Tag-Cloud. Tags have been selected and visually weighted according to its frequency of use.

4 RESULTS

lisp perl python ruby rails
 database wordpress fonts wiki gtd
 books writing language math science philosophy religion history politics
 media news blog blogs internet technology business web2.0 rss search gcogle
 firefox accessibility usability php xml ajax javascript html css webdesign
 design web reference howto tutorial java programming development tools software opensource free
 windows linux unix security networking hardware apple mac osx
 game games fun funny humor art photography flash animation com cs
 cinema film movies movie video tv
 audio music mp3 ipod radio podcast podcasting
 mobile treo psp xbox fashion shopping
 travel food health marketing advertising

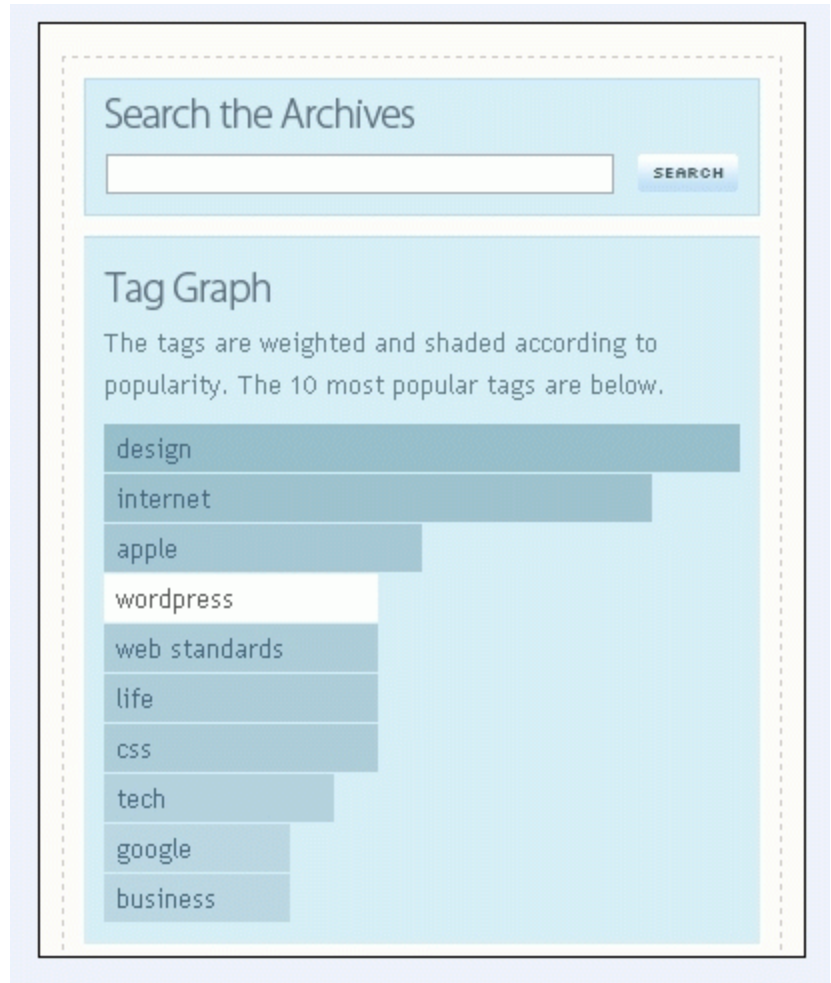
Figure 2: Improved Tag-Cloud. Tags have been selected and visually weighted according to function 1.

Tag Cloud Alternatives

Provided by Martin Wattenberg

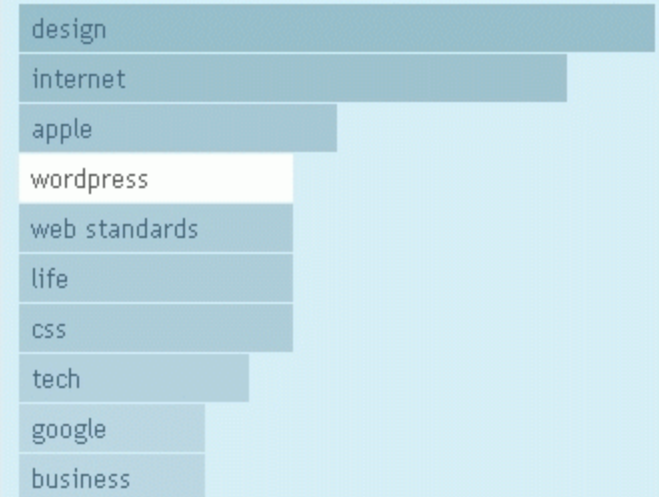
Ofresh architecture **art** artist
bio blog census chi class
cogsci color company
constraint **cs** **data** dataportrait
design diagram enron ethics
example eyetracking
framework gis golan google
graph graphics hci history
homepage hypergraph ibm
infovis infovisgroup ir java
language law map math music
network newmedia news
numbers person portrait
psych publicrecords rdf sna
software softwarevis stats
stream tech text toolkit
treemap ui venn video vision
visualization visualreasoning
vna web wiki

abcd	391	■
archived	14	
articles	1369	■
bjadn.the	13	
categories	19	
cleanup	15	
copyright	28	
jesus	47	
list	137	■
mind	32	
redirects	53	
requested	17	
requests	220	■



Tag Graph

The tags are weighted and shaded according to popularity. The 10 most popular tags are below.





