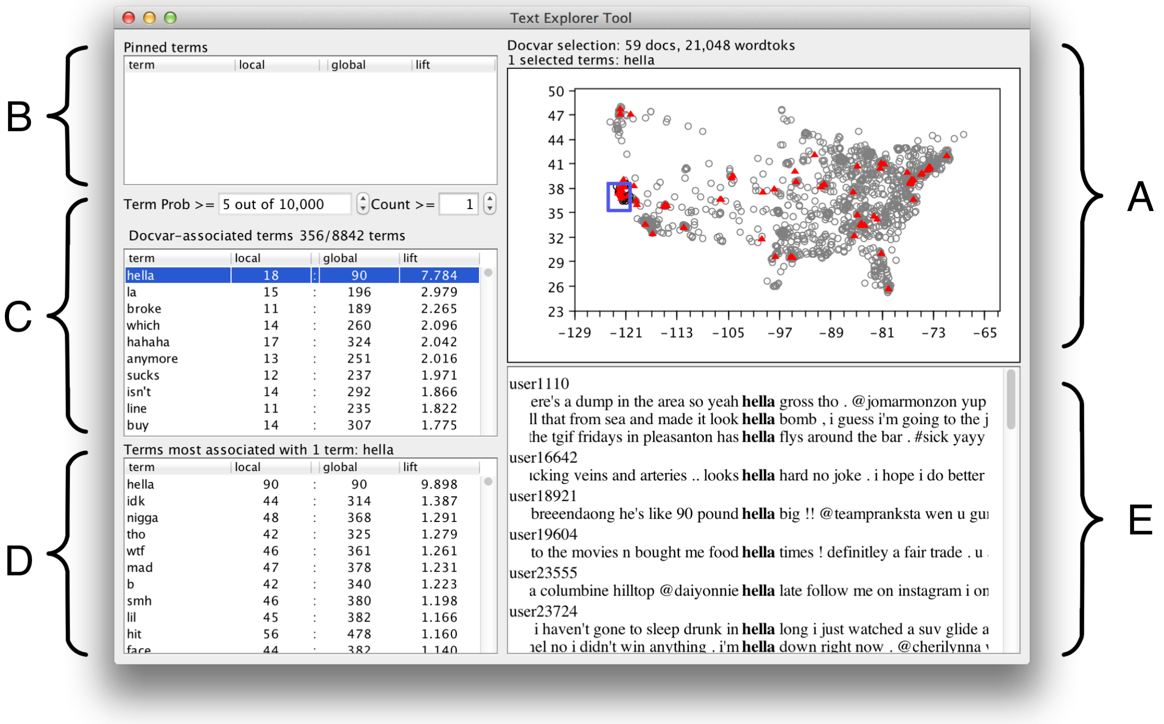
MiTextExplorer: Linked brushing and mutual information for exploratory text data analysis



Pinned terms

term	local		global	lift

Term Prob >= 5 out of 10,000 Count >= 1

Docvar-associated terms 356/8842 terms

term	local		global	lift
hella	18	:	90	7.784
la	15	:	196	2.979
broke	11	:	189	2.265
which	14	:	260	2.096
hahaha	17	:	324	2.042
anymore	13	:	251	2.016
sucks	12	:	237	1.971
isn't	14	:	292	1.866
line	11	:	235	1.822
buy	14	:	307	1.775

Terms most associated with 1 term: hella

term	local		global	lift
hella	90	:	90	9.898
idk	44	:	314	1.387
nigga	48	:	368	1.291
tho	42	:	325	1.279
wtf	46	:	361	1.261
mad	47	:	378	1.231
b	42	:	340	1.223
smh	46	:	380	1.198
lil	45	:	382	1.166
hit	56	:	478	1.160
face	44	:	382	1.140

Docvar selection: 59 docs, 21,048 wordtoks
1 selected terms: hella

user1110
ere's a dump in the area so yeah **hella** gross tho . @jomarmonzon yup
ll that from sea and made it look **hella** bomb , i guess i'm going to the j
the tgif fridays in pleasanton has **hella** flys around the bar . #sick yayy
user16642
ıcking veins and arteries .. looks **hella** hard no joke . i hope i do better
user18921
breeendaong he's like 90 pound **hella** big !! @teampranksta wen u gu
user19604
to the movies n bought me food **hella** times ! definitley a fair trade . u
user23555
a columbine hilltop @daiyonnie **hella** late follow me on instagram i on
user23724
i haven't gone to sleep drunk in **hella** long i just watched a suv glide a
ıel no i didn't win anything . i'm **hella** down right now . @cherilynna

Figure 1: Screenshot of MiTextExplorer, analyzing geolocated tweets.

Brendan O'Connor
Machine Learning Department
Carnegie Mellon University
brenocon@cs.cmu.edu
http://brenocon.com

Abstract

In this paper I describe a preliminary experimental system, MiTextExplorer, for *textual linked brushing*, which allows an analyst to interactively explore statistical relationships between (1) terms, and (2) document metadata (covariates). An analyst can graphically select documents embedded in a temporal, spatial, or other continuous space, and the tool reports terms with strong statistical associations for the region. The user can then drill down to specific term and term groupings, viewing further associations, and see how terms are used in context. The goal is to rapidly compare language usage across interesting document covariates.

I illustrate examples of using the tool on several datasets: geo-located Twitter messages, presidential State of the Union addresses, the ACL Anthology, and the King James Bible.

[This is a preprint/draft, to appear in Proceedings of the ACL 2014, at the Workshop on Interactive Language Learning, Visualization, and Interfaces (June 2014): http://nlp.stanford.edu/events/illvi2014/]

1 Introduction: Can we "just look" at statistical text data?

Exploratory data analysis (EDA) is an approach to extract meaning from data, which emphasizes learning about a dataset through an iterative process of many analyses which suggest and refine possible hypotheses. It is vital in early stages of a data analysis for data cleaning and sanity checks,

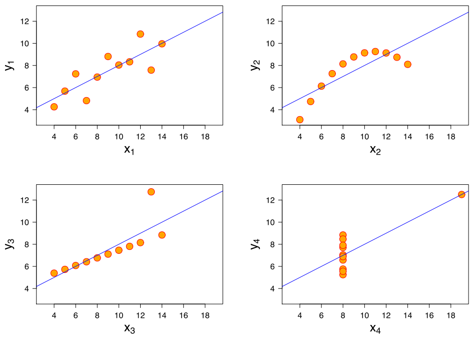


Figure 2: Anscombe Quartet. (Source: Wikipedia)

which are crucial to help ensure a dataset will be useful. Exploratory techniques can also suggest possible hypotheses or issues for further investigation.

