CS448G :: 25 May 2011

Text Analysis

Jeffrey Heer Stanford University

What should we hope to gain from visual analysis of text?

Goals of Visual Text Analysis

Understanding – get the "gist" of documents

Grouping – cluster for overview or classification

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

Correlate – compare patterns in text to those in other data, e.g., correlate with social network

Some Text Analysis Topics

Vector Space Model
Descriptive Phrases
Named Entity Recognition
Clustering & Topic Modeling

Vector Space Model

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

Tokenization: segment text into *terms*

- Special cases? e.g., "San Francisco", "L'ensemble", "U.S.A."
- Remove stop words? e.g., "a", "an", "the", "to", "be"?

Stemming: one means of normalizing terms

- Reduce terms to their "root"; Porter's algorithm for English
- e.g., *automate(s), automatic, automation* all map to *automat*
- For visualization, want to reverse stemming for labels
  - Simple solution: map from stem to the most frequent word

Result: ordered stream of terms

The 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 x term matrix

- Document vector space model

Document x Term matrix

Each document is a vector of term weights

Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Keyword Weighting

Term Frequency

tf_{td} = count(t) in d

TF.IDF: Term Freq by Inverse Document Freq

$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

df_t = # docs containing t; N = # of docs

G^2: Probability of different word frequency

$E_1 = |d| \times (tf_{td} + tf_{t(C-d)}) / |C|$

$E_2 = |C\text{-}d| \times (tf_{td} + tf_{t(C-d)}) / |C|$

$G^2 = 2 \times (tf_{td} \log(tf_{td}/E_1) + tf_{t(C-d)} \log(tf_{t(C-d)}/E_2))$

Useful Libraries

Apache Lucene

http://lucene.apache.org/

Tokenization, stemming, indexing & search

Stanford Core NLP

http://nlp.stanford.edu/software/

Tokenization, stemming, POS tagging, entity recognition, parsing & coreference resolution

Descriptive Phrases

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

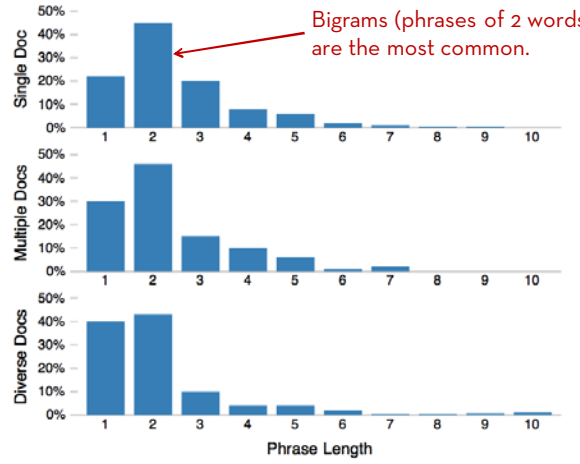
How do people describe text?

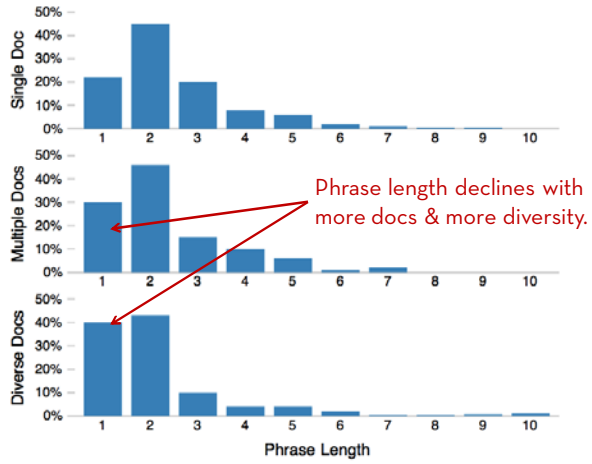
We asked 69 subjects (all Ph.D. students) to read and describe dissertation abstracts.

