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
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, ...

Natural Languages
French, English, Korean, Chinese, Tagalog, ...

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

- 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?’
  – kynässänsääkäänkö ‘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

Children’s comprehension of relative clauses.
De Villiers et al. J. Psych Res 8(5) 2005
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 \textsubscript{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

“Generative” Model

Algorithm

Application to Natural Language

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
New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.
The New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.
1) Select "valid" if the passage contains strong evidence of an experimentally determined localization.

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.

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)

<table>
<thead>
<tr>
<th>Protein: ExbB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>Invalid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organism: Escherichia coli</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>Invalid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location: cytoplasmic membrane</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>Invalid</td>
</tr>
</tbody>
</table>
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 Quebeckers 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. ...
The control panel looks the same but responds more quickly to commands and menu choices.
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

10 Happiest Tweets

- @WRiTEMiND 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: Selamatt ulang tahun yaaa RT @Phitz_bow: Selamat siangg RT @isan_coy: Slamat pagiiii
Sentiment detection


### 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 buitin :
- 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 gatel 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.

Example:

[Boeing Inc.] reached agreements to sell [its remaining seven aircraft] to buyers that weren't disclosed.

- **A0:** seller
- **A1:** goods
- **A2:** buyer
Architecture of a SRL system
Architecture of a SRL system

• On a given parse tree, run the pruning component: some candidate spans are potential arguments, the others are labeled NONE.

• Run a binary classifier for identification: some spans labeled ARG and the rest NONE.

• Run binary classifiers for classification: A0 vs not-A0, A1 vs not-A1, etc. on the nodes labeled ARG.

• Combine output of binary classifiers using 1 vs all: for each ARG node pick the binary classifier with highest confidence and decide the label of each node: A0, A1, A2, A3, A4, …, or AM.
## Accuracy of Semantic Role Labeling

<table>
<thead>
<tr>
<th></th>
<th>SRL</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
<td>F1</td>
</tr>
<tr>
<td>Overall</td>
<td>81.90</td>
<td>78.81</td>
<td>80.32</td>
</tr>
<tr>
<td>A0</td>
<td>88.37</td>
<td>88.91</td>
<td>88.64</td>
</tr>
<tr>
<td>A1</td>
<td>81.50</td>
<td>81.27</td>
<td>81.38</td>
</tr>
<tr>
<td>A2</td>
<td>73.44</td>
<td>68.74</td>
<td>71.01</td>
</tr>
<tr>
<td>A3</td>
<td>75.00</td>
<td>55.49</td>
<td>63.79</td>
</tr>
<tr>
<td>A4</td>
<td>74.74</td>
<td>69.61</td>
<td>72.08</td>
</tr>
<tr>
<td>A5</td>
<td>100.00</td>
<td>80.00</td>
<td>88.89</td>
</tr>
<tr>
<td>AM-*</td>
<td>78.19</td>
<td>69.98</td>
<td>73.86</td>
</tr>
<tr>
<td>R-AM-*</td>
<td>73.91</td>
<td>61.44</td>
<td>67.10</td>
</tr>
</tbody>
</table>
Sen. Mitchell added that the agreement requires that the Contras not initiate any military action.

The State Department said 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 wasn't an organized effort </thing_insisted>
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?
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 zimbabwe 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)
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

* i247: Information Visualization and Presentation by Marti Hearst
Text Mining
TAKMI, by Nasukawa and Nagano, 2001

- Documents containing “windows 98”
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”

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CANON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUJITSU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUJI XEROX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RICOH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEIKO EPSON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BROTHER IND.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HITACHI IND.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KAO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FUJI FILM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HITACHI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MURATA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MITSUBISHI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOKYO ELECTRICITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEIKOHSAYA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Text Mining: Jigsaw by Stasko et al.
Text Mining: WebFountain

Your Search of ( ibm webfountain ) returned the following results in 5.086 seconds.