The classical approach to EDA, as pioneered in works such as Tukey (1977) and Cleveland (1993) (and other work from the Bell Labs statistics group during that period) emphasizes visual analysis under nonparametric, model-free assumptions, in which visual attributes are a fairly direct reflection of numerical or categorical aspects of data. As a simple example, consider the well-known Anscombe Quartet (1973), a set of four bivariate example datasets. The Pearson correlation, a very widely used measure of dependence that assumes a linear Gaussian model of the data, finds that each dataset has an identical amount of dependence ($r = 0.82$). However, a scatterplot instantly reveals that very different dependence relationships hold in each dataset (Figure 2). The scatterplot is possibly the simplest visual analysis tool for investigating the relationship between two variables, in which the variables' numerical values are mapped to horizontal and vertical space. While the correlation coefficient is a model-based analysis tool, the scatterplot is model-free (or at least, it is effective under an arguably wider range of data generating assumptions), which is crucial for this example.

This nonparametric, visual approach to EDA has been encoded into many data analysis packages, including the now-ubiquitous R language (R Core Team, 2013), which descends from earlier software by the Bell Labs statistics group (Becker and Chambers, 1984). In R, tools such as histograms, boxplots, barplots, dotplots, mosaicplots, etc. are built-in, basic operators in the language. (Wilkinson (2006)'s grammar of graphics more extensively systematizes this approach; see also (Wickham, 2010; Bostock et al., 2011).)

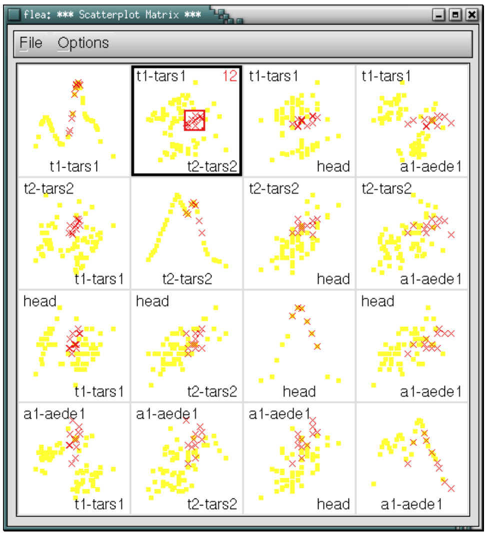


Figure 3: Linked brushing with the analysis software *GGobi*. More references at source: http://www.infovis-wiki.net/index.php?title=Linking_and_Brushing

In the meantime, *textual data* has emerged as a resource of increasing interest for many scientific, business, and government data analysis applications. Consider the use case of automated content analysis (a.k.a. text mining) as a tool for investigating social scientific and humanistic questions (Grimmer and Stewart, 2013; Jockers, 2013; Shaw, 2012; O'Connor et al., 2011). The content of the data is under question: analysts are interested in what/when/how/by-whom different concepts, ideas, or attitudes are expressed in a corpus, and the trends in these factors across time, space, author communities, or other document-level covariates (often called metadata). Comparisons of word statistics across covariates are absolutely essential to many interesting questions or social measurement problems, such as

- What topics tend to get censored by the Chinese government online, and why (Bamman et al., 2012; King et al., 2013)? *Covariates*: whether a message is deleted by censors, time/location of message.

- What drives media bias? Do newspapers slant their coverage in response to what readers want (Gentzkow and Shapiro, 2010)? *Covariates*: political preferences of readers, competitiveness of media markets.

There exist dozens, if not more, of other examples in social scientific and humanities research; see references in O'Connor et al. (2011); O'Connor (2014).

In this work, I focus on the question: What should be the baseline exploratory tools for textual data, to discover important statistical associations between *text* and *document covariates*? Ideally, we'd like to "just look" at the data, in the spirit of scatterplotting the Anscombe Quartet. An analysis tool to support this should not require any statistical model assumptions, and should display the data in as direct a form as possible.

For low-dimensional, non-textual data, the base functionality of R prescribes a broad array of useful defaults: one-dimensional continuous data can be histogrammed (*hist(x)*), or kernel density plotted (*plot(density(x))*), while the relationship between two dimensions of continuous variables can be viewed as a scatterplot (*plot(x,y)*); or perhaps a boxplot for discrete *x* and continous *y* (*boxplot(x,y)*); and so on. Commercial data analysis systems such as Excel, Stata, Tableau, JMP, etc., have similar functionality.

These visual tools can be useful for analyzing derived content statistics from text—for example, showing a high-level topic or sentiment frequency trending over time—but they cannot visualize the text itself. Text data consists of a linear sequence of high-dimensional discrete variables (words). The most aggressive and common analysis approach, bag-of-words, eliminates the problematic sequential structure, by reducing a document to a high-dimensional discrete counts over words. But still, none of the above visual tools makes sense for visualizing a word distribution; many popular tools simply crash or become very slow when given word count data. And besides the issues of discrete high-dimensionality, text is unique in that it has to be manually *read* in order to more reliably understand its meaning. Natural language processing tools can sometimes extract partial views of text meaning, but full understanding is a long ways off; and the quality of available NLP tools varies greatly across corpora and languages. A useful exploratory tool should be able to work with a variety of levels of sophistication in NLP tooling, and allow the user to fall back to manual reading when necessary.

2 MITEXTEXPLORER: linked brushing for text and covariate correlations

The analysis tool presented here, MITEXTEXPLORER, is designed for exploratory analysis of relationships between document covariates—such as time, space, or author community—against textual variables—words, or other units of meaning, that can be counted per document. Unlike topic model approaches to analyzing covariate-text relationships (Mimno, 2012; Roberts et al., 2013), there is no dimension reduction of the terms. Instead, interactivity allows a user to explore more of the high-dimensional space, by specifying a *document selection* (Q) and/or a *term selection* (T). We are inspired by the *linking and brushing* family of techniques in interactive data visualization, in which an analyst can select a group of data points under a query in one covariate space, and see the same data selection in a different covariate space (Figure 3; see Buja et al. (1996), and e.g. Becker and Cleveland (1987); Buja et al. (1991); Martin and Ward (1995); Cook and Swayne (2007)). In our case, one of the variables is text.