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, 2010]



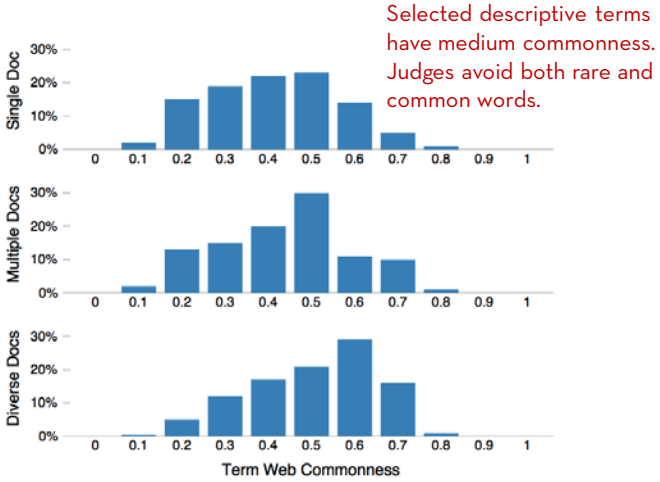


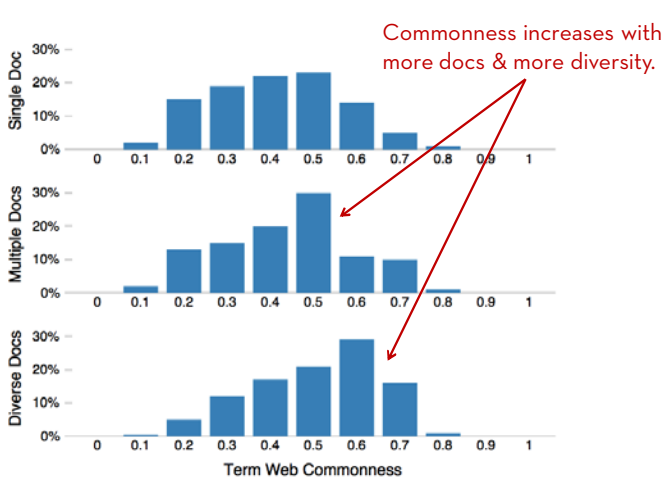
Term Commonness

$$\log(tf_w) \ / \ \log(tf_{the})$$

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)





Grammar: Technical Term Patterns

Technical Term	*T = (A|N)+ (N|C) | N*
Compound Tech. Term	*X = (A|N)* N of T*

Regular expressions over part-of-speech tags.

A = adjective, *N* = noun, C = cardinal number.

Prior work suggests these patterns can be used to identify important terms in text.

Over 4/5 of selected terms match pattern!

