Text Visualization

Jason Chuang  University of Washington
Text visualization:
What-Why-How
What is text data?

Documents
Articles, books and novels
E-mails, web pages, blogs

Text Snippets
Tweets, SMS messages
Tags, comments, profiles

And More...
Computer programs, logs
This slide!
Collections of documents
Why visualize text?

**Understanding** – read a document

**Summaries** – get the “gist” of a document

**Clustering** – group together similar contents

**Quantify** – convert to numerical measures

**Correlate** – compare patterns in text to those in other data, e.g., correlate with social network
Example: Health Care Reform

Recent history
  Initiatives by President Clinton
  Overhaul by President Obama

Text data
  News articles
  Speech transcriptions
  Legal documents

What questions might you want to answer?
What visualizations might help?
September 10, 2009

TEXT

Obama’s Health Care Speech to Congress

Following is the prepared text of President Obama’s speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of you
Tag Clouds: Word Count

President Obama’s Health Care Speech to Congress [New York Times]
Barack Obama 2009
Bill Clinton 1993
WordTree: Word Sequences

Visualizations: Word Tree President Obama's Address to Congress on Health Care
I will let up until those Americans who seek jobs can find them -- (applause) -- until those who are working hard can finally benefit from the hard work they are putting in.

Back down on the basic principle that if Americans can't find affordable coverage, we can't make that mistake with health care.

Sign a plan that adds one dime to our deficits -- either now or in the future, period.

Make sure that no government bureaucrat or insurance company bureaucrat gets between you and the health care you need to protect Medicare.

Continue to seek common ground in the weeks ahead.

Be there to listen.

Still believe we can act even when it's hard.

Replace acrimony with civility, and gridlock with progress.

Do great things, and that here and now we will meet history's test.

I still believe that we can act when it's hard.
A Double Gulf of Evaluation

Many (most?) text visualizations do not represent the text directly. They represent the output of a language model (word counts, word sequences, etc.).

- Can you interpret the visualization? How well does it convey the properties of the model?
- Do you trust the model? How does the model enable us to reason about the text?
Topics

Summarizing with Words
Visualizing Themes in a Document Collection
Quantifying Textual Content
Performing Text Analysis
Summarize with Words
Words are (not) nominal?

High dimensional (10,000+)
More than equality tests
Words have meanings and relations

- Correlations: Hong Kong, San Francisco, Bay Area
- Order: April, February, January, June, March, May
- Membership: Tennis, Running, Swimming, Hiking, Piano
- Hierarchy, antonyms & synonyms, entities, ...
Text Processing Pipeline

1. Tokenization
   Segment text into terms.
   Remove stop words?  a, an, the, of, to, be
   Numbers and symbols?  #gocard, @stanfordfbball, Beat Cal!!!!!!!
   Entities?  San Francisco, O’Connor, U.S.A.

2. Stemming
   Group together different forms of a word.
   Porter stemmer?  visualization(s), visualize(s), visually  →  visual
   Lemmatization?  goes, went, gone  →  go

3. Ordered list of terms
Tips: Tokenization and Stemming

Well-formed text to support stemming?

*txt u l8r!*

Word meaning or entities?

#berkeley → #berkelei

Reverse stems for presentation.

*Ha apl made programm cool?*

*Has Apple made programmers cool?*
Bag of Words Model

Ignore ordering relationships within the text

A document ≈ vector of term weights
  • Each dimension corresponds to a term (10,000+)
  • Each value represents the relevance
    For example, simple term counts

Aggregate into a document-term matrix
  • Document vector space model
Document-Term Matrix

Each document is a vector of term weights
Simplest weighting is to just count occurrences

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>157</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brutus</td>
<td>4</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>232</td>
<td>227</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>worser</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
WordCount (Harris 2004)

http://wordcount.org
Tag Clouds

Strengths
Can help with gisting and initial query formation.