The interface consists of several *linked views*, which contain:

(A) a view of the documents in a two-dimensional covariate space (e.g. scatterplot),

(B) an optional list of pinned terms,

(C) *document-associated terms*: a view of the relatively most frequent terms for the current document selection,

(D) *term-associated terms*: a view of terms that relatively frequently co-occur with the current term selection; and

(E) a keyword-in-context (KWIC) display of textual passages for the current term selection.

Figure 1 shows the interface viewing a corpus of 201,647 geo-located Twitter messages from 2,000 users during 2009-2012, which have been tagged with their author's spatial coordinates through a mobile phone client and posted publicly; for data analysis, their texts have been lowercased and tokenized appropriately (Owoputi et al., 2013; O'Connor et al., 2010). Since this type of corpus contains casual, everyday language, it is a dataset that may illuminate geographic patterns of slang and lexical variation in local dialects (Eisenstein et al., 2012, 2010).

The document covariate display (A) uses (longitude, latitude) positions as the 2D space. The corpus has been preprocessed to define a document as the concatenation of messages from a single author, with its position the average location of the author's messages. When the interface loads, all points in (A) are initially gray, and all other panels are blank.

2.1 Covariate-driven queries

A core interaction, *brushing*, consists of using the mouse to select a rectangle in the (x,y) covariate space. Figure 1 shows a selection around the Bay Area metropolitan area (blue rectangle). Upon selection, the document-driven term display (C) is updated to show the relatively most frequent terms in the document selection. Let Q denote the set of documents that are selected by the current covariate query. The tool ranks terms w by their (exponentiated) pointwise mutual information, a.k.a. *lift*, for Q:

$$\text{lift}(w;Q) = \frac{p(w|Q)}{p(w)} \quad \left(= \frac{p(w,Q)}{p(w)p(Q)} \right) \tag{1}$$

This quantity measures how much more frequent the term is in the queryset, compared to the baseline global probability in the corpus ($p(w)$). Probabilities are calculated with simple MLE relative frequencies, i.e.

$$\frac{p(w|Q)}{p(w)} = \frac{\sum_{d \in Q} n_{dw}}{\sum_{d \in Q} n_d} \frac{N}{n_w} \tag{2}$$

where d denotes a document ID, n_{dw} the count of word w in document d, and N the number of tokens in the corpus. PMI gives results that are much more interesting than results from ranking w on raw probability within the query set ($p(w|Q)$), since that simply shows grammatical function words or other terms that are common both in the queryset and across the corpus, and not distinctive for the queryset.[1]

A well-known weakness of PMI is overemphasis on rare terms; terms that appear only in the queryset, even if they appear only once, will attain the highest PMI value. One way to address this is through a smoothing prior/pseudocounts/regularization, or through statistical significance ranking (see §3). For simplicity, we use a minimum frequency threshold filter. The user interface allows minimums for either local or global term frequencies, and to easily adjust them, which naturally shifts the emphasis between specific and generic language. All methods to protect against rare probabilistic events necessarily involve such a tradeoff parameter that the user ought to experiment with; given this situation, we might prefer a transparent mechanism instead of mathematical priors (though see also §3).

Figure 1 shows that *hella* is the highest ranked term for this spatial selection (and freqency threshold), occurring 7.8 times more frequently compared to the overall corpus; this comports with surveyed intuitions of Californian English speakers (Bucholtz et al., 2007). For full transparency to the user, the local and global term counts are shown in the table. (Since *hella* occurred 18 times in the queryset and 90 times globally, this implies the simple conditional probability $p(Q|w) = 18/90$; and indeed, ranking on $p(Q|w)$ is equivalent to ranking on PMI, since exponentiated PMI is $p(Q|w)/p(Q)$.) The user can also sort by local count to see the raw most-frequent term report for the document selection. As the user reshapes the query box, or drags it around the space, the terms in panel (C) are updated.

Not shown are options to change the term frequency representation. For exposition here, probabilities are formulated as counts of tokens, but this can be problematic for social media data, since a single user might use a term a very large number of times. The above analysis is conducted with an indicator representation of terms per user, so all frequencies refer to the probability that a user uses the term at least once. However, the other examples in this paper use token-level frequencies, which seem to work fine. It is an interesting statistical analysis question how to derive a single range of methods to work across these situations.

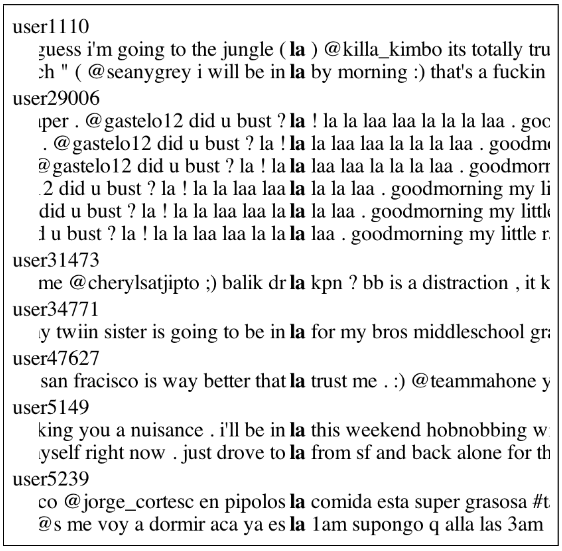
2.2 Term selection and KWIC views

Terms in the table (C) can be clicked and selected, forming a term selection as a set of terms T. This action drives several additional views:

(A) documents containing the term are highlighted in the document covariate display (here, in red),

(E) examples of the term's usage, in Keyword-in-Context style with vertical alignment for the query term; and

[1] The term "lift" is used in business applications (Provost and Fawcett, 2013), while PMI has been used in many NLP applications to measure word associations.



user1110
guess i'm going to the jungle (**la**) @killa_kimbo its totally tru
ch " (@seanygrey i will be in **la** by morning :) that's a fuckin
user29006
per . @gastelo12 did u bust ? **la** ! la la laa laa la la la laa . goc
. @gastelo12 did u bust ? la ! **la** la laa laa la la la laa . goodm
@gastelo12 did u bust ? la ! la **la** laa laa la la la laa . goodmorr
2 did u bust ? la ! la la laa laa **la** la la laa . goodmorning my li
did u bust ? la ! la la laa laa la **la** la laa . goodmorning my littl
d u bust ? la ! la la laa laa la la **la** laa . goodmorning my little r
user31473
me @cherylsatjipto ;) balik dr **la** kpn ? bb is a distraction , it k
user34771
y twiin sister is going to be in **la** for my bros middleschool gr
user47627
san fracisco is way better that **la** trust me . :) @teammahone y
user5149
king you a nuisance . i'll be in **la** this weekend hobnobbing w
yself right now . just drove to **la** from sf and back alone for th
user5239
co @jorge_cortesc en pipolos **la** comida esta super grasosa #t
@s me voy a dormir aca ya es **la** 1am supongo q alla las 3am

Figure 4: KWIC examples of "la" usage in tweets selected in Figure 1.