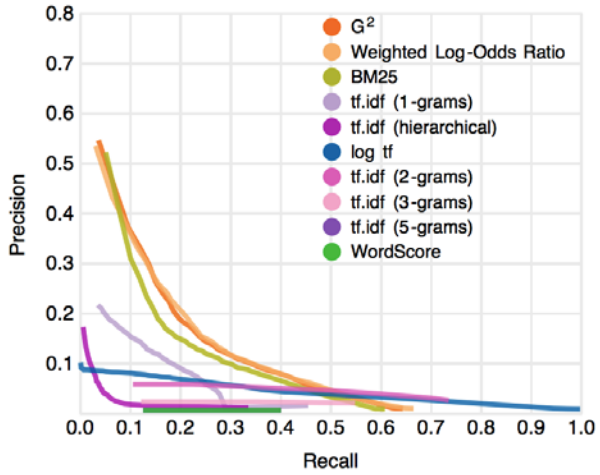
Method – Part 2

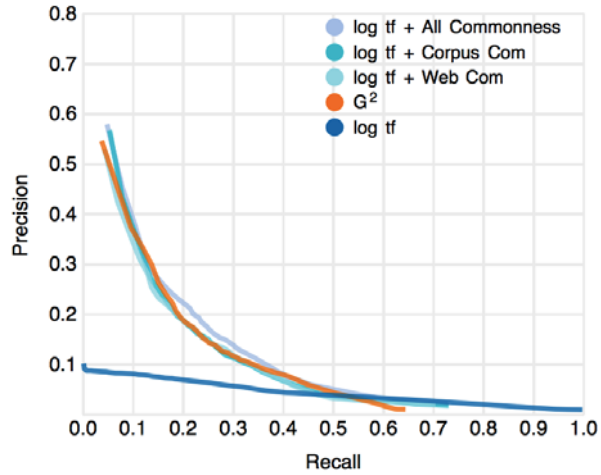
Build a statistical model of keyphrase quality

Train a logistic regression model
Positive examples: selected phrases
Negative examples: randomly sampled phrases

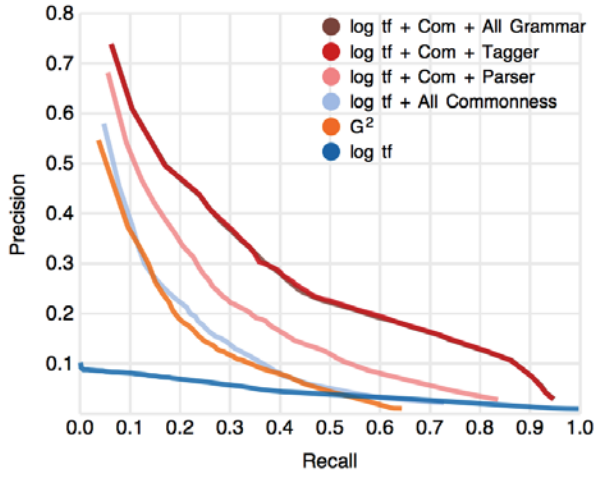
Assess contributions of four classes of features:
Freq stats, commonness, grammar & position

Evaluate the phrases selected by our model using precision & recall measures

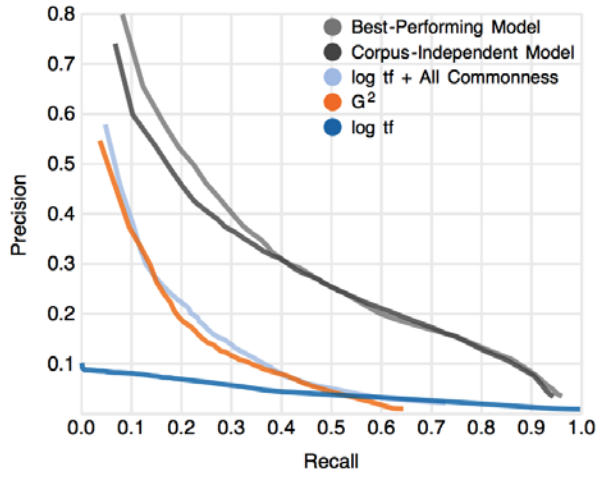




Adding Grammatical Features



Scoring Terms with Freq, Grammar & Position



Model Feature	Coefficients
constant	-2.3550***
log(tf)	0.9390***
WC ∈ (0%, 20%]	0.1770
WC ∈ (20%, 40%]	0.2304*
WC ∈ (40%, 60%]	0.0158
WC ∈ (60%, 80%]	-0.6205***
WC ∈ (80%, 100%]	-1.9081***
relative first occurrence	0.4800**
first sentence	0.9386***
full tech. term	-0.5015
partial tech. term	1.4461**
full compound tech. term	1.1373
partial compound tech. term	1.1806*

Fitted Parameters for a Corpus-Independent Model
WC: web commonness bins, *: p < 0.05, **: p < 0.01, ***: p < 0.001