Weaknesses
Sub-optimal visual encoding (size vs. position)
Inaccurate size encoding (long words are bigger)
May not facilitate comparison (unstable layout)
Term frequency may not be meaningful
Does not show the structure of the text
Keyword Weighting

Term Frequency
\[ \text{tf}_{td} = \text{count}(t) \text{ in } d \]
Can take log frequency: \( \log(1 + \text{tf}_{td}) \)
Can normalize to show proportion: \( \frac{\text{tf}_{td}}{\sum_t \text{tf}_{td}} \)
Keyword Weighting

Term Frequency
\[ tf_{td} = \text{count}(t) \text{ in } d \]

TF.IDF: Term Freq by Inverse Document Freq
\[ tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t) \]
\[ df_t = \# \text{ docs containing } t; \quad N = \# \text{ of docs} \]
Partisan Words, 106th Congress, Abortion
(Log-Odds-Ratio, Smoothed Log-Odds-Ratio)
Keyword Weighting

Term Frequency
\[ tf_{td} = \text{count}(t) \text{ in } d \]

**TF.IDF: Term Freq by Inverse Document Freq**
\[ tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t) \]
\[ df_t = \# \text{ docs containing } t; \ N = \# \text{ of docs} \]

**G^2: Probability of different word frequency**
\[ E_1 = |d| \times (tf_{td} + tf_{t(C-d)}) / |C| \]
\[ E_2 = |C-d| \times (tf_{td} + tf_{t(C-d)}) / |C| \]
\[ G^2 = 2 \times (tf_{td} \log(tf_{td}/E_1) + tf_{t(C-d)} \log(tf_{t(C-d)}/E_2)) \]
Limitations of Frequency Statistics?

Typically focus on unigrams (single terms)

Often favors frequent (TF) or rare (IDF) terms
   Not clear that these provide best description

A “bag of words” ignores additional information
   Grammar / part-of-speech
   Position within document
   Recognizable entities
How do people describe text?

We asked 69 subjects (graduate students) to read and describe dissertation abstracts.

Students were given 3 documents in sequence; they then described the collection as a whole.

Students were matched to both familiar and unfamiliar topics; topical diversity within a collection was varied systematically.

[Chuang, Manning & Heer, 2012]
Bigrams (phrases of 2 words) are the most common.
Phrase length declines with more docs & more diversity.
Term Commonness

\[ \log(tf_w) / \log(tf_{the}) \]

The normalized term frequency relative to the most frequent n-gram, e.g., the word “the”.

Measured across an entire corpus or across the entire English language (using Google n-grams)
Selected descriptive terms have medium commonness. Judges avoid both rare and common words.
Commonness increases with more docs & more diversity.
A fighter jet rain check

Story and video by Chamila Jayaweera

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out more about how the process works and watch the action unfold, click above to see the video story.
G² Regression Model

fighter
F/A
Hornet
Super
Boeing
-18
rain
St.
jet
Louis
15-inch-per-hour
douse
hangar
water-tight
Check
Baxter
sea-based
aircraft
Rich
sleep
click
Navy
sure
Water
moisture
watch
enormous
stay

Super Hornet
F/A -18
fighter jet
Boeing engineers
special process
rain check
electronics suite
Program experts
simulated rain
ultimate customers
enormous hangar
water-tight integrity
Rich Baxter
15-inch-per-hour rate
video story
aircraft
U.S. Navy fighter pilots
Super Hornet final assembly manager
Yelp: Review Spotlight [Yatani 2011]

- amazing
- around
- baked
- bar
- bass
- best
- chef
- delicious
- eat
- elite
- everything
- favorite
- fish
- food
- fresh
- going
- hamachi
- hawaiian
- hour
- line
- love
- mango
- minutes
- mussels
- name
- nigiri
- order
- people
- prices
- really
- restaurant
- roll
- expensive or cheap?
- sushi
- sake
- salmon
- sea
- seated
- service
- spicy
- stars
- sure
- table
- think
- tuna
- waitress
- worth
- “long wait” or “no wait”?
- what type of sushi roll?
possess sage of the halos wisdom, and know in advance sushi zone only accepts cash and the waits will be long and arduous.

yes, it's a long wait, learn the master of zen if you want to eat here.
Tips: Descriptive Keyphrases

Understand the limitations of your language model.

Bag of words:
   Easy to compute
   Single words
   Loss of word ordering

Select appropriate model and visualization
   Generate longer, more meaningful phrases
   Adjective-noun word pairs for reviews
   Show keyphrases within source text
Visualize Themes in a Document Collection
Topical Analysis

Large document collections
  Too large to manually read the source documents
  Deeper analysis than the most common theme

Statistical topic modeling
  Analysis of word relationships
  Extract latent topics belonging to the documents
Statistical Topic Modeling

Latent topics

- optical
- quantum
- results
- frequency
- provide
- laser
- demonstrate
- dna
- replication
- interaction
- rna

Topic 1
Topic 2
Topic 3
...
Topic K
Computational Linguistics

[Hall et al. 2008]

ACL conferences and journals
14,000 papers over 40 years

Statistics
46 latent topics

History
Topics
Rise and fall of research areas

Speech Recognition
speech, recognition, word, system, language, data, speaker, error, test, spoken, ...