(D) other terms that frequently co-occur with T (§2.3).

The KWIC report in (E) shows examples of term's usage. For example, why is the term "la" in the PMI list? My initial thought was that this was an example of "LA", short for "Los Angeles". But clicking on "la" instantly disproves this hypothesis—Figure 4, showing the Los Angeles sense, but also the "la la la" sense, as well as the Spanish function word.

The KWIC alignment makes it easier to rapidly browse examples, and think about a rough assessment of their word sense or how they are used. Figure 5 compares how the term "God" is used by U.S. presidents Ronald Reagan and Barack Obama, in a corpus of State of the Union speeches, from two different displays of the tool. The predominant usage is the invocation of "God bless America" or similar, nearly ornamental, expressions, but Reagan also has substantive usages, such as references to the role of religion in schools. The vertical alignments of the right-side context words makes it easy to see the "God bless" word sense. I initially found this example simply by browsing the covariate space, and noticing "god" as a frequent term for Reagan, though still occurring for other presidents; the KWIC drilldown better illuminated these distinctions, and suggests differences in political ideologies between the presidents.

In lots of exploratory text analysis work, especially in the topic modeling literature, it is common to look at word lists produced by a statistical analysis method and think about what they might mean. At least in my experience doing this, I've often found that seeing examples of words in context has disproved my initial intuitions. Hopefully, supporting this activity in an interactive user interface might make exploratory analysis more effective. Currently, the interface simply shows a sample of in-context usages from the document query-set; it would be interesting to perform grouping and stratified sampling based on local contextual statistics. Summarizing local context by frequencies could be done as a trie visualization (Wattenberg and Viégas, 2008); see §5.

2.3 Term-association queries

When a term is selected, its interaction with covariates is shown by highlighting documents in (B) that contain the term. This can be thought of as another document query: instead of being specified as a region in the covariate space, is specified as a fragment of the discrete lexical space. As illustrated in much previous work (e.g. Church and Hanks (1990); Turney (2001, 2002)), word-to-word PMI scores can find other terms with similar meanings, or having interesting semantic relationships, to the target term.[2]

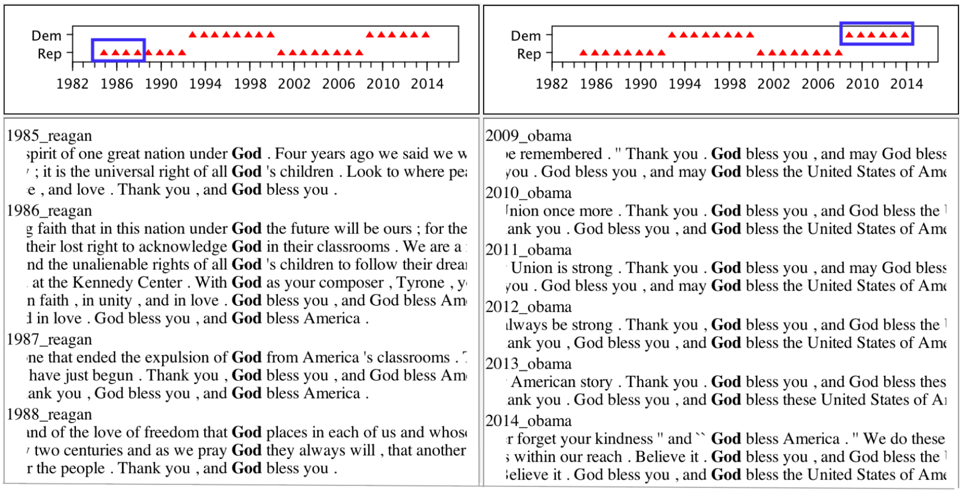
This panel ranks terms u by their association with the query term v. The simplest method is to analyze the relative frequencies of terms in documents that contain v,

$$\text{bool-tt-epmi}(u,v) = \frac{p(w_i = u | v \in \text{supp}(d_i))}{p(w_i = u)}$$

Here, the subscript i denotes a token position in the entire corpus, for which there is a wordtype w_i and a document ID d_i. In this notation, the covariate PMI in 2.1 would be $p(w_i = u | d_i \in Q)/p(w_i = u)$. $\text{supp}(d_i)$ denotes the set of terms that occur at least once in document d_i.

This measure is a very simple extension of the document covariate selection mechanism, and easy to understand. However, it is less satisfying for longer documents, since a larger number of occurrences of v do not lead to a stronger association score. A possible extension is to consider the joint random event of selecting two tokens i

[2] For finding terms with similar semantic meaning, distributional similarity may be more appropriate (Turney and Pantel, 2010); this could be interesting to incorporate into the software.



1985_reagan
spirit of one great nation under **God** . Four years ago we said we w
; it is the universal right of all **God** 's children . Look to where pe
e , and love . Thank you , and **God** bless you .
1986_reagan
g faith that in this nation under **God** the future will be ours ; for the
their lost right to acknowledge **God** in their classrooms . We are a
nd the unalienable rights of all **God** 's children to follow their drea
at the Kennedy Center . With **God** as your composer , Tyrone , y
n faith , in unity , and in love . **God** bless you , and God bless Am
d in love . God bless you , and **God** bless America .
1987_reagan
ne that ended the expulsion of **God** from America 's classrooms .
have just begun . Thank you , **God** bless you , and God bless Am
ank you , God bless you , and **God** bless America .
1988_reagan
nd of the love of freedom that **God** places in each of us and whos
two centuries and as we pray **God** they always will , that another
r the people . Thank you , and **God** bless you .