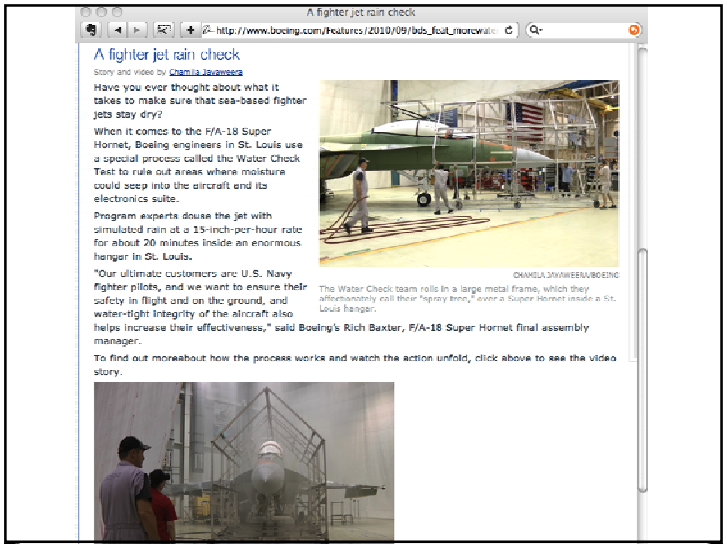
Compare to User-Selected Phrases

Automatic Keyphrase Extraction

Phase 1 Score candidate terms using our keyphrase quality regression model

Phase 2 Eliminate redundancy by grouping similar terms based on word overlap plus entity and acronym resolution.

- *"analysis", "data analysis", ...*
- *"Barack Obama", "Obama", ...*



G^2	*Regression Model*
fighter	Super Hornet
F/A	F/A -18
Hornet	fighter jet
Super	Boeing engineers
Boeing	special process
-18	rain check
rain	electronics suite
St.	Program experts
jet	simulated rain
Louis	ultimate customers
15-inch-per-hour	enormous hangar
douse	water-tight integrity
hangar	Rich Baxter
water-tight	15-inch-per-hour rate
Check	video story
Baxter	aircraft
sea-based	U.S. Navy fighter pilots
aircraft	Super Hornet final assembly manager
Rich	
seep	
click	
Navy	
suni	
Water	
moisture	
watch	
enormous	
stay	

Named Entity Recognition

Named Entity Recognition

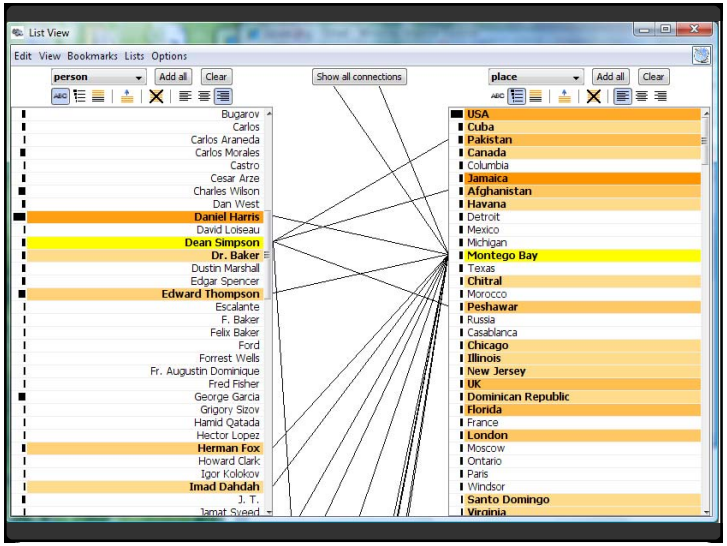
Input:

Jim bought 300 shares of ACME Corp. in 2006.

Output:

<PERSON>Jim</PERSON> bought <QUANTITY>300</QUANTITY> shares of <ORG>ACME Corp.</ORG> in <YEAR>2006</YEAR>.

Current state-of-the-art uses statistical methods: conditional random fields (CRF).



Entity Recognition Uses/Issues

Extract structured data from unstructured text.

Analyze documents w.r.t. constituent entities.

