Combining latent topics with document attributes in text analysis

Nelson Auner
Prof. Matt Taddy[1], Prof. Stephen Stigler[2]

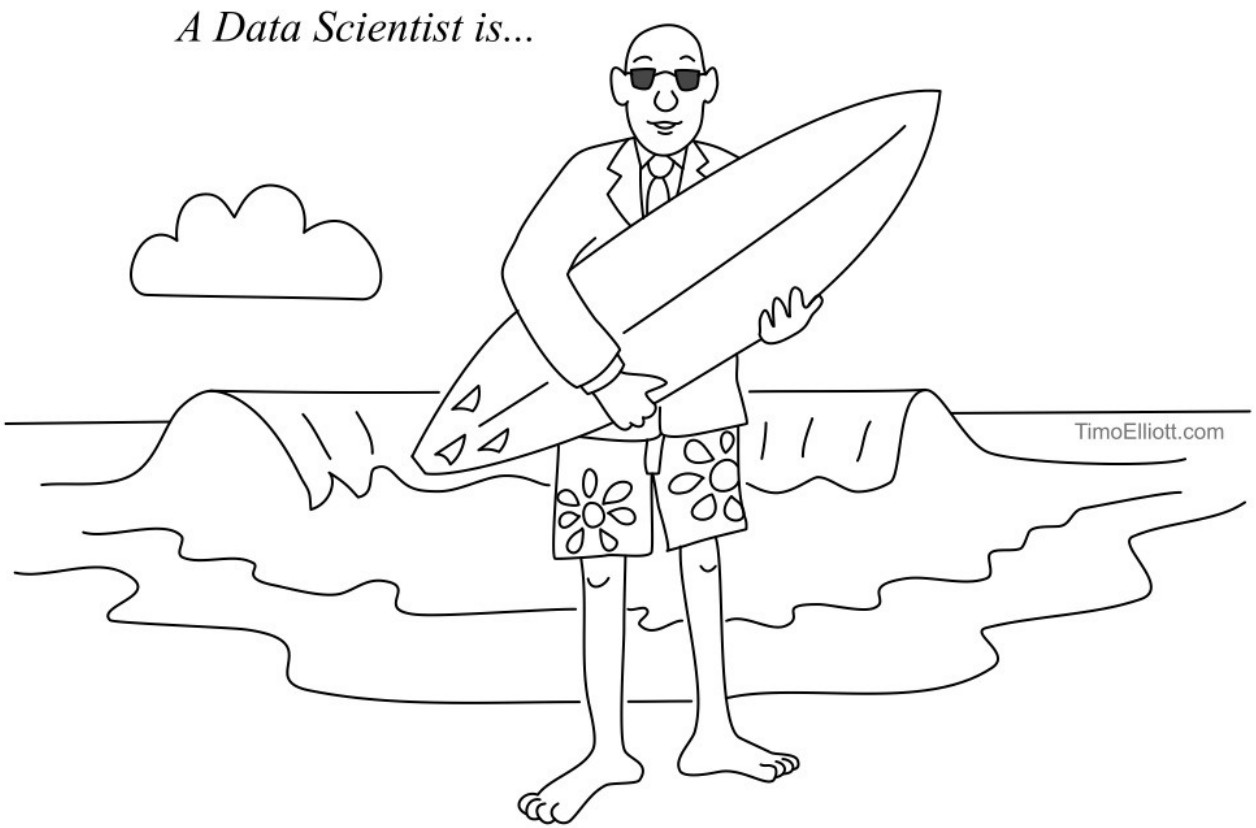
University of Chicago

June 16, 2014

[1]Associate Professor of Econometrics and Statistics at Chicago Booth School of Business