2009_obama
e remembered . " Thank you . **God** bless you , and may God bless
you . God bless you , and may **God** bless the United States of Am
2010_obama
nion once more . Thank you . **God** bless you , and God bless the
ank you . God bless you , and **God** bless the United States of Am
2011_obama
Union is strong . Thank you . **God** bless you , and may God bless
you . God bless you , and may **God** bless the United States of Am
2012_obama
lways be strong . Thank you , **God** bless you , and God bless the
ank you , God bless you , and **God** bless the United States of Am
2013_obama
American story . Thank you . **God** bless you , and God bless thes
ank you . God bless you , and **God** bless these United States of A
2014_obama
r forget your kindness " and `` **God** bless America . " We do these
s within our reach . Believe it . **God** bless you , and God bless the
elieve it . God bless you , and **God** bless the United States of Am

Figure 5: KWIC examples of "God" in speeches by Reagan versus Obama.

and j in the corpus, and consider if the two tokens being in the same document is informative for whether the tokens are the words (u, v), i.e. PMI$[(w_i, w_j) = (u, v); d_i = d_j]$,

$$\text{freq-tt-epmi}(u, v) = \frac{p(w_i = u, w_j = v | d_i = d_j)}{p(w_i = u, w_j = v)}$$

In terms of word counts, this expression has the form

$$\text{freq-tt-epmi}(u, v) = \frac{\sum_d n_{du} n_{dv}}{n_u n_v} \frac{N^2}{\sum_d n_d^2}$$

The right-side term is a normalizing constant invariant to u and v. The left-side term is interesting: it can be viewed as a similarity measure, where the numerator is the inner product of the inverted term-document vectors $n_{.,u}$ and $n_{.,v}$, and the denominator is the product of their ℓ_1 norms. This is a very similar form as cosine similarity, which is another normalized inner product, except its denominator is the product of the vectors' ℓ_2 norms.

Term-to-term associations allow a navigation of the term space, complementing the views of terms driven by document covariates. This part of the tool is still at a more preliminary stage of development. One important enhancement would be adjustment of the context window size allowed for co-occurrences; the formulations above assume a context window the size of the document. Medium sized context windows might capture more focused topical content, especially in very long discourses such as speeches; and the smallest context windows, of size 1, should be more like collocation detection (though see §3; this is arguably better done with significance tests, not PMI).

2.4 Pinned terms

The term PMI views of (C) and (D) are very dynamic, which can cause interesting terms to disappear when their supporting query is changed. It is often useful to select terms to be constantly viewed when the document covariate queries change.

Any term can be double-clicked to be moved to the the table of *pinned terms* (B). The set of terms here does not change as the covariate query is changed; a user can fix a set of terms and see how their PMI scores change while looking at different parts of the covariate space. One possible use of term pinning is to manually build up clusters of terms—for example, topical or synonymous term sets—whose aggregate statistical behavior (i.e. as a disjunctive query) may be interesting to observe. Manually built sets of keywords are a very useful form of text analysis; in fact, the WordSeer corpus analysis tool has explicit support to help users create them (Shrikumar, 2013).

3 Statistical term association measures

There exist many measures to measure the statistical strength of an association between a term and a document covariate, or between two terms. A number of methods are based on significance testing, looking for violations of a null hypothesis that term frequencies are independent. For collocation detection, which aims to find meaningful non-compositional lexical items through frequencies of neighboring words, likelihood ratio (Dunning, 1993) and chi-square tests have been used (see review in Manning and Schütze (1999)). For term-covariate associations, chi-square tests were used by Gentzkow and Shapiro (2010) to find politically loaded phrases often used by members of one political party; this same method is often used as a feature selection method for supervised learning (Guyon and Elisseeff, 2003).

The approach we take here is somewhat different, being a point estimate approach, analyzing the estimated difference (and giving poor results when counts are small). Some related work for topic model analysis, looking at statistical associations between words and latent topics (as opposed to between words and observed covariates in this work) includes Chuang et al. (2012b), whose term saliency function measures one word's associations against all topics; a salient term tends to have most of its probability mass in a small set of topics. The measure is a form of mutual information,[3] and may be useful for our purposes here if the user wishes to see a report of distinctive terms for a group of several different observed covariate values at once. Blei and Lafferty (2009) ranks words per topic by a measure inspired by TFIDF, which like PMI downweights words that are generically common across all topics.

Finally, hierarchical priors and regularizers can also be used; for example, by penalizing the log-odds parameterization of term probabilities (Eisenstein et al., 2011; Taddy, 2013). These methods are better in that they incorporate both protection against small count situations, while paying attention to effect size, as well as allowing overlapping covariates and regression control variables; but unfortunately, they are more computationally intensive, as opposed to the above measures which all work directly from sufficient count statistics. An association measure that fulfilled all these desiderata would be very useful. For term-covariate analysis, Monroe et al. (2008) contains a review of many different methods, from both political science as well as computer science; they also propose a hierarchical prior method, and to rank by statistical significance via the asymptotic standard error of the terms' odds ratios.

Given the large amount of previous work using the significance approach, it merits further exploration for this system.

4 Phrase selection

The simplest approach to defining the terms is to use all words (unigrams). This can be insightful, but single words are both too coarse and too narrow a unit of analysis. They can be too narrow when there are multiple ways of saying the same thing, such as synonyms—for example, while we have evidence about differing usages of the term "god" in presidential rhetoric, in order to make a claim about religious themes, we might need to find other terms such as "creator", "higher power", etc. Another problematic case is alternate names or anaphoric references to an entity. In general, any NLP tool that extracts interesting discrete variable indicators of word meaning could be used for mutual information and covariate exploratory analysis—for example, a coreference system's entity ID predictions could be browsed by the system as the term variables. (More complex concepts, of course, would also require more UI support.)

At the same time, words can be too coarse compared to the longer phrases they are contained within, which often contain more interesting and distinctive concepts: for example, "death tax" and "social security" are important concepts in U.S. politics that get missed under a unigram analysis. In fact, Sim et al. (2013)'s analysis of U.S. politicians' speeches found that domain experts had a hard time understanding unigrams out-of-context, but bigrams and trigrams worked much better; Gentzkow and Shapiro (2010) similarly focus on partisan political phrases.