Stats Machine Translation
english, word, alignment, language, source, target, sentence, machine, bilingual, mt, ...

Figure 7: Applications over time

Figure 8: Six applied topics over time

Figure 9: Speech recognition over time

Differences and Similarities Among COLING, ACL, and EMNLP

The computational linguistics community has two distinct conferences, COLING and ACL, with different histories, organizing bodies, and philosophies. Traditionally, COLING was larger, with parallel sessions and presumably a wide variety of topics, while ACL had single sessions and a more narrow scope. In recent years, however, ACL has moved to parallel sessions, and the conferences are of similar size. Has the distinction in breadth of topics also been blurred? What are the differences and similarities in topics and trends between these two conferences?

More recently, the EMNLP conference grew out of the Workshop on Very Large Corpora, sponsored by the Special Interest Group on Linguistic Data and corpus-based approaches to NLP (SIGDAT).
Computational Linguistics

[Hall et al. 2008]

ACL conferences and journals
14,000 papers over 40 years

Statistical topic modeling
46 latent topics

History of ideas
Topical proportions per year
Rise and fall of research areas

Figure 7: Applications over time

Applications

Figure 8: Six applied topics over time
looked at trends over time for the following applications:
Machine Translation, Spelling Correction, Dialogue Systems, Information Retrieval, Call Routing, Speech Recognition, and Biomedical applications.

Figure 7 shows a clear trend toward an increase in applications over time. The figure also shows an interesting bump near 1990. Why was there such a sharp temporary increase in applications at that time? Figure 8 shows details for each application, making it clear that the bump is caused by a temporary spike in the Speech Recognition topic.

In order to understand why we see this temporary spike, Figure 9 shows the unsmoothed values of the Speech Recognition topic prominence over time.

Figure 9 clearly shows a huge spike for the years 1989–1994. These years correspond exactly to the DARPA Speech and Natural Language Workshop, held at different locations from 1989–1994. That workshop contained a significant amount of speech until its last year (1994), and then it was revived in 2001 as the Human Language Technology workshop with a much smaller emphasis on speech processing. It is clear from Figure 9 that there is still some speech research appearing in the Anthology after 1995, certainly more than the period before 1989, but it’s equally clear that speech recognition is not an application that the ACL community has been successful at attracting.

6 Differences and Similarities Among COLING, ACL, and EMNLP

The computational linguistics community has two distinct conferences, COLING and ACL, with different histories, organizing bodies, and philosophies. Traditionally, COLING was larger, with parallel sessions and presumably a wide variety of topics, while ACL had single sessions and a more narrow scope. In recent years, however, ACL has moved to parallel sessions, and the conferences are of similar size. Has the distinction in breadth of topics also been blurred? What are the differences and similarities in topics and trends between these two conferences?

More recently, the EMNLP conference grew out of the Workshop on Very Large Corpora, sponsored by the Special Interest Group on Linguistic Data and corpus-based approaches to NLP (SIGDAT).

Automated data analysis
via machine learning?

Generate 100 latent topics
Retain 36 topics
Remove 64 topics
Manually insert 11 new topics
Re-generate topics (with seeds)
Label topics
NIH Grants & Funding Agencies

Biomedical research
110,000 NIH grant awards
220,000 MEDLINE journal articles

Statistical topic modeling
700 latent topics

Clustering & correlations
Project areas & Funding institutes
Trends in research funding
Changes in research topics

[Image: Cluster map of biomedical research topics and funding agencies]

[Talley et al. 2011]
NIH Grants & Funding Agencies

[110,000 NIH grant awards
220,000 MEDLINE journal articles

Biomedical research

Statistical topic modeling
700 latent topics
Clustering & correlations
Project areas & Funding institutes
Trends in research funding
Changes in research topics

Generate 700 latent topics
Remove 15% “nonsensical” topics
Modify vocabulary (phrases, acronyms)
Extensive parameter search
Expert validation and topic curation

Tagging, plus ~215 categorical designations
NIH reporting requirements, rather than
Manual efforts involved in
tion is discovered using two unsupervised
model-driven data analysis?

