Text as Data

Documents
Articles, books and novels
E-mails, web pages, blogs
Tags, comments
Computer programs, logs

Collections of Documents
Messages (e-mail, blogs, tags, comments)
Social networks (personal profiles)
Academic collaborations (publications)
Why Visualize Text?
Why Visualize Text?

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

**Grouping** – cluster for overview or classification

**Comparison** – compare document collections, or inspect evolution of collection over time

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

Background
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 in one sentence about the direction we must take together.
Tag Clouds: Word Count

President Obama’s Health Care Speech to Congress [NY Times]

Barack Obama 2009

Bill Clinton 1993

Visualizations: Word Tree President Obama's Address to Congress on Health Care
I will not let up until those Americans who seek jobs can find them -- (applause) -- until those back down on the basic principle that if Americans can't find affordable coverage, we will sign a plan that adds one dime to our deficits -- either now or in the future, period. I will not make that same mistake with health care. I will not waste time with those who have made the calculation that it's better politics to kill this. And I will not accept the status quo as a solution.

I will make sure that no government bureaucrat or insurance company bureaucrat gets between you and the protection Medicare.

I will continue to seek common ground in the weeks ahead.

I still believe we can act even when it's hard.
Gulfs of Evaluation

Many 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?
Text Visualization Challenges

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.
Topics

Text as Data
Visualizing Document Content
Visualizing Conversation
Document Collections
Text as Data
Words as nominal data?

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

- **Correlations:**  *Hong Kong, Puget Sound, 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?  #huskies, @UW, OMG!!!!!!!
Entities?  Washington State, O’Connor, U.S.A.
Text Processing Pipeline

1. Tokenization
   Segment text into terms.
   - Remove stop words: *a, an, the, of, to, be*
   - Numbers and symbols: *#huskies, @UW, OMG!!!!!!!*
   - Entities: *Washington State, 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*
Text Processing Pipeline

1. Tokenization
   Segment text into terms.
   Remove stop words?  a, an, the, of, to, be
   Numbers and symbols?  #huskies, @UW, OMG!!!!!!!
   Entities?  Washington State, 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
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>
WordCounts (Harris ’04)

http://wordcount.org
Google Books Ngram Viewer

Graph these comma-separated phrases: automobile, locomotive, airplane

between 1800 and 2000 from the corpus English

with smoothing of 3

Search lots of books

https://books.google.com/ngrams/
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
Given a text, what are the best descriptive words?
Keyword Weighting

**Term Frequency**

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

Can take log frequency: \( \log(1 + tf_{td}) \)

Can normalize to show proportion: \( \frac{tf_{td}}{\sum_t tf_{td}} \)
Keyword Weighting

**Term Frequency**

\[ \text{tf}_{td} = \text{count}(t) \text{ in } d \]

**TF.IDF: Term Freq by Inverse Document Freq**

\[ \text{tf.idf}_{td} = \log(1 + \text{tf}_{td}) \times \log(N/\text{df}_t) \]

\[ \text{df}_t = \# \text{ docs containing } t; \ N = \# \text{ of docs} \]
Keyword Weighting

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

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

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

Require comparison across full corpus!
Limitations of Freq. 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 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

$$\log(\text{tf}_w) \div \log(\text{tf}_{\text{the}})$$

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

Measured across a 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.
Scoring Terms with Freq, Grammar & Position

- Best-Performing Model
- Corpus-Independent Model
- $\log \text{tf} + \text{All Commonness}$
- $G^2$
- $\log \text{tf}$

Precision

Recall
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
seep
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
- night
- nigiri
- order
- people
- prices
- really
- restaurant
- roll
- expensive or cheap?
- sushi
- wait
- “long wait” or “no wait”? what type of sushi roll?
Yelp Review Spotlight (Yatani 2011)

Mentioned 63 times

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 Phrases

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
Document Content
Information Retrieval

Search for documents
Match query string with documents
Visualization to contextualize results
Rom 9:5 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

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.