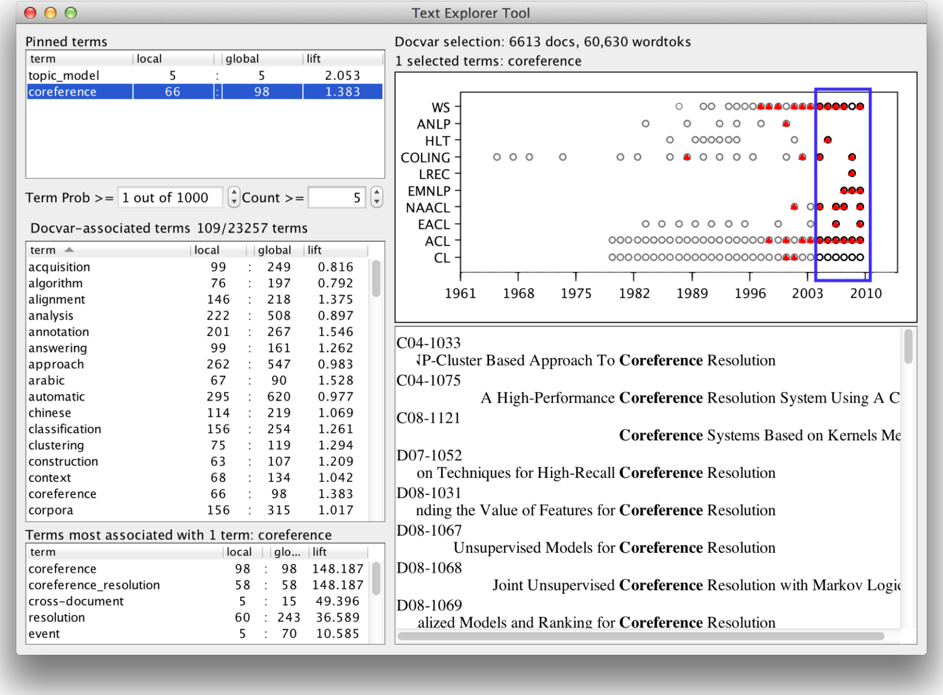
It sometimes works to simply add overlapping n-grams as more terms, but sometimes odd phrases get selected that cross constituent boundaries from their source sentences, and are thus not totally meaningful. I've experimented with a very strong filtering approach to phrase selection: besides using all unigrams, take all n-grams up to length 5 that have nominal part-of-speech patterns: either the sequence consists of zero or more adjectives followed by one or more noun tokens, or all tokens were classified as names by a named entity recognition system.[4] This tends to yield

[3] This is apparent as follows, using notation from their section 3.1:

$$\begin{aligned}\text{saliency}(w) &= p(w)\sum_T p(T|w)\log[p(T|w)/p(T)] \\ &= \sum_T p(w,T)\log[p(w,T)/[p(w)p(T)]]\end{aligned}$$

This might be called a "half-pointwise" mutual information: between a specific word w and the topic random variable T. Mutual information is $\sum_w \text{saliency}(w)$.

[4] For traditional text, the tool currently uses Stanford CoreNLP; for Twitter, CMU ARK TweetNLP.



Pinned terms

term	local		global	lift
topic_model	5	:	5	2.053
coreference	66	:	98	1.383

Term Prob >= 1 out of 1000 Count >= 5

Docvar-associated terms 109/23257 terms

term	local		global	lift
acquisition	99	:	249	0.816
algorithm	76	:	197	0.792
alignment	146	:	218	1.375
analysis	222	:	508	0.897
annotation	201	:	267	1.546
answering	99	:	161	1.262
approach	262	:	547	0.983
arabic	67	:	90	1.528
automatic	295	:	620	0.977
chinese	114	:	219	1.069
classification	156	:	254	1.261
clustering	75	:	119	1.294
construction	63	:	107	1.209
context	68	:	134	1.042
coreference	66	:	98	1.383
corpora	156	:	315	1.017

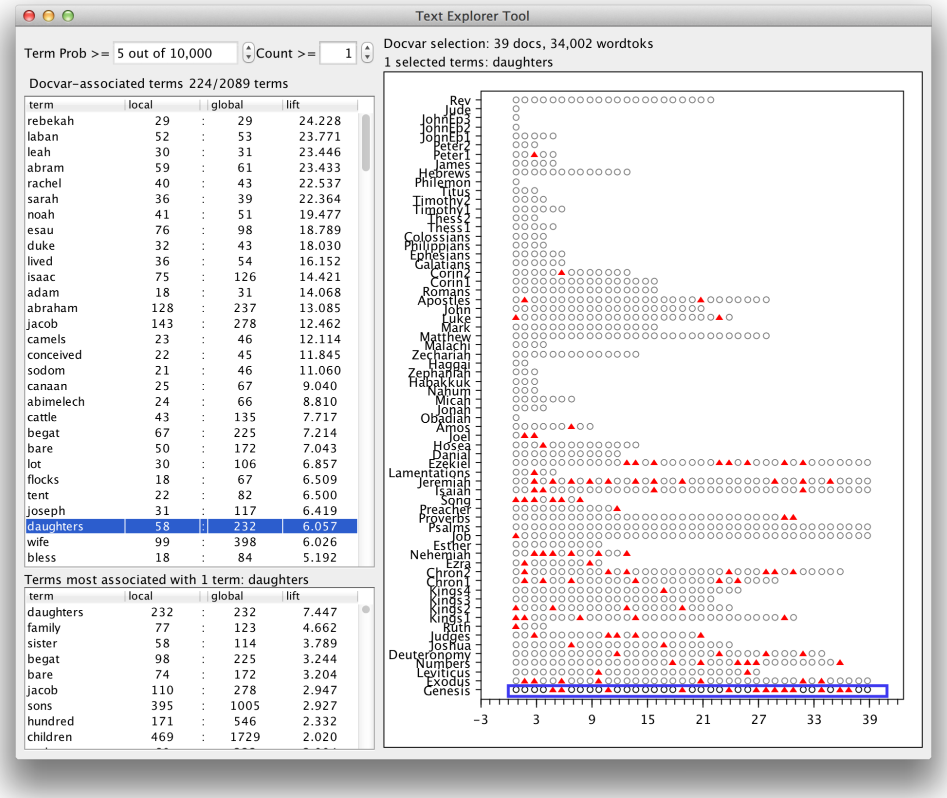
Terms most associated with 1 term: coreference

term	local		glo...	lift
coreference	98	:	98	148.187
coreference_resolution	58	:	58	148.187
cross-document	5	:	15	49.396
resolution	60	:	243	36.589
event	5	:	70	10.585

Docvar selection: 6613 docs, 60,630 wordtoks
1 selected terms: coreference

C04-1033
ꞮP-Cluster Based Approach To **Coreference** Resolution
C04-1075
A High-Performance **Coreference** Resolution System Using A C
C08-1121
Coreference Systems Based on Kernels Me
D07-1052
on Techniques for High-Recall **Coreference** Resolution
D08-1031
nding the Value of Features for **Coreference** Resolution
D08-1067
Unsupervised Models for **Coreference** Resolution
D08-1068
Joint Unsupervised **Coreference** Resolution with Markov Logi
D08-1069
alized Models and Ranking for **Coreference** Resolution

Figure 6: MITEXTEXPLORER for paper titles in the ACL Anthology (Radev et al., 2009). Y-axis is venue (conference or journal name), X-axis is year of publication. Unlike the other figures, docvar-associated terms are sorted alphabetically.