>> subject of lecture on Jigsaw (see also Palantir)

Challenges

- Entity resolution (which entities are the same?)
- Entity linkage (how are entities connected? Can one label the relationship types?)
- Most analysis tools require human-in-the-loop

Clustering / Topic Modeling

Cluster & Compare Documents

STRATEGY 1: Cluster term vectors directly

- K-means clustering of document vectors
- Hierarchical agglomerative clustering

STRATEGY 2: Map documents to lower-dim space

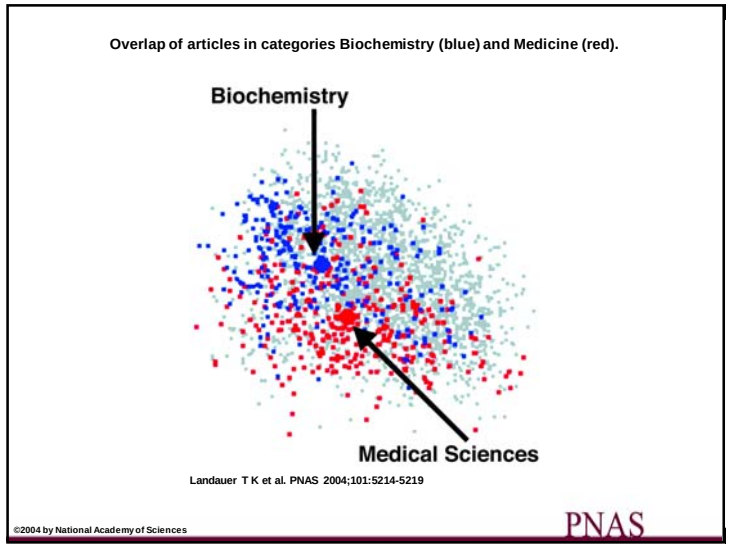
- Latent Semantic Analysis (LSA)
- Statistical Methods
  (e.g., Latent Dirichlet Allocation or LDA)
- Use topics as clusters or compare in new space

Latent Semantic Analysis (LSA)

IDEA: Construct a low-rank approximation of the document-term matrix. Map the individual terms (10,000+) to a smaller set of dimensions (typically 100-300) representing "concepts"

Roughly, the idea is to find the eigenvectors and use the top-K vectors to represent the data.

BUT, doc-term matrices aren't square. Thus a different way to find principal components is used: *singular value decomposition (SVD)*.



Statistical Topic Modeling

Discover latent topics in a collection by finding words that co-occur in the same documents.

Popular model: *Latent Dirichlet Allocation (LDA)*
Assumes documents are mixtures of topics, and the words are drawn from these topics.

Given K (# topics) and a set of documents D with words from a vocabulary V, determine:

β_k ~ Multinomial(V) – topics as dist. over words

θ_d ~ Multinomial(K) – per-doc topic mixtures

Example LDA Output

Topics	Terms
Physics	optical, quantum, frequency, laser, high, electron, single
Biology	dna, replication, rna, repair, complex, interaction, base
Generic	results, show, finally, present, provide, demonstrate

Topic Identification

LDA is an example of a *latent* topic model: topics are inferred from the data. The learned topics have no names: each is exactly and only defined by its distribution over words.

What if we already have label metadata?

Labeled LDA: assume each label *is* a topic, and infer the probability of the words associated with each label.