[Reference: Talley et al. 2011]
# List of Words

<table>
<thead>
<tr>
<th>Anaphora Resolution</th>
<th>resolution anaphora pronoun discourse antecedent pronouns coreference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automata</td>
<td>string state set finite context rule algorithm strings language symbol</td>
</tr>
<tr>
<td>Biomedical</td>
<td>medical protein gene biomedical wkh abstracts medline patient clinical</td>
</tr>
<tr>
<td>Call Routing</td>
<td>call caller routing calls destination vietnamese routed router destination</td>
</tr>
<tr>
<td>Categorial Grammar</td>
<td>proof formula graph logic calculus axioms axiom theorem proofs lamb</td>
</tr>
<tr>
<td>Centering*</td>
<td>centering cb discourse cf utterance center utterances theory coherence</td>
</tr>
<tr>
<td>Classical MT</td>
<td>japanese method case sentence analysis english dictionary figure japan</td>
</tr>
<tr>
<td>Classification/Tagging</td>
<td>features data corpus set feature table word tag al test</td>
</tr>
<tr>
<td>Comp. Phonology</td>
<td>vowel phonological syllable phoneme stress phonetic phonology pronu</td>
</tr>
<tr>
<td>Comp. Semantics*</td>
<td>semantic logical semantics john sentence interpretation scope logic for</td>
</tr>
</tbody>
</table>

[Hall et al. 2008]

<table>
<thead>
<tr>
<th>Topic Words:</th>
<th>vegf angiogenesis vascular_endothelial_growth_factor angiogenic endo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Words:</td>
<td>angiogenesis, vegf, vascular_endothelial_growth_factor, angiogenic, tu</td>
</tr>
<tr>
<td>Phrases:</td>
<td>vascular_endothelial_growth_factor vegf, vegf angiogenesis, vegf recep</td>
</tr>
</tbody>
</table>

[Talley et al. 2011]
Termite | Topic Model Visualization

Matrix Visualization

Drill-Down to Source Documents

Word Frequency

Representative Documents

A Comparison of the Readability of Graphs Using Node-Link and Matrix Representations

Nivedita R. Kadaba, Pourang P. Irani, and Jason Leboe

Balancing Systematic and Flexible Exploration of Social Networks

Adam Perer, Ben Shneiderman

SpicyNodes: A Radial Layout Authoring for the General Public

Niklas Elmqvist, Philippas Tsagias
All papers published in IEEE InfoVis Conference
(16 years, 372 papers, 50 latent topics)

Word frequency per topic
Overview of all topics
Identify junk topics
force directed graph layout algorithms algorithm layouts maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps
force directed graph layout algorithms maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps
force directed graph layout algorithms maintaining stability
force directed graph layout algorithms maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps
force directed graph layout algorithms maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles
rectangular treemaps

Compare topics
Reveal multi-word phrases
force directed graph layout algorithms maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps

Compare topics
Reveal multi-word phrases
force directed graph layout algorithms layouts maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps

Compare topics
Reveal multi-word phrases
force directed graph layout algorithms maintaining stability

dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps

Compare topics
Reveal multi-word phrases
force

directed

graph

layout

algorithms

algorithm

layouts

maintaining

stability

dimension

ordering

ordered

treemap

favorable

aspect

ratios

rectangles

rectangular
treemaps
force directed graph layout algorithms maintaining stability
dimension ordering ordered treemap favorable aspect ratios rectangles rectangular treemaps

Compare topics
Reveal multi-word phrases
Termite | Topic Model Visualization

- Provide overview of all topics
- Examine words in a topic
- Identify junk topics
- Compare topics
- Reveal multi-word phrases
- Access to source documents
Filtering & Seriation
Filtering: What words to show?

**Frequent words are not necessarily discriminative**
data, visualization, information, visual, techniques, users, visualizations, ...

**Saliency**
Score for word $w$ based on frequency and distinctiveness

$$\text{saliency}(w) = \text{frequency}(w) \times \text{distinctiveness}(w)$$

**Distinctiveness**
Knowing a word $w$, how much does it tell us about a topic?

$$\text{distinctiveness}(w) = \text{KL}( P(T|w) || P(T) )$$

$P(T|w)$ = generating topic $T$ for a given word $w$

$P(T)$ = generating topic $T$ for a randomly-selected word in the corpus
Seriation: How to show the words?
Seriation: How to show the words?

**Clustering of related words**
large, node, networks, social, link, diagrams, online, dataset, communities, ...

**Preservation of reading order**
online communities, social networks, node link diagrams, large datasets, ...

**Word similarity matrix (asymmetric)**
document co-occurrence
sentence co-occurrence
collocation (word transition probability)

**Text seriation**
based on bond energy algorithm
accepts asymmetric similarity matrices
aware of salient terms
early termination
Seriation: How to show the words?