Term Prob >= 5 out of 10,000 Count >= 1

Docvar-associated terms 224/2089 terms

term	local		global	lift
rebekah	29	:	29	24.228
laban	52	:	53	23.771
leah	30	:	31	23.446
abram	59	:	61	23.433
rachel	40	:	43	22.537
sarah	36	:	39	22.364
noah	41	:	51	19.477
esau	76	:	98	18.789
duke	32	:	43	18.030
lived	36	:	54	16.152
isaac	75	:	126	14.421
adam	18	:	31	14.068
abraham	128	:	237	13.085
jacob	143	:	278	12.462
camels	23	:	46	12.114
conceived	22	:	45	11.845
sodom	21	:	46	11.060
canaan	25	:	67	9.040
abimelech	24	:	66	8.810
cattle	43	:	135	7.717
begat	67	:	225	7.214
bare	50	:	172	7.043
lot	30	:	106	6.857
flocks	18	:	67	6.509
tent	22	:	82	6.500
joseph	31	:	117	6.419
daughters	58	:	232	6.057
wife	99	:	398	6.026
bless	18	:	84	5.192

Terms most associated with 1 term: daughters

term	local		global	lift
daughters	232	:	232	7.447
family	77	:	123	4.662
sister	58	:	114	3.789
begat	98	:	225	3.244
bare	74	:	172	3.204
jacob	110	:	278	2.947
sons	395	:	1005	2.927
hundred	171	:	546	2.332
children	469	:	1729	2.020

Docvar selection: 39 docs, 34,002 wordtoks
1 selected terms: daughters

Figure 7: MITEXTEXPLORER for the King James Bible. Y-axis is book, X-axis is chapter (truncated to 39).

(partial) constituents, and nouns tend to be more interesting than other content words (perhaps because they are relatively less reliant on predicate-argument structure to express their semantics—as opposed to adjectives or verbs, say—and a bag-of-terms analysis does not allow expression of argument structure.) However, for many corpora, POS or NER taggers work poorly—for example, I've seen paper titles from the ACL Anthology have capitalized prepositions tagged as names—so simpler stopword heuristics are necessary.

The phrase selection approach could be improved in many ways; for example, a real noun phrase recognizer could get important (NP PP) constructs like "war on terror." Furthermore, Chuang et al. (2012a) find that while these sorts of syntactic features are helpful in choosing useful keyphrases to summarize of scientific abstracts, it is also very useful to add in collocation detection scores. Similarly to the PMI calculations used here, likelihood ratio or chi-square collocation detection statistics are also very rapid to compute and may benefit from interactive adjustment of decision thresholds. More generally, any type of lexicalized linguistic structures could potentially be used, such as dependency paths or constituents from a syntactic parser, or predicate-argument structures from a semantic parser. Linguistic structures extracted from more sophisticated NLP tools may indeed be better-generalized units of linguistic meaning compared to words and phrases, but they will still bear the same high-dimensionality issues for data analysis purposes.

5 Related work: Exploratory text analysis

Many systems and techniques have been developed for interactive text analysis. Two such systems, WordSeer and Jigsaw, have been under development for several years, each having had a series of user experiments and feedback. Recent and interesting review papers and theses are available for both of them.

The WordSeer system (Shrikumar, 2013)[5] contains many different interactive text visualization tools, including syntax-based search, and was initially designed for the needs of text analysis in the humanities; the WordSeer 3.0 system includes a word frequency analysis component that can compare word frequencies along document covariates. Interestingly, Shrikumar found in user studies with literary experts that data comparisons and annotation/note-taking support were very important capabilities to add to the system. Unique to the work in this paper is the emphasis on conditioning on document covariates to analyze relative word frequencies, and encouraging the user to change the statistical parameters that govern text correlation measurements. (The term pinning and term-to-term association techniques are certainly less developed than previous work.)

Another text analysis system is Jigsaw (Görg et al., 2013),[6] originally developed for investigative analysis (as in law enforcement or intelligence), which again has many features. It emphasizes visualizations based on entity extractions, such as for names, places, and dates. Görg et al. note that errors in entity extraction were a major problem for users; this might be a worthwhile argument to focus on getting something to first work with simple words/phrases before tackling more complex units of meaning. A section of the review paper is entitled "Reading the documents still matters", pointing out that analysts did not want just to visualize high-level relationships, but also wanted to read documents in context; this capability was added to later versions of Jigsaw, and supports the emphasis here on the KWIC display.

Both these systems also use variants of Wattenberg and Viégas (2008)'s word tree visualization, which gives a sequential word frequencies as a tree (i.e., what computational linguists might call a trie representation of a high-order Markov model). The "God bless" word sense example from §2 indicates that such statistical summarization of local contextual information may be useful to integrate; it is worth thinking how to integrate this against the important need of document covariate analysis, while being efficient with the use of space.

Many other systems, especially ones designed for literary content analysis, emphasize concordances and keyword searches within a text; for example, Voyeur/Voyant (Rockwell et al., 2010),[7] which also features some document covariate analysis through temporal trend analyses for individual terms. Another class of approaches emphasizes the use of document clustering or topic models (Gardner et al., 2010; Newman et al., 2010;

[5] http://wordseer.berkeley.edu/

[6] http://www.cc.gatech.edu/gvu/ii/jigsaw/

[7] http://voyant-tools.org/, http://hermeneuti.ca/voyeur

Grimmer and King, 2011; Chaney and Blei, 2013), while Overview[8] emphasizes hierarchical document clustering paired with manual tagging.

Finally, considerable research has examined exploratory visual interfaces for information retrieval, in which a user specifies an information need in order to find relevant documents or passages from a corpus (Hearst (2009), Ch. 10). Information retrieval problems have some similarities to text-as-data analysis in the need for an exploratory process of iterative refinement, but the text-as-data perspective differs in that it requires an analyst to understand content and contextual factors across multiple or many documents.

6 Future work

The current MiTextExplorer system is an extremely simple prototype to explore what sorts of "bare words" text-and-covariates analyses are possible. Several major changes will be necessary for more serious use.

First, essential basic capabilities must be added, such as a search box the user can use to search and filter the term list.