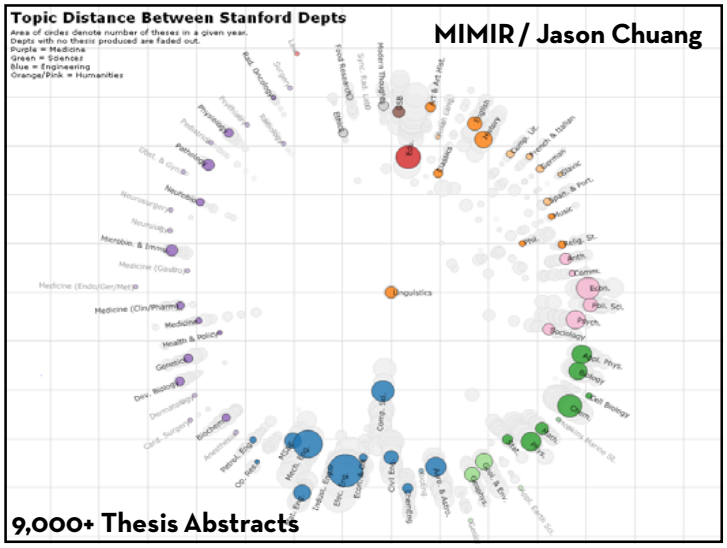
Document Similarity

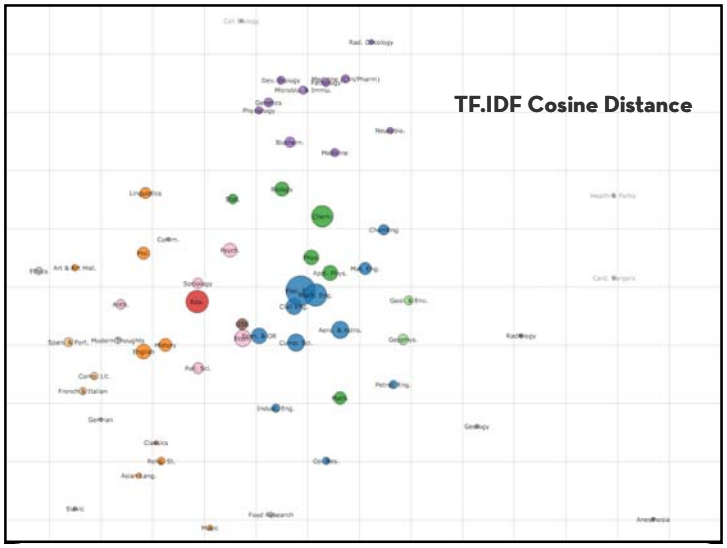
Compute distance between document vectors

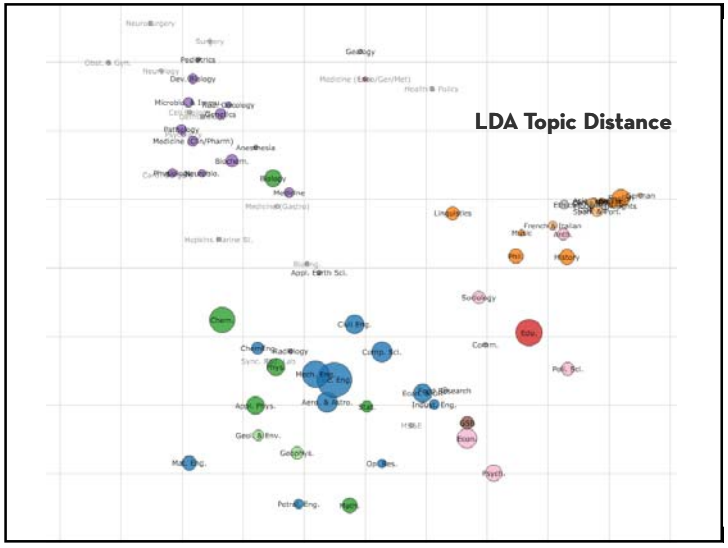
- Common: cosine of angle b/w term vectors (just the dot product if vectors are normalized)
- Similarity affected by term weighting (e.g., tfidf)