**Clustering of related words**
large, node, networks, social, link, diagrams, online, dataset, communities, ...

**Preservation of reading order**
one online communities, social networks, node link diagrams, large datasets, ...

**Word similarity matrix (asymmetric)**
document co-occurrence
sentence co-occurrence
collocation (word transition probability)

**Text seriation**
based on bond energy algorithm
accepts asymmetric similarity matrices
aware of salient terms
early termination
Ordered by frequency

My text seriation method

parallel coordinates coordinate axes scatterplot matrix scatterplots dimensions multidimensional scaling
Quantify Textual Content
### London Riot 2012

Select a rumour to see how unsubstantiated claims are spread on Twitter before being confirmed or denied.

<table>
<thead>
<tr>
<th>Rumour</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rioters attack London zoo and release animals</td>
<td><img src="image" alt="Tiger" /></td>
</tr>
<tr>
<td>Rioters cook their own food in McDonald's</td>
<td><img src="image" alt="Hamburger" /></td>
</tr>
<tr>
<td>Police 'beat a 16-year-old girl'</td>
<td><img src="image" alt="Fire" /></td>
</tr>
<tr>
<td>London Eye set on fire</td>
<td><img src="image" alt="London Eye" /></td>
</tr>
<tr>
<td>Rioters attack a children's hospital in Birmingham</td>
<td><img src="image" alt="Children's Hospital" /></td>
</tr>
<tr>
<td>Army deployed in Bank</td>
<td><img src="image" alt="Army" /></td>
</tr>
<tr>
<td>Miss Selfridge set on fire</td>
<td><img src="image" alt="Store" /></td>
</tr>
</tbody>
</table>
How riot rumors spread?

Our last challenge was to classify each tweet according to a 'common sense understanding' of its main role as a communicative act. Did it support, oppose, query or comment on a rumour? In addition to an algorithmic analysis by our academic partners, each tweet was independently coded by three sociology PhD students in order to enable us to check for reliability. All the results were then subject to final review for quality assurance purposes. These categories could then be used to colour code each tweet so that readers get an overall picture of what direction the dialogue is taking.
FR88513–0157
AP: Groups Seek $1 Billion a Year for Aging Research
SJMN: WOMEN’S HEALTH LEGISLATION PROPOSED
AP: Older Athletes Run For Science
FR: Committee Meetings
FR: October Advisory Committees; Meetings
FR88120–0046
FR: Chronic Disease Burden and Prevention Models; Program
AP: Survey Says Experts Split on Diversion of Funds for AIDS
FR: Consolidated Delegations of Authority for Policy Development
SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P
<table>
<thead>
<tr>
<th>Psychological Processes</th>
<th>social</th>
<th>Mate, talk, they, child</th>
<th>455</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>family</td>
<td>Daughter, husband, aunt</td>
<td>64</td>
</tr>
<tr>
<td>Friends</td>
<td>friend</td>
<td>Buddy, friend, neighbor</td>
<td>37</td>
</tr>
<tr>
<td>Humans</td>
<td>human</td>
<td>Adult, baby, boy</td>
<td>61</td>
</tr>
<tr>
<td>Affective processes</td>
<td>affect</td>
<td>Happy, cried, abandon</td>
<td>915</td>
</tr>
<tr>
<td>Positive emotion</td>
<td>posemo</td>
<td>Love, nice, sweet</td>
<td>406</td>
</tr>
<tr>
<td>Negative emotion</td>
<td>negemo</td>
<td>Hurt, ugly, nasty</td>
<td>499</td>
</tr>
<tr>
<td>Anxiety</td>
<td>anx</td>
<td>Worried, fearful, nervous</td>
<td>91</td>
</tr>
<tr>
<td>Anger</td>
<td>anger</td>
<td>Hate, kill, annoyed</td>
<td>184</td>
</tr>
<tr>
<td>Sadness</td>
<td>sad</td>
<td>Crying, grief, sad</td>
<td>101</td>
</tr>
<tr>
<td>Cognitive processes</td>
<td>cogmech</td>
<td>cause, know, ought</td>
<td>730</td>
</tr>
<tr>
<td>Insight</td>
<td>insight</td>
<td>think, know, consider</td>
<td>195</td>
</tr>
<tr>
<td>Causation</td>
<td>cause</td>
<td>because, effect, hence</td>
<td>108</td>
</tr>
<tr>
<td>Perception</td>
<td></td>
<td>think, know, consider</td>
<td>36</td>
</tr>
</tbody>
</table>
Visual Thesaurus [ThinkMap]
Named Entity Recognition

