Combining latent topics with document attributes in text analysis

Nelson Auner
Advisors: Prof. Matt Taddy & Prof. Stephen Stigler

University of Chicago

May 13, 2014

Outline

1. Text as Data
   - Multinomial Models
   - Metadata and Computation
   - Topic Models
2. Cluster Model
   - Algorithm
   - Cluster Initialization
3. Application
   - Congressional Speech Data
   - Restaurant Review Data
4. Extensions

Text as Data

Text as Data

- A document is a collection of phrases.

Text as Data

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

Text as Data

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

Table: What did homework consist of?

Text as Data

- A document is a collection of 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

Multinomial Models

- If order doesn't matter, then we can treat each document as a "bag of words".

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

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

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

A better model: Metadata

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

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?

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

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.

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

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

Topic Models

A topic is a distribution of words.

In a topic model, documents are made of a mixtures of topics.

Running Topic	Bike Topic	Swimming
Stride, Pacing, Stretch	Pedal, Helmet, Gears	Stroke, Air, Water

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

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

Outline

1. Text as Data
   - Multinomial Models
   - Metadata and Computation
   - Topic Models
2. Cluster Model
   - Algorithm
   - Cluster Initialization
3. Application
   - Congressional Speech Data
   - Restaurant Review Data
4. Extensions

Cluster Model

Goal

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

Cluster Model

Goal

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

Approach

- Restrict each document to only one topic ⇒ "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 u_i for $i = 1, \ldots, n$

Algorithm for Cluster Membership Model with Gamma Lasso Penalty

1. Initialize 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)

Algorithm for Cluster Membership Model with Gamma Lasso Penalty

1. Initialize u_i for $i = 1, \ldots, n$
2. Determine parameters $\alpha_, \phi, \Gamma$ 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 $argmax_{k=1,..,K}\left[\ell(u_i|\alpha, \phi, \Gamma)\right]$

Algorithm for Cluster Membership Model with Gamma Lasso Penalty

1. Initialize u_i for $i = 1, \ldots, n$
2. Determine parameters $\alpha_, \phi, \Gamma$ 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.

How do we initialize the clusters?

We test three different approaches:

How do we initialize the clusters?

We test three different approaches:

1. Randomly assign each observation to a cluster

How do we initialize the clusters?

We test three different approaches:

1. Randomly assign each observation to a cluster
2. Group documents by k-means, then assign clusters

How do we initialize the clusters?

We test three different approaches:

1. Randomly assign each observation to a cluster
2. Group documents by k-means, then assign clusters
3. Regress metadata on text, then group residual's by k-means to clusters

Outline

1. Text as Data
   - Multinomial Models
   - Metadata and Computation
   - Topic Models
2. Cluster Model
   - Algorithm
   - Cluster Initialization
3. Application
   - Congressional Speech Data
   - Restaurant Review Data
4. Extensions

Congressional Speech and Restaurant Reviews

- We apply the algorithm to two datasets:
  - Congressional Speech records (Moskowitz and Shapiro, 2010)
  - A corpus of restaurant reviews called we8there.

Congressional Speech and Restaurant Reviews

- We apply the algorithm to two datasets:
  - Congressional Speech records (Moskowitz 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

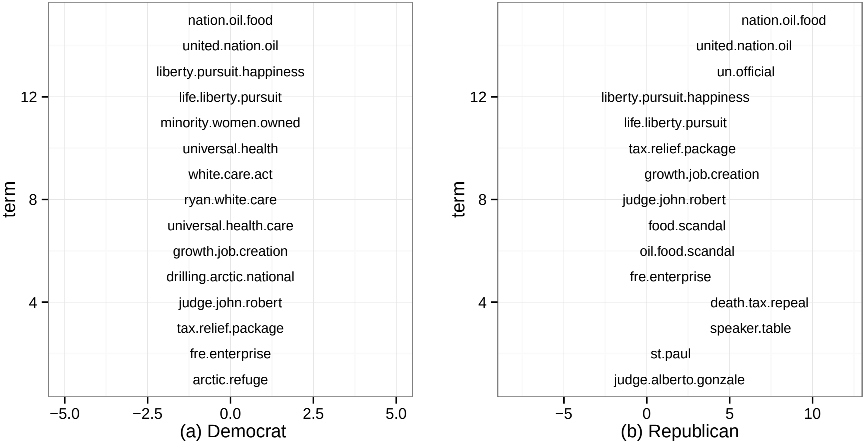
Good news: We are able to recover similar topics with our model:

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)

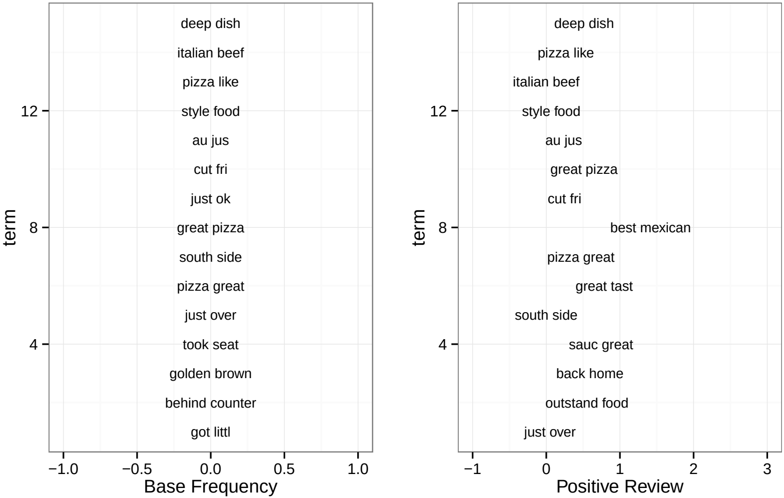
Incorporating metadata: Congressional Speech



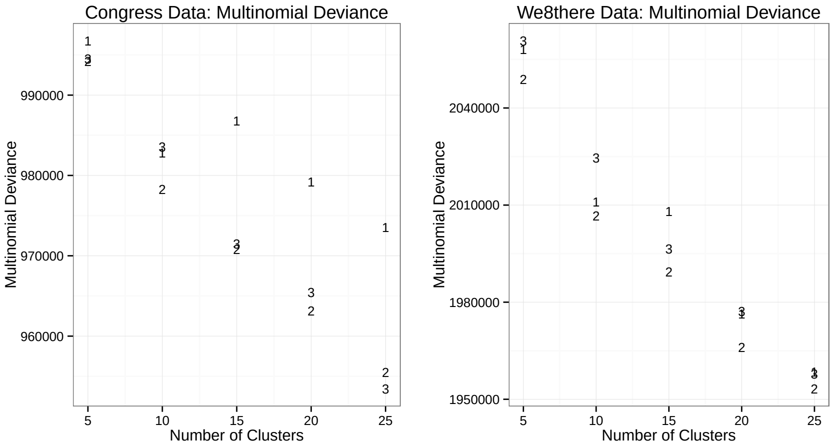
Example Topic from Restaurant Review

	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



Evaluating Cluster Initialization



Outline

1. Text as Data
   - Multinomial Models
   - Metadata and Computation
   - Topic Models
2. Cluster Model
   - Algorithm
   - Cluster Initialization
3. Application
   - Congressional Speech Data
   - Restaurant Review Data
4. Extensions

1. Relationship Between Clusters and Metadata

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

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!

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

Results

Results

	term	loading
1	yeezus	0.36
2	challeng	0.3
3	dreamer	0.3
4	grass	0.3
5	loud	0.3
6	opportun	0.3
7	reduc	0.3
8	speak	0.3
9	frustrat	0.3
10	origin	0.3
11	perfect	0.29
12	rest	0.27
13	bound	0.23
14	seat	0.22

Thank You

Thank You