

TEXT AS DATA: WHAT YOU NEED TO KNOW

Nelson Auner

Prepared for TGG

October 16, 2014

A quick aside...



A quick aside...

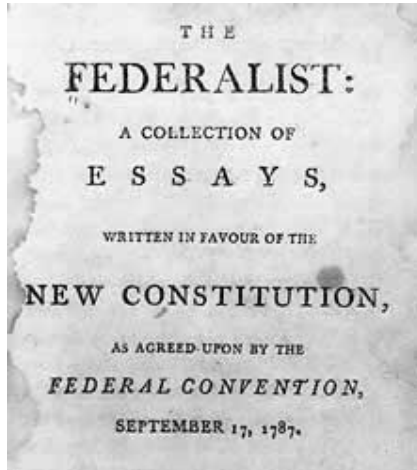


Outline

- 1 Motivation
- 2 Goals
- 3 Text as Data
 - Overview
 - Parsing
 - Multinomial Models
 - Topic Models
- 4 Cluster Model
 - Algorithm
- 5 Application
 - Congressional Speech Data

Motivation: Historical

Motivation: Historical



Motivation: In the News

Motivation: In the News

BloombergBusinessweek Technology

Global
Economies

Companies &
Industries

Politics & Policy

Technology

Markets &
Finance

Innovation &
Design

Lifestyle

Focus On Big Data

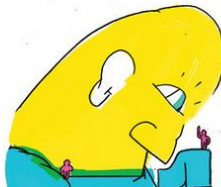
What's the Dollar Value of Online Patient Chatter?

By Caroline Chen | October 09, 2014



SEND TO kindle

On the Internet, people talk about everything, even their illnesses. Treato, an Israeli data-mining company, monitors conversations on Facebook (FB), Twitter (TWTR), and patient forums for information on drug side effects and prescription patterns. It then sends weekly or monthly analyses of the chatter to hedge funds and money managers that invest in pharma stocks.



How do they do it?

How do they do it?

- "Treato distills the collective patient voice from blogs and forums using Natural Language Processing, Big Data and a proprietary patient language..."

Public Sector Use

Public Sector Use



Current Providers

What Will You Receive in Collection
(Surveillance and Stored Comms)?

It varies by provider. In general:

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN

Outline

- 1 Motivation
- 2 Goals
- 3 Text as Data
 - Overview
 - Parsing
 - Multinomial Models
 - Topic Models
- 4 Cluster Model
 - Algorithm
- 5 Application
 - Congressional Speech Data

Are you comfortable talking about text?

Are you comfortable talking about text?

Can you...

Are you comfortable talking about text?

Can you...

- explain the basics of text analysis to a potential client?

Are you comfortable talking about text?

Can you...

- explain the basics of text analysis to a potential client?
- identify opportunity to utilize text analysis?

Are you comfortable talking about text?

Can you...

- explain the basics of text analysis to a potential client?
- identify opportunity to utilize text analysis?
- communicate why text analysis is difficult?

Outline

- 1 Motivation
- 2 Goals
- 3 Text as Data**
 - Overview
 