Lensing Wikipedia

<http://lensingwikipedia.cs.sfu.ca>

Anoop Sarkar
Maryam Siahbani
Max Whitney
Ravikiran Vadlapudi
Rohit Dholakia

SFU Natural Language Lab

<http://natlang.cs.sfu.ca>

Text is tough (to visualize)

- With text, it is easy to extract and show the main trends
- But in text analytics we often want to highlight the rare but unexpected and important event

Explore new visualizations that exploit parsed language

Lensing Language

- Semantic parsing of natural language: going beyond topic models and clustering bags of words
- Exploit language understanding: *who* did *what* to *whom*, *where*, *when* and *how* ...
- "Embodied" visualization: place spatial, temporal and social entities into an intuitive low dimensional space

Lensing Wikipedia

- Provide a summary visualization of all of history ... as represented in Wikipedia
- The information is all in natural language (English)
- The task: **query based visual summarization of history events (in Wikipedia)**
 - e.g. “Describe Roman interactions with Carthage between 200BC and 15BC”
 - e.g. “Between 500 BCE and 2012 CE what events occurred in Siberia”
 - e.g. what was created, burned, bought, sold in a particular region or at a particular time?

Lensing Wikipedia

- Web crawl of wikipedia:
- select all pages that summarize events that happened in a year or decade, e.g.



Article [Talk](#)

1470s

From Wikipedia, the free encyclopedia

This is a list of events occurring in the 1470s, ordered by year.

Contents [\[hide\]](#)

[1470](#) • [1471](#) • [1472](#) • [1473](#) • [1474](#) • [1475](#) • [1476](#) • [1477](#) • [1478](#) • [1479](#)

1470

January–December

- [March 12](#) – [Wars of the Roses](#) – [Battle of Lose-coat Field](#): The [House of York](#) defeats the [House of Lancaster](#).
- [May 15](#) – [Charles VIII of Sweden](#), who had served three terms as [King of Sweden](#), dies. [Sten Sture the Elder](#) becomes Regent of Sweden.
- [October](#) – A rebellion orchestrated by King [Edward's](#) former ally, the [Earl of Warwick](#), forces the King to flee England to seek support from his brother.
- [October 30](#) – Warwick releases [Henry VI of England](#) from the Tower and restores him to the throne.

Natural Language Processing Pipeline

- Tokenization (splitting up into word tokens)
- Part of speech tagging (nouns, verbs, ...)
- Parsing (finding syntactic structure: noun phrases, verb phrases, ...)
- Finding person, location names (person=Pericles, location=Rome, ...)
- Semantic annotation: Finding predicates like *start* and arguments like *Athenion* (**initiator**) and *slave rebellion* (**thing_started**)
- Pronoun referents and noun phrase coreference

Natural Language Processing Pipeline

- End result: we crawl thousands of Wikipedia pages (the precise number varies)
- We can find the time for each event easily in this dataset (the URL contains the year)
- About 64K events can be assigned a geo-location out of 82K events in total.
- Our current dataset visualizes the entire 64K extracted events from Wikipedia history summary pages.

Lensing Wikipedia

104BC: Athenion starts a slave rebellion in Segesta

First Step: Parsing Wikipedia

Second Step: Project spatially

Predicate

Arg0: initiator

Arg1: thing_started

Location

(Sicily: 2D Plot, lat=37.93; long=12.83)

Time

Plot 104BC on a 1D timeline

```
{  
  "arg0": "Mina de Ouro",  
  
  "arg1": "the chief center for the gold trade and a major source of revenue  
for the crown",  
  
  "description": "Portuguese sailors reach Mina de Ouro on the Gold Coast  
(present-day Ghana) and explore Cape St. Catherine, two degrees south of the  
equator. Mina de Ouro becomes the chief center for the gold trade and a major  
source of revenue for the crown.",  
  
  "event": "become",  
  
  "latitude": 5.5499977999999999,  
  "longitude": -0.249999,  
  
  "roleArg0": "Agent",  
  "roleArg1": "entity_changing",  
  
  "title": "Ghana",  
  
  "year": 1471  
}
```

- **Show the text as soon as possible.** (click on Sparta, chronologically arranged events shown to user)
- NLP is hidden from the user. No parses shown
- All views are always synchronized.
- Information is assumed to be verb-centric. (20K verbs in our 64K event dataset)
- **Map:** different views (flat, globe, butterfly). Toggle to select one cluster, drag to select many. Pan to move map around.
- **Timeline.** Shows global timeline and local selection of time interval simultaneously.

- **Faceted browsing.** Each list is a facet. Choices in the list are added as a constraint. Constraints can be removed in any order. Sparta, entity_refusing. Remove Sparta, add Texas.
- **Location Facet.** All the locations identified in the data as playing a role in some event. Italy v.s. Rome.
- **Current Country Facet.** Names of contemporary countries by reverse lookup of geo-locations.
- **Role Facet.** Taken from the semantic role labels. Underlying parse structure is not shown to the user.
- **Group By:** Facets can also be used to group results into a two dimensional grid of events. Narrows down what you read attentively. Texas, entity_refusing => thing_tried
- Search box to permit text-based search.

- Person facet. Names of people in the dataset (automatically identified). Clear all, select Ptolemy.
- Timeline view. Restrict to a time interval: Select 350BC to 250BC. Move entire selection rightwards to 325BC-225BC.
- All views and facets are synchronized.
- Can clear constraints out of order just like in faceted browsing (Marti Hearst).
- Clear all constraints and start again.
- Typically can find surprising facts about history in about 5 to 6 interactions with this interface.
- Try it out! It's on the web:
 - <http://lensingwikipedia.cs.sfu.ca>

What next?

- Multiple faceted lists can be used. More or less hierarchical.
- Some ideas for evaluation of a visual browser:
 - Exploit the fact that it is on the web. Now. No need to distribute or install anything.
 - Potentially large number of users to test text vis ideas.
 - Use multivariate analysis on the web site.
 - Track usage of different facets.
 - Track time to find an “interesting” page on Wikipedia.
 - Try to attract a large number of users.