1. John Battelle's Searchblog: WebFountain, the Long Version  [view Cache]
   Enter 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 subject-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, 2005  Rank: 0.9227114

2. John Battelle's Searchblog: Book Related Archives  [View Cache]
   Enter IBM's 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 subject-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: 0.89429064

3. How to build a WebFountain: An architecture for very large-s...  [View Cache]
   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: 0.86230644

4. IBM sets out to make sense of the Web | CNET News.com  [View Cache]
   Salesforce.com sold on IBM Sybase technologies. With that insight WebFountain was born. IBM thinks Fast for bioscience searches. Trail over safety at IBM in jurors hands. By contrast IBM's WebFountain wants to help find meaning in the glut of online data. [28 more]
   Date: Feb 5, 2004  Rank: 0.8489775

5. Feature Article: Cover Story  [View Cache]
Visualization Support for SenseMaking

- DLITE by Cousins et al. ‘97
Visualization in Sensemaking

TRIST (The Rapid Information Scanning Tool) is the work space for Information Retrieval and Information Triage.

TRIST, Jonkers et al 05 User Defined and Automatic Categorization

Comparative Analysis of Answers and Content

Rapid Scanning with Context

Entities

Linked Multi-Dimensional Views Speed Scanning

Annotated Document Browser

Launch Queries

Query History

Dimensions
Visualization for Sensemaking

Sandbox, Wright et al '06

Who wants to XXXX?

- Who attacked in the past?
- Who would benefit from death?
- Motivated enemy

Allies

Who are 's allies?

Enemies

- Sponsors
- Enemies
- Motivated enemies

Conclusions

- It is still in great danger, and faces enemies on many fronts including internal corruption within the army, violent militants, and...

Quick Emphasis of Items of Importance.

Dynamic Analytical Models.

Assertions with Proving/Disproving Gates.

Direct interaction with Gestures (no dialog, no controls).
Concordances & Word Frequencies

From www.concordancesoftware.co.uk
Concordances & Word Frequencies

SeeSoft by Eick et al.
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

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 analyse what Mr. Bush has said.

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

Compared with other words

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.

-- 2007 (Paragraph 14 of 73)

Putting it together: Werschkul of the NYTimes
Concordances & Word Frequencies

Search: hope

John Edwards calls on us to hope.
I'm not talking about blind optimism here — the almost willful ignorance that thinks un
slaves sitting around a fire singing freedom songs; the hope of immigrants setting ou
immigrants setting out for distant shores; the hope of a young naval lieutenant brave
young naval lieutenant bravely patrolling the Mekong Delta; the hope of a millworker's
son who dares to defy the odds; the hope of a skinny kid with
skinny kid with a funny name who believes that America has a place for h
WordSeer (Hearst et al 2013)
Search for relationships
WordSeer (Hearst et al 2013)
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

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

All time most popular tags

- africa
- amsterdam
- animals
- architecture
- art
- 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
- day
- dog
- england
- europe
- fall
- family
- festival
- fl
- 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
- 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
- road
- 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.

.net  ajax  apple  architecture  art  article  articles  audio  bit200w07  blog  blogs  books  business  code  comics  community  computer  cooking  cool  css  culture  database  design  development  dy  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  rubynails  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  wishist  wordpress  writing  youtube

(red tags are tags you share with everyone else)
Alternative: “Semantic” Layout


- Tags grouped by “similarity, based on clustering techniques and co-occurrence analysis”

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

4 RESULTS

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

Provided by Martin Wattenberg

![Tag Cloud Alternatives](image-url)
February 12 – North Korea conducts its third underground nuclear test, prompting tightened sanctions by the international community.

April 13 – Kwangmyŏngsŏng-3, a North Korean Earth observation satellite, explodes shortly after launch. The United States and other countries had called the impending launch a violation of United Nations Security Council demands. The launch was planned to mark the centenary of the birth of Kim Il-sung, the founder of the republic.
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

- **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 Language**
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.
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.
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
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.
• **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](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.