Second, the document covariate display needs to support more than just scatterplots. When there are hundreds or more documents, summarization is necessary in the form of histograms, kernel density plots, or other tools. For example, for a large corpus of documents over time, a lineplot or temporal histogram is more appropriate, where each timestep has a document count. The ACL Anthology scatterplot (Figure 6, Radev et al. (2009)), which has hundreds of overplotted points at each (year,venue) position, makes clear the limitations of the current approach.

Better visual feedback for term selections here could be useful—for example, sizing document points monotonically with the term's frequency (rather than just presence/absence), or using stacked line plots—though certain visual depictions of frequency may be difficult given the Zipfian distribution of word frequencies.

Furthermore, document structures may be thought of as document covariates. A single book has interesting internal variation that could be analyzed itself. Figure 7 shows the King James Bible, which has a hierarchical structure of book, chapter, and verse. Here, the (y,x) coordinates represent books and chapters. A more specialized display for book-level structures, or other discourse structures, may be appropriate for book-length texts.

Finally, a major goal of this work is to use analysis methods that can be computed on the fly, but the current prototype only works with small datasets. Hierarchical spatial indexing techniques (e.g. r-trees), may make it possible to interactively compute sums for covariate PMI scoring over very large numbers of documents. Text indexing is also important for term-driven queries and KWIC views. Techniques from ad-hoc data querying systems may be necessary for further scale (e.g. Melnik et al. (2010)).

Many other directions are possible. The prototype tool, as described in §2, will be available as open-source software at: http://brenocon.com/MiTextExplorer. It is a desktop application written in Java.

Acknowledgments

Thanks to Michael Heilman and Bryan Routledge, for many discussions and creative text analysis scripts that inspired this work. Thanks also to the anonymous reviewers for very helpful feedback. This research was supported in part by NSF grant IIS-1211277 and CAREER grant IIS-1054319.

References

Francis J Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

David Bamman, Brendan O'Connor, and Noah A. Smith. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3), 2012.

Richard A Becker and John M Chambers. *S: an interactive environment for data analysis and graphics*. CRC Press, 1984.

Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2): 127–142, 1987.

David M. Blei and John D. Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10:71, 2009.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. URL http://vis.stanford.edu/papers/d3.

[8] https://www.overviewproject.org/ http://overview.ap.org/

Mary Bucholtz, Nancy Bermudez, Victor Fung, Lisa Edwards, and Rosalva Vargas. Hella Nor Cal or totally So Cal? the perceptual dialectology of California. *Journal of English Linguistics*, 35(4):325–352, 2007. URL http://people.duke.edu/~eec10/hellanorcal.pdf.

Andreas Buja, John Alan McDonald, John Michalak, and Werner Stuetzle. Interactive data visualization using focusing and linking. In *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on*, pages 156–163. IEEE, 1991.

Andreas Buja, Dianne Cook, and Deborah F Swayne. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5(1):78–99, 1996.

Allison J.B. Chaney and David M. Blei. Visualizing topic models. In *Proceedings of ICWSM*, 2013.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. "without the clutter of unimportant words": Descriptive keyphrases for text visualization. *ACM Trans. on Computer-Human Interaction*, 19:1–29, 2012a. URL http://vis.stanford.edu/papers/keyphrases.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012b. URL http://vis.stanford.edu/papers/termite.

K. W Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):2229, 1990.

William S. Cleveland. *Visualizing data*. Hobart Press, 1993.

Dianne Cook and Deborah F. Swayne. *Interactive and dynamic graphics for data analysis: with R and GGobi*. Springer, 2007.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61—74, 1993. doi: 10.1.1.14.5962. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.5962.

J. Eisenstein, A. Ahmed, and E.P. Xing. Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048, 2011.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277—1287, 2010.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. Mapping the geographical diffusion of new words. In *NIPS Workshop on Social Network and Social Media Analysis*, 2012. URL http://arxiv.org/abs/1210.5268.

M.J. Gardner, J. Lutes, J. Lund, J. Hansen, D. Walker, E. Ringger, and K. Seppi. The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization. MIT Press*, 2010.

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Carsten Görg, Zhicheng Liu, and John Stasko. Reflections on the evolution of the jigsaw visual analytics system. *Information Visualization*, 2013.

Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011.

Justin Grimmer and Brandon M Stewart. Text as Data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013. URL http://www.stanford.edu/~jgrimmer/tad2.pdf.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.

Matthew L Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

Gary King, Jennifer Pan, and Margaret E. Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107:1–18, 2013.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Allen R. Martin and Matthew O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the 6th Conference on Visualization'95*, page 271. IEEE Computer Society, 1995.

Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, 3(1-2):330–339, 2010.

David Mimno. *Topic regression*. PhD thesis, University of Massachusetts Amherst, 2012.

B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin'Words: lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372, 2008.

D. Newman, T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2):169–175, 2010.

Brendan O'Connor. *Statistical Text Analysis for Social Science*. PhD thesis, Carnegie Mellon University, 2014.

Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

Brendan O'Connor, David Bamman, and Noah A. Smith. Computational text analysis for social science: Model assumptions and complexity. In *Second Workshop on Comptuational Social Science and the Wisdom of Crowds (NIPS 2011)*, 2011.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, 2013.

Foster Provost and Tom Fawcett. *Data Science for Business*. O'Reilly Media, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. The ACL anthology network corpus. In *Proc. of ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, 2009.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. Structural topic models. 2013. URL http://scholar.harvard.edu/bstewart/publications/structural-topic-models. Working paper.

Geoffrey Rockwell, Stéfan G Sinclair, Stan Ruecker, and Peter Organisciak. Ubiquitous text analysis. *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture*, 2 (1), 2010.

Ryan Shaw. Text-mining as a research tool, 2012. URL http://aeshin.org/textmining/.

Aditi Shrikumar. *Designing an Exploratory Text Analysis Tool for Humanities and Social Sciences Research*. PhD thesis, University of California at Berkeley, 2013.

Yanchuan Sim, Brice Acree, Justin H Gross, and Noah A Smith. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*, 2013.

Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

John W. Tukey. *Exploratory data analysis*. 1977.

P. D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 417424, 2002.

P. D Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37 (1):141188, 2010. ISSN 1076-9757.

Peter Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the Twelth European Conference on Machine Learning*, 2001. URL http://nparc.cisti-icist.

nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=5765594.

Martin Wattenberg and Fernanda B Viégas. The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1221–1228, 2008.

Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):328, 2010. doi: 10.1198/jcgs.2009.07098.

Leland Wilkinson. *The grammar of graphics*. Springer, 2006.