Identify and classify named entities in text:

John Smith → PERSON
Soviet Union → COUNTRY
353 Serra St → ADDRESS
(555) 721-4312 → PHONE NUMBER

Entity relations: how do the entities relate?
Simple approach: do they co-occur in a small window of text?
List View

Edit View Bookmarks Lists Options

person

Add all Clear

Show all connections

place

Add all Clear

USA
Cuba
Pakistan
Canada
Columbia
Jamaica
Afghanistan
Havana
Detroit
Mexico
Michigan
Montego Bay
Texas
Chitral
Morocco
Peshawar
Russia
Casablanca
Chicago
Illinois
New Jersey
UK
Dominican Republic
Florida
France
London
Moscow
Ontario
Paris
Windsor
Santo Domingo
Virginia

Buganov
Carlos
Carlos Araneda
Carlos Morales
Castro
Cesar Arze
Charles Wilson
Dan West

Daniel Harris
David Loiseau

Dean Simpson
Dr. Baker
Dustin Marshall
Edgar Spencer

Edward Thompson
Escalante
F. Baker
Felix Baker
Ford
Forrest Wells
Fr. Augustin Dominique
Fred Fisher
George Garcia
Grigory Sizov
Hamid Qataada
Hector Lopez

Herman Fox
Howard Clark
Igor Kolokov

Imad Dahdah
J. T.
Jamal Sveed
Doc. Similarity & Clustering

In vector model, compute distance among docs
• For TF.IDF, typically cosine distance
• Similarity measure can be used to cluster

Topic modeling approaches
• Assume documents are a mixture of topics
• Topics are (roughly) a set of co-occurring terms
• Latent Semantic Analysis (LSA): reduce term matrix
• Latent Dirichlet Allocation (LDA): statistical model
ThemeRiver [Havre et al 99]
TIARA [Wei et al. 09]
Whose are the fathers, and of whom as concerning the flesh Christ came, who is over all, God blessed for ever. Amen.
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 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

<table>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>29</td>
<td>7</td>
<td>13</td>
<td>21</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Compared with other words

<table>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax</td>
<td>26</td>
<td>7</td>
<td>13</td>
<td>21</td>
<td>11</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

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)

New York Times
Concordance

What is the common local context of a term?
Rabbit with pink eyes ran close by her. I think it so VERY much out of the way to hear the Rabbit say to me, "But when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT!-before seen a rabbit with either a waistcoat-pocket, or a watch to put in a large rabbit-hole under the hedge."

The rabbit-hole went straight on like a tunnel for some way, and then another long passage, and the White Rabbit was still at its coming, but the Rabbit was no longer to be seen. She found it was the White Rabbit returning, splendidly dressed, with a large grin, that she was ready to ask help of any one; so, when the Rabbit asked, "Who are you?" it was the White Rabbit returning, splendidly dressed, with a large grin, that she was ready to ask help of any one; so, when the Rabbit asked, "Who are you?"

"I'm Nobody's Nobody," said Alice. "I don't think you can mean that," said the Rabbit.

"I can't think what you mean," said Alice. "I think it's a good deal worse than meaning nothing," said the Rabbit. "Are you sure that you aren't meaning something?" Alice asked, a little shyly. "Certainly not," said the Rabbit. "Why, you might as well say that a horse is a tree."

"I sometimes think that," said Alice, "especially when I see some of our older people talking."

"What is the use of having eyes if you don't put them to use?" said the Rabbit. "It's a capital thing to find out just before you start to do something, that you can't do it."

"I think you'll find," said the Rabbit, "that if you can't do something well, it's better not to do it at all."

"I've often seen a rabbit run across the road when I was going to school," Alice said, very much to her own surprise, for she had never in all her life before seen a rabbit running across the road. "I suppose it's just as easy to run across the road as go down a rabbit-hole!", she thought. "I wish I had a rabbit-hole to run into!", she said aloud.

"Yes, I've often seen a rabbit run across the road when I was going to school," Alice said, very much to her own surprise, for she had never in all her life before seen a rabbit running across the road. "I suppose it's just as easy to run across the road as go down a rabbit-hole!"
if love be rough with you, be rough with love.
if love be blind, love cannot hit the mark.
if love be blind, it best agrees with night.
Filter infrequent runs
Recurrent themes in speech

Today, I have a dream that one day

- This nation will rise up and live out the true meaning of its creed:
  "We hold these truths to be self-evident,
  on the red hills of Georgia the sons of former slaves and the sons of former slave owners will be able to sit down together at
  even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression, will be transformed into
  down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of interposition and nullification - one day
  every valley shall be exalted, and every hill and mountain shall be made low; the rough places will be made plain, and the
  my four little children will one day live in a nation where they will not be judged by the color of their skin but by the
  down in Alabama, with its vicious racists, with its governor having his lips dripping with the words of interposition and nullification - one day
  every valley shall be exalted, and every hill and mountain shall be made low; the rough places will be made plain, and the
Endings and Job Progressions

- Word Transition Probability
Glimpses of structure

Concordances show local, repeated structure
But what about other types of patterns?