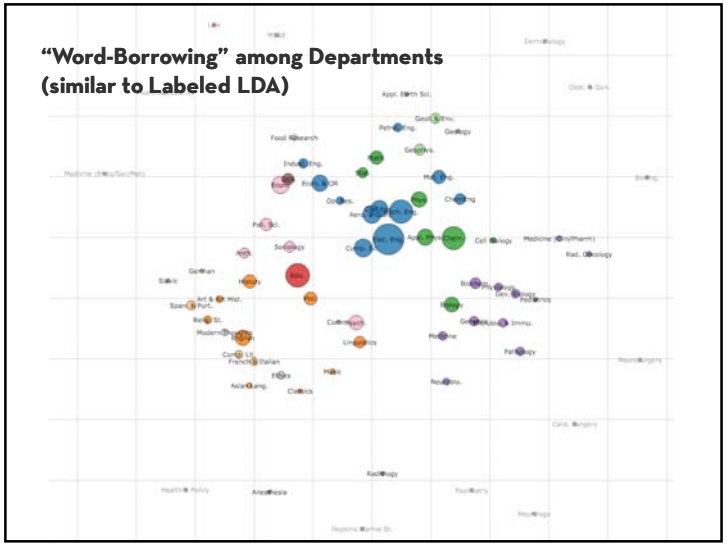
With LSA or LDA models, can compute distance in lower dimensional space, e.g., $\cos(\theta_{d1}, \theta_{d2})$

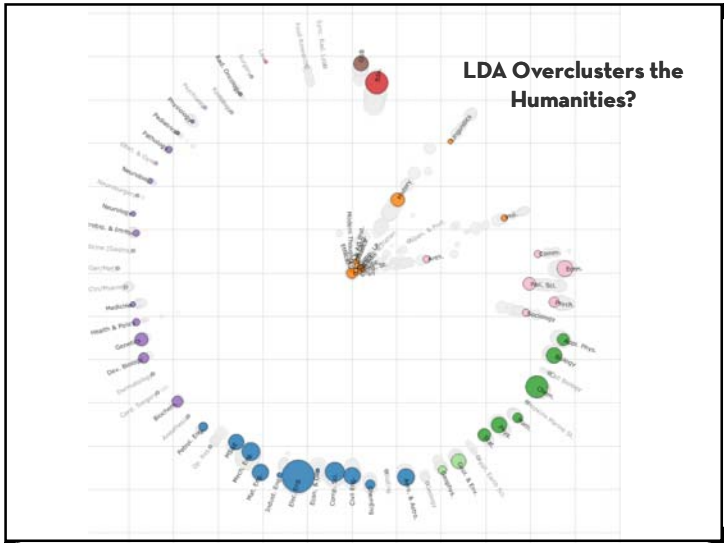
Example: Thesis Explorer


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
MIMIR / Jason Chuang
9,000+ Thesis Abstracts


TF.IDF Cosine Distance


LDA Topic Distance


"Word-Borrowing" among Departments
(similar to Labeled LDA)



Implications for Vis Design

Standard InfoVis Reference Model:

Input -> Data -> Visualization <-> Analyst

Model-Driven Visualization?

Input -> Data -> Model -> Visualization <-> Analyst

What considerations guide the visualization process for increasingly abstracted data?

Interpretation & Trust

Interpretation: the facility with which an analyst makes inferences about the underlying data, which is affected by all stages in the process (choice of model, choice of visual encodings...)

Trust: the actual and perceived validity of an analysts's inferences. Are the interpretations valid and actionable?

Strategy 1: Alignment

Understand the domain and organize the display around the *primary entities of interest*. Define the model accordingly. Anticipate sources of error to *reduce false inferences*.

In the Thesis Explorer this led to:

- Focus on similarity between departments
- Choice of word-borrowing model
- Abandonment of PCA projection overview

Strategy 2: Progressive Disclosure

Enable analysts to *shift among levels of abstraction on-demand*, by drilling down from overview, to intermediate abstractions, and eventually to the underlying data itself.

In the Thesis Explorer this led to:

- Department level view
- Thesis level drill-down view, access to text
- Needed future work to assess word overlap

Strategy 3: Social Verification (?)

Observe and record domain experts' *met and unmet expectations* as they explore the data. We used this information to drive modeling decisions for the thesis explorer.

In the future, visualizations might enable a larger, crowd-sourced verification of the model's ability to quantify and discover facts known to distinct subsets of a community.