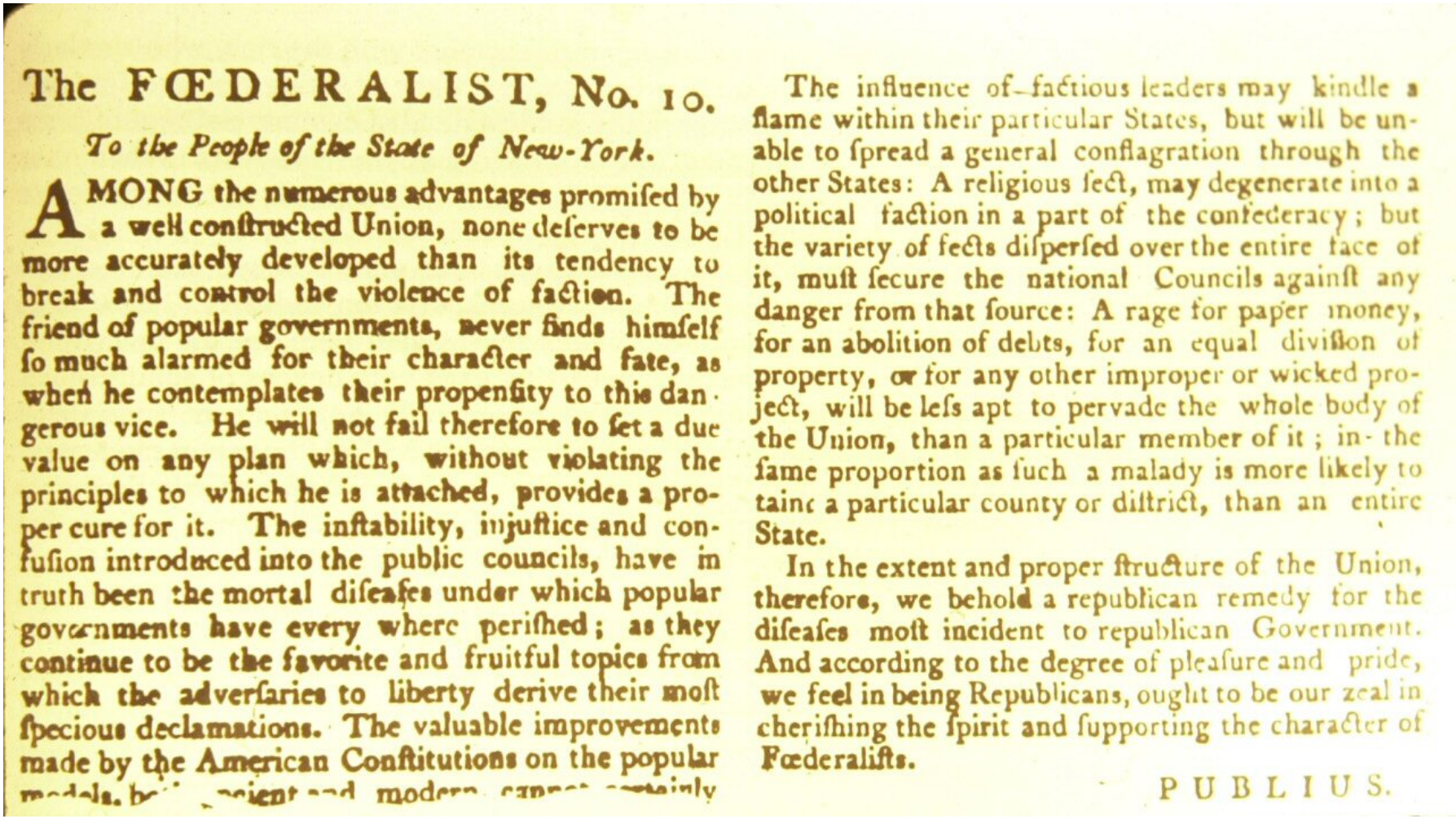
[2]Ernest DeWitt Burton Distinguished Service Professor at the Department of Statistics of the University of Chicago

What is a data scientist?



A Business Analyst that lives in California.

Motivation: Political Science



The FŒDERALIST, No. 10.

To the People of the State of New-York.

AMONG the numerous advantages promiſed by a well conſtructed Union, none deſerves to be more accurately developed than its tendency to break and control the violence of faction. The friend of popular governments, never finds himſelf ſo much alarmed for their character and fate, as when he contemplates their propenſity to this dangerous vice. He will not fail therefore to ſet a due value on any plan which, without violating the principles to which he is attached, provides a proper cure for it. The inſtability, injuſtice and confuſion introduced into the public councils, have in truth been the mortal diſeaſes under which popular governments have every where periſhed; as they continue to be the favorite and fruitful topics from which the adverſaries to liberty derive their moſt ſpecious declamations. The valuable improvements made by the American Conſtitutions on the popular models, both ancient and modern, cannot certainly

The influence of factious leaders may kindle a flame within their particular States, but will be unable to ſpread a general conflagration through the other States: A religious ſect, may degenerate into a political faction in a part of the confederacy; but the variety of ſects diſperſed over the entire face of it, muſt ſecure the national Councils againſt any danger from that ſource: A rage for paper money, for an abolition of debts, for an equal diviſion of property, or for any other improper or wicked project, will be leſs apt to pervade the whole body of the Union, than a particular member of it; in the ſame proportion as ſuch a malady is more likely to taint a particular county or diſtrict, than an entire State.

In the extent and proper ſtructure of the Union, therefore, we behold a republican remedy for the diſeaſes moſt incident to republican Government. And according to the degree of pleaſure and pride, we feel in being Republicans, ought to be our zeal in cheriſhing the ſpirit and ſupporting the character of Fœderaliſts.

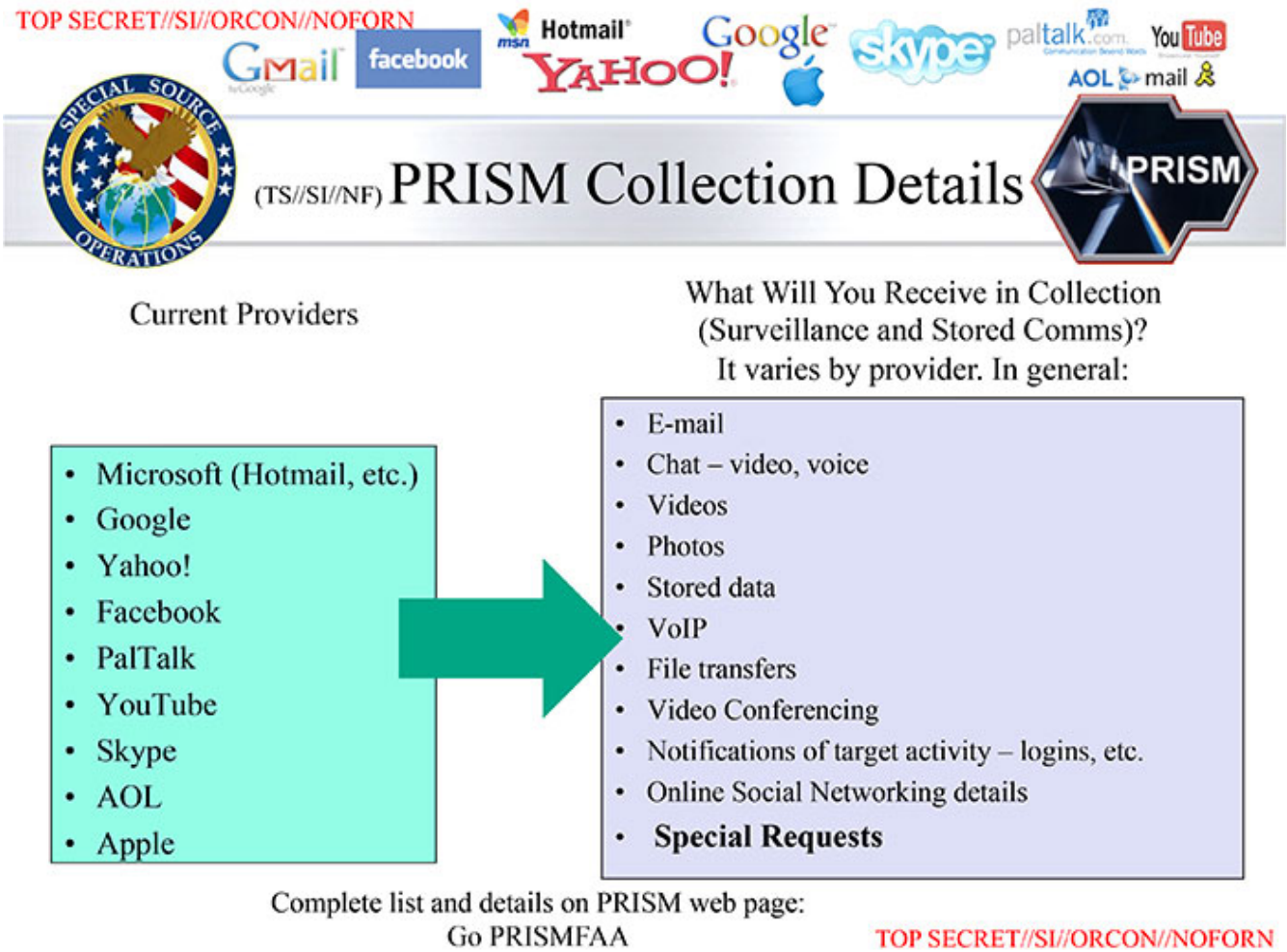
PUBLIUS.

Motivation: Finance

Twitter Can Predict The Stock Market, If You're Reading The Right Tweets

In a world where one tweet can send Wall Street into a panic, social analytics company Dataminr tries to be there first, scanning all of Twitter to find individual messages with the right combination of language, context, and location that might end up being breaking–and money-making–news.

Motivation: Public Service



Text as Data

- A document is a collection of words or phrases.
- Our datasets are collections of documents

Table: What did homework consist of?

Document	Content
1	Some computation and formula proving, a lot of R code
2	Problems, computation using R
3	Some computations and writing R code
4	Proofs, problems, and programming work

Multinomial Models

- If order doesn't matter, then we can treat each document as a "bag of words".
- The number of words can be modeled $\sim$ multinomial

Table: Creating a word-count matrix from text

Document	Some	comp	formula	prov	R	code	use	problem	writ	program	work
1	1	1	1	1	1	1	0	0	0	0	0
2	0	1	0	0	1	0	1	1	0	0	0
3	1	1	0	0	1	0	0	0	1	0	0
4	0	0	0	1	0	0	0	1	0	1	1

A better model: Metadata

- We would like to add structure to the model for inference or prediction
- Metadata is data that accompanies a document

Table: What did homework consist of?

Grade	**Content**
A+	Some computation and formula proving, a lot of R code
B	Problems, computation using R
B	Some computations and writing R code
C+	Proofs, problems, and programming work

Metadata and Computation

- n documents with metadata that takes m discrete values:
- Normally, $n >> m$
- $\Rightarrow$ Collapse observations by outcome variables.
- Model as m observations, instead of n

Document	Some	comp	formula	prov	R	code	use	problem	writ	program	work
A+	1	1	1	1	1	1	0	0	0	0	0
B	1	2	0	0	2	0	1	1	1	0	0
C	0	0	0	1	0	0	0	1	0	1	1

Reality: There are thousands of course reviews

Topic Models

A topic is a distribution of words.
In a topic model, documents are made of a mixtures of topics.

Running Topic
Stride, Pacing, Stretch

Bike Topic
Pedal, Helmet, Gears

Swimming
Stroke, Air, Water

- A book about triathalon training $\sim \theta_1$ Running $+ \theta_2$ Biking $+ \theta_3$ Swimming
- Problem: We can no longer collapse observations, must use all n observations

Cluster Model

Goal

- Want to use the Topic Model but incorporate Metadata
- Also want computational ease

Approach

- Restrict each document to only one topic $\Rightarrow$ "cluster"
- Can collapse observations over unique (metadata, cluster) combination
- $x_i \sim MN(q_{ij}, m_{ij})$; $\quad q_{ij} = \frac{exp(\alpha_j + y_i\phi_j + u_i\Gamma_{kj})}{\sum_{l=1}^{P} exp(\alpha_l + y_i\phi_l + u_i\Gamma_{kl})}$

Algorithm for Cluster Membership Model with Gamma Lasso Penalty

1. Initialize cluster membership u_i for $i = 1, \ldots, n$
2. Determine parameters α, ϕ, Γ by fitting a multinomial regression on $y_i|x_i, u_i$ with a gamma lasso penalty (Taddy 2013)
3. For each document i, determine new cluster u_i membership as $argmax_{k=1,..,K}\left[\ell(u_i|\alpha, \phi, \Gamma)\right]$
4. Check if current cluster assignment is different from previous cluster assignment , $(\mathbf{u}^{(t)} = \mathbf{u}^{(t-1)})$.If so, return to step 2. If not, end algorithm.

Congressional Speech and Restaurant Reviews

- We apply the algorithm to two datasets:
  - Congressional Speech records (Gentzkow and Shapiro, 2010)
  - A corpus of restaurant reviews called we8there.
- Questions:
  - Can this simple model capture the variation explained by a topic model?
  - How does choice of cluster initialization affect the fit?

An Example Cluster

	term	loading
1	nation.oil.food	20.09
2	united.nation.oil	12.09
3	liberty.pursuit.happiness	8.11
4	life.liberty.pursuit	8.11
5	minority.women.owned	6.73
6	universal.health	6.67
7	white.care.act	6.64
8	ryan.white.care	6.6
9	universal.health.care	5.99
10	growth.job.creation	5.39
11	drilling.arctic.national	5.3
12	tax.relief.package	5.29
13	judge.john.robert	5.26
14	fre.enterprise	5.07
15	arctic.refuge	4.93

Comparison with the Topic Model

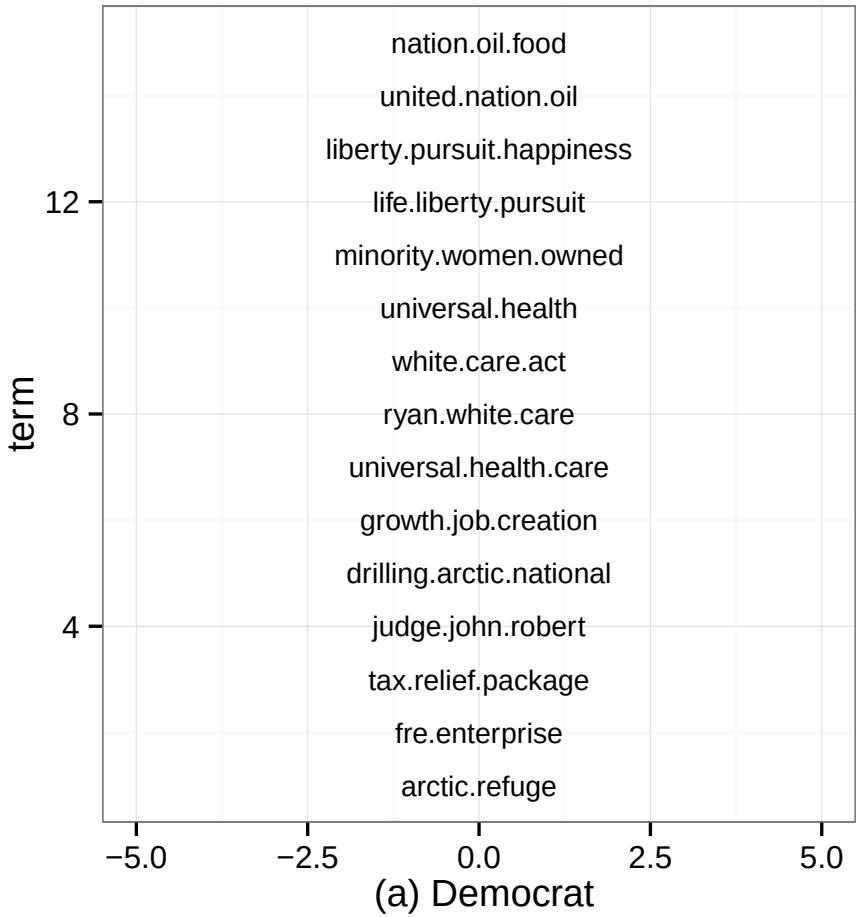
Good news: We are able to recover similar topics with our model:

Table: Comparison of top word loadings on a stem-cell topic

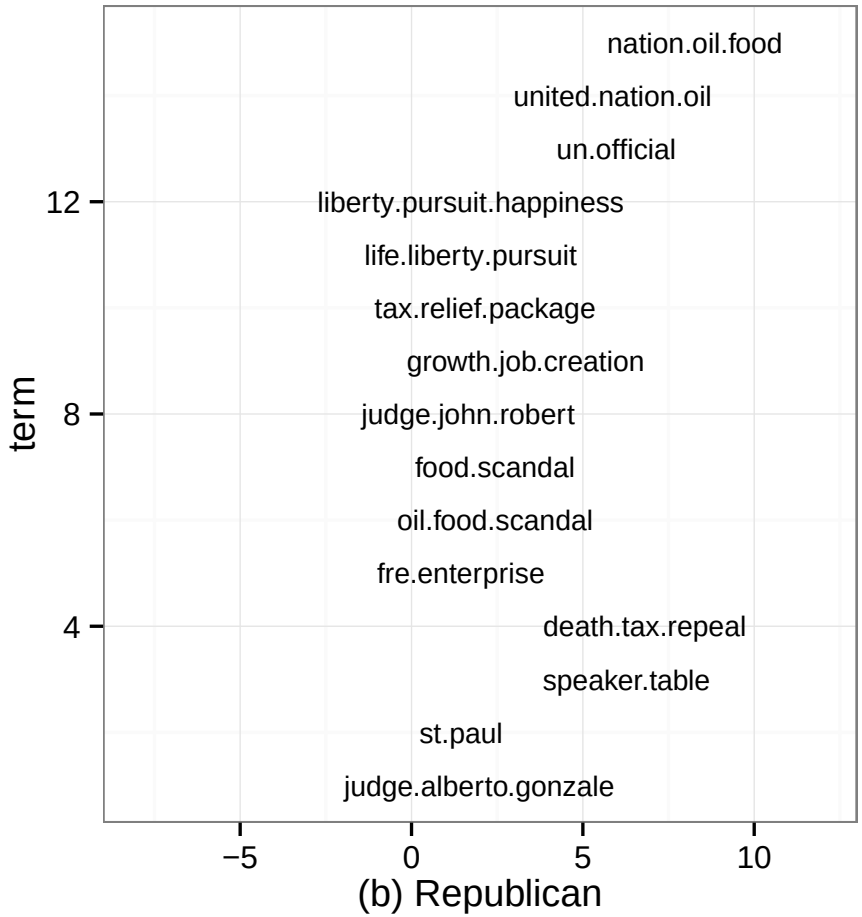
Cluster Membership	Topic Model (LDA)*
umbilic.cord.blood	pluripotent.stem.cel
cord.blood.stem	national.ad.campaign
blood.stem.cel	cel.stem.cel
adult.stem.cel	stem.cel.line

*Results reported in Taddy (2012)

Incorporating metadata: Congressional Speech



(a) Democrat

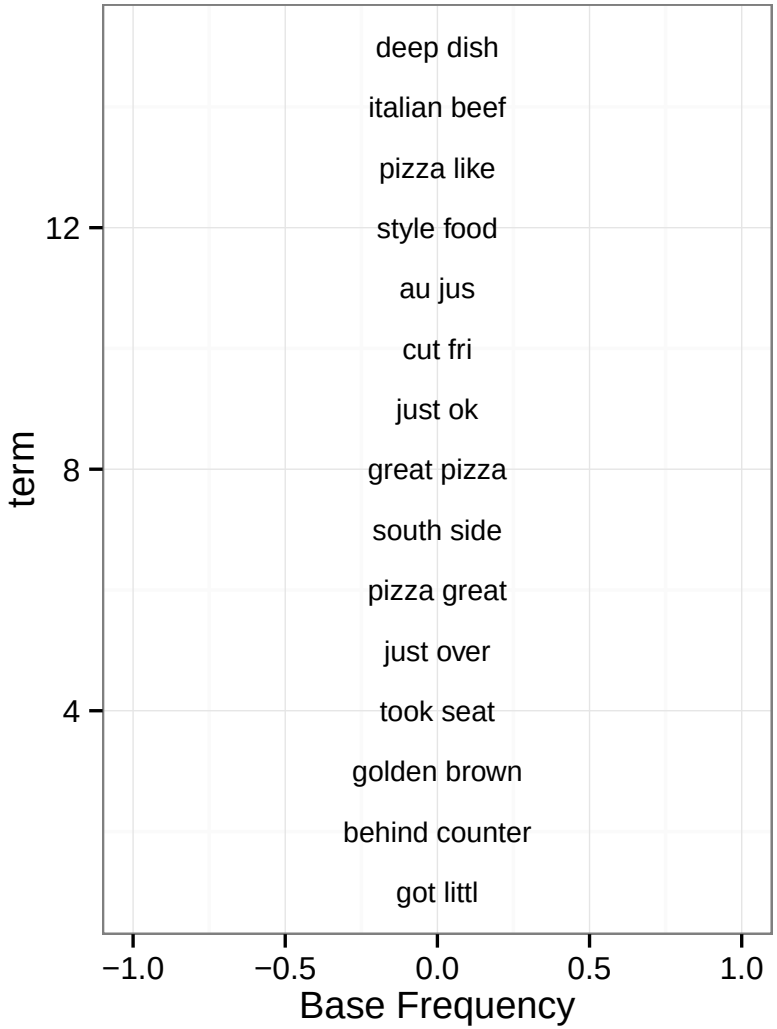


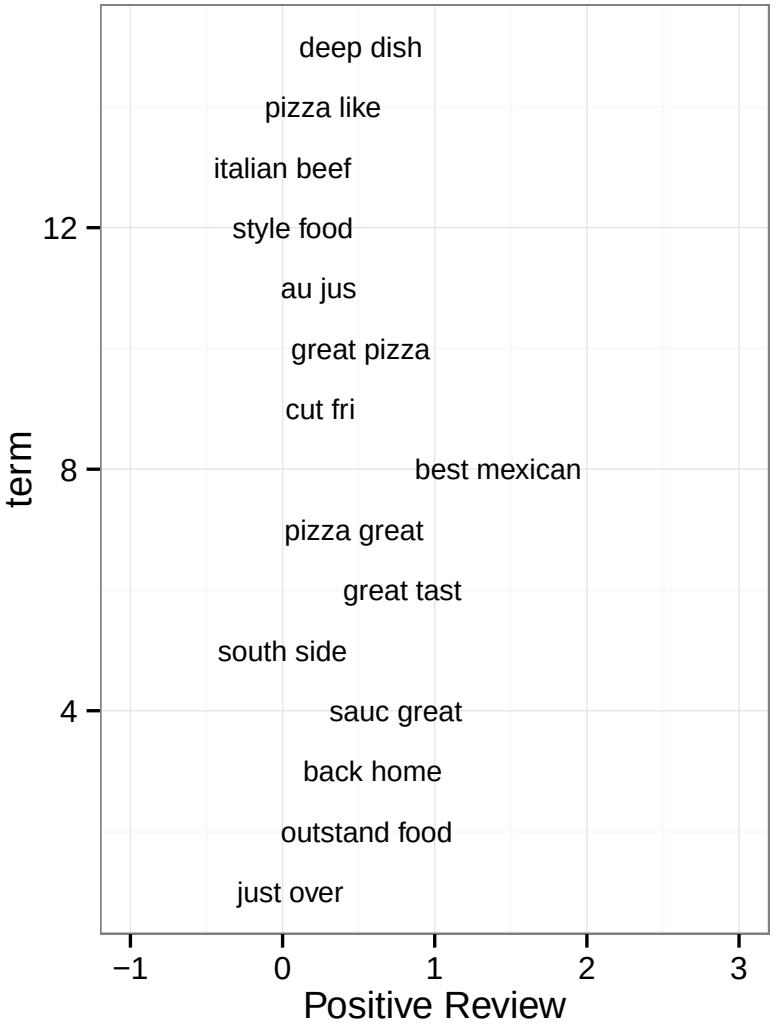
(b) Republican

Example Topic from Restaurant Review

	term	loading
1	deep dish	7.76
2	italian beef	7.07
3	pizza like	6.85
4	style food	6.69
5	au jus	6.33
6	cut fri	6.16
7	just ok	6.01
8	great pizza	5.96
9	south side	5.94
10	pizza great	5.82
11	just over	5.75
12	took seat	5.72
13	golden brown	5.61
14	behind counter	5.58
15	got littl	5.52

Incorporating metadata: Restaurant Review





1. Relationship Between Clusters and Metadata
2. Feature Allocations: Allow an obervation to be a member of multiple clusters
3. Prediction and Cross Validation

Imma Let you Finish, but the Dirichlet was the greatest prior of all time!

Results

	term	loading
1	yeezus	5.48
2	constel	3.79
3	homm	3.79
4	preach	3.79
5	bound	3.6
6	thoma	3.38
7	thirti	3.32
8	rocka	3.31
9	rowland	3.25
10	jamaican	3.23
11	blocka	3.22
12	movement	3.22
13	unlik	3.08
14	yknow	3.08

Extending with OSDC

Million Song Dataset

- A freely-available collection of audio features and metadata for a million contemporary popular music tracks
- Variables:
  - year
  - danceability..(?)
- Lyrics
  - Available from related datasets also on OSDC – linking them?