For example

Lexical:  <A> at <B>
Syntactic: <Noun> <Verb> <Object>
Phrase Nets [van Ham et al 2009]

Could be output of regexp or parser.

Visualize extracted patterns in a node-link view

Occurrences ➔ Node size
Pattern position ➔ Edge direction
Portrait of the Artist as a Young Man
X and Y
Node Grouping
The Bible

X begat Y
Pride & Prejudice
X at Y
Lexical Parser, < 1sec running time
Pride & Prejudice
X at Y
Syntactic Parser, > 24 hours running time
18th & 19th Century Novels
X's Y
New Testament
X of Y
Text Visualization Summary

High Dimensionality
Where possible use text to represent text...
... which terms are the most descriptive?

Context & Semantics
Provide relevant context to aid understanding.
Show (or provide access to) the source text.

Modeling Abstraction
Determine your analysis task.
Understand abstraction of your language models.
Match analysis task with appropriate tools and models.
Perform Text Analysis
Information Retrieval

Search for documents
  • Match query string with documents

Contextualized search
Track Changes Icons appear once you select "Track Changes" from the "Tools Menu".

This is a test document to demonstrate the use of tracking changes. The characters in black font represent the original document while the characters in red font represent the changes which are being tracked.
# Visualizing Revision History

How to depict contributions over time?

Example: Wikipedia history log

## Chocolate

Revision history

Legend: (cur) = difference with current version, (last) = difference with preceding version, M = minor edit

- (cur) (last) . 12:01, 20 Aug 2003 . Dysprosia *(neaten to do, rearrange see also)*
- (cur) (last) . 11:59, 20 Aug 2003 . Patrick
- (cur) (last) . 11:52, 20 Aug 2003 . 81.203.98.109
- (cur) (last) . M 18:36, 6 Aug 2003 . Manika *(corrected spelling)*
- (cur) (last) . 18:32, 6 Aug 2003 . Daniel Quinlan *(removing obscure heraldry information, belongs on [[heraldry]] if anywhere)*
- (cur) (last) . 15:21, 6 Aug 2003 . Rmhermen
- (cur) (last) . 15:08, 6 Aug 2003 . Cyp *(Chocolate often has odd shapes.)*
- (cur) (last) . 19:14, 3 Aug 2003 . Daniel C. Boyer *("chocolate" as shade of gules in heraldry)*
- (cur) (last) . M 02:00, 30 Jul 2003 . Evercat *(fnt)*
ON THE ORIGIN OF SPECIES The Preservation of Favoured Traces

http://benfry.com/traces/
public class WelcomePageDispatchAction {

    // associated forward definitions
    public final static String COLLECTION;
    public final static String LOGOUTACTION;
    public final static String TEST1_TO_TEST2;

    // inherited forward definitions

    // dispatch action methods declaration
    public ActionForward enter(ActionMapping mapping, ActionForm form, HttpServletRequest request, HttpServletResponse response) {
        if (form != null) {
            WelcomePageForm currentForm = (WelcomePageForm) form;
            currentForm.onExit();
            int i_ENTERXXX=0;
            CollectInformationForm collectInformationForm = (CollectInformationForm) currentForm;
            CollectInformationForm collectInformationForm = (CollectInformationForm) currentForm;
            return null;
        }
    }

    // Start of user code :

Files Changed:

1. sshconsole.js: 1 change [1]