-- 2001 (Paragraph 14 of 73)

New York Times
Concordance

What is the common local context of a term?
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.
thine heart, and with all thy soul, and with all thy might, that thou mayest live. 

keep his charge, and his statutes, and his judgments, and his commandments, alway. 
to walk ever in his ways; then shalt thou add three cities more for thee, beside these three: 
that thou mayest obey his voice, and that thou mayest cleave unto him: for he is thy life, and the 
to walk in his ways, and to keep his commandments and his statutes and his judgments, that thou mayest live. 
serve him with all your heart and with all your soul, 11:14 that I will give you the rain of your land, 
to walk in all his ways, and to keep his commandments, and to cleave unto him, and to serve him. 
with all your heart and with all your soul. 

all ye his saints: for the Lord preserveth the faithful, and plentifully rewardeth the proud doer. 
hate evil: he preserveth the souls of his saints; he delivereth them out of the hand of the wicked. 
because he hath heard my voice and my supplications. 

name of the Lord, to be his servants, every one that keepeth the sabbath from polluting it, and taketh hold of my covenant 
good, and establish judgment in the gate: it may be that the Lord God of hosts will be gracious unto the remnant of Joseph 
evil; who pluck off their skin from off them, and their flesh from off their bones; 3:3 who also eat the 
truth and peace. 

other; or else he will hold to the one, and despise the other. ye cannot serve God and mammon. 

rooms at feasts, and the chief seats in the synagogues, 23:7 and greetings in the markets, and to be called of 
uppermost seats in the synagogues, and greetings in the markets. 

father hath bestowed upon us, that we should be called the sons of God: therefore the world knoweth us not, because it knew him 
brotherhood. 
world, the love of the father is not in him. 
brethren. 
children of God, when we love God, and keep his commandments.
Filter Infrequent Runs
Recurrent Themes in Speeches

I have a dream that one day...

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...

this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident,"...
Visualizations: Word tree / Alberto Gonzales

Creator: Martin Wattenberg
Tags:

Search: i don't

Data source: CQ Transcript Wire via the Washington Post

Comments (4)
Glimpses of Structure...

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

Lexical:  <A> at <B>

Syntactic:  <Noun> <Verb> <Object>
Phrase Nets [van Ham et al.]

Look for specific linking patterns in the text:


Could be output of regexp or parser.

Visualize 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

(a) 

(b) 

(c)
X begat Y
18th & 19th Century Novels
X’s Y
Old Testament

X of Y
New Testament
X of Y
Understand Your Analysis Task

Visually: Word position, browsing, brush & link
Semantically: Word sequence, hierarchy, clustering
Both: Spatial layout reflects semantic relationships

The Role of Interaction

Language model supports visual analysis cycles
Allow modifications to the model: custom patterns for expressing contextual or domain knowledge
Conversations
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 [Viegaas & Smith]

Show correspondence patterns in text forums
Initiate vs. reply; size and duration of discussion
Email Mountain [Viegas]

Conversation by person over time (who x when).
Themail [Viegas]

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

[Heer]
steven.kean@enron.com

Subject: Re: California hearings and Daseyvich

Date: 2000-09-02 10:14:00.0

To: <kmagru@enron.com>
Cc: Richard Shapiro <richard.shapiro@enron.com>

Got your message. I'm testifying at the Congressional hearing and Daseyvich 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 (i.e., 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)
Document Collections
Named Entity Recognition

Label 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?
<table>
<thead>
<tr>
<th>Person</th>
<th>Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel Harris</td>
<td>Jamaica, Montego Bay</td>
</tr>
<tr>
<td>Dr. Baker</td>
<td>USA, Texas, Chitral</td>
</tr>
<tr>
<td>Edward Thompson</td>
<td>Mexico, Peshawar</td>
</tr>
<tr>
<td>Herman Fox</td>
<td>Florida, London</td>
</tr>
<tr>
<td>Imad Dahdah</td>
<td>Santo Domingo, Virginia</td>
</tr>
</tbody>
</table>
Parallel Tag Clouds [Collins et al.]
Theme River [Havre et al.]
Similarity & Clustering

Compute vector distance among docs
For TF-IDF, typically cosine distance
Similarity measure can be used to cluster

Topic modeling
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
Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning


Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.
PCA projections for four similarity measures. Which to prefer?
Topic Distance Between Stanford Depts

Area of circles denote number of theses in a given year.
Depts with no thesis produced are faded out.

Purple = Medicine
Green = Sciences
Blue = Engineering
Orange/Pink = Humanities

Undistorted distances from focal point.
Oh, the humanities!
(Insufficient number of topics)
Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning


Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks---sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

Drill down to specific theses and most-related departments.
PCA projections for four similarity measures. Which to prefer?

Answer: None of the above. We need a new affinity measure!
Asymmetric affinities…

“Word Borrowing” via Labeled LDA
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
Understand abstraction of your language models.
Match analysis task with appropriate tools and models.
Currently: from bag-of-words to vector space embeddings