/home/loddw/src/sshconsole-read-only/content/sshconsole.js

```
51    _term = new VT100(80, 24, "term");
52    // _term.debug = 1;
53    _term.curs_set(true, true, _term_box_element);
54    _term.noecho();

// Replace the go getch function with our own, this is called
// for every keypress that is passed through the terminal to the
// remote server. The character is already converted into the
// required VT100 character sequence(s).
VT100.go_getch = function() {
    var vt = VT100.the_vt;

    if (vt === undefined) {
        return;
    }
    var ch = vt.key_buf.shift();

    if (ch === undefined) {
        return;
    }
    if (vt.echo && ch.length == 1) {
        vt.addch(ch);
    }
    if (_ssh_channel) {
        _ssh_channel.sendStdin(ch);
    }
}
```

```
51    _term = new VT100(80, 24, "term");
52    // _term.debug = 1;
53    _term.curs_set(true, true, _term_box_element);
54    _term.noecho();

// Replace the go getch function with our own, this is called
// for every keypress that is passed through the terminal to the
// remote server. The character is already converted into the
// required VT100 character sequence(s).
VT100.go_getch = function() {
    var vt = VT100.the_vt;

    if (vt === undefined) {
        return;
    }
    var ch = vt.key_buf.shift();

    if (ch === undefined) {
        return;
    }
    if (vt.echo && ch.length == 1) {
        vt.addch(ch);
        vt.refresh();
    }
    if (_ssh_channel) {
        _ssh_channel.sendStdin(ch);
    }
}
```
Visualizing Conversation

Many dimensions to consider:

- Who (senders, receivers)
- What (the content of communication)
- When (temporal patterns)

Interesting cross-products:

- What x When ➔ Topic “Zeitgeist”
- Who x Who ➔ Social network
- Who x Who x What x When ➔ Information flow
Naming Names
Names used by major presidential candidates in the series of Democratic and Republican debates leading up to the Iowa caucuses.

- Each thin line represents one candidate speaking the last name of another candidate.
- Every line ends at an arrow, which points to the name that was spoken.
- The length of each circle segment represents the total number of words spoken by the candidate during the debates. Each tick mark represents 1,000 words.
- Each slice represents one debate, arranged clockwise from the first to the final debate.

Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times
Usenet Visualization (Viégas & Smith)

Show correspondence patterns in text forums
Initiate vs. reply; size and duration of discussion
Mountain (Viégas)

Conversation by person over time (who x when).
Themail (Viégas et al)

One person over time, TF.IDF weighted terms
Enron E-Mail Corpus

[Heer]
steven_kean@enron.com

2000-09-01 04:25:00.0 Linda Jenkins on "Jerry's Show" 
2000-09-02 10:14:00.0 Re: The Governors' Natural Gas Summit 
2000-09-08 10:03:00.0 "California" 
2000-09-10 14:07:00.0 CPUC Hearing in SD on 9/8 
2000-09-10 16:20:00.0 Re: Fletcher School/Enron 
2000-09-13 06:47:00.0 Re: Contact

ID: 174285
Subject: [steven_kean@enron.com]
From: steven_kean@enron.com
To: kmagrudr@enron.com
Cc: Richard Shapiro <richard.shapiro@enron.com>

Got your message. I'm testifying at the Congressional hearing and Dave Vick is covering FERC. I think Jeff's comments were taken out of context. He said policymakers need to take care of small customers whose bills are tripling. Frankly, we'd get slaughtered if we said anything else. But he also said there is a right way and a wrong way to do it. Enron and others had provided a market-based answer by offering a fixed price deal to PG&E (which would have enabled them to cap rates to those who had not switched). California elected instead to cap rates and deficit spend (ie create a deferral account). I don't think we can stand for anything that doesn't protect the small customers, but we can continue to emphasize the market-based solutions. One of the messages in my testimony will be: customers should be encouraged to choose. Those who did are doing fine.
Enron 'Mastermind' Pleads Guilty

SAN FRANCISCO, Oct. 17, 2002

(AP) A former top energy trader, considered the mastermind of Enron Corp.'s scheme to drive up California's energy prices, pleaded guilty Thursday to a federal conspiracy charge.

Deputy Attorney General Larry Thompson, center, head of the Justice Department's Corporate Fraud Task Force, comments Thursday on the guilty plea by Timothy N. Belden, Enron's chief energy trader. (Photo: CBS/AP)

Timothy Belden, the former head of trading in Enron's Portland, Ore., office, admitted to one count of conspiracy to commit wire fraud and promised to cooperate with state and federal prosecutors as well as any non-criminal effort to investigate the energy industry.

"I did it because I was trying to maximize profit for Enron," Belden told U.S. District Judge Martin Jenkins.
Tips: Document Contents

Understand your task, and handle high dimensionality accordingly...

*Visually*: Word position, browsing, brushing + linking

*Semantically*: Word sequence, hierarchy, clustering

*Both*: Spatial layout reflect semantic relationships

Role of Interaction:

Sufficient language model to enable visual analysis cycles

Allow modifications to the model: custom patterns for expressing contextual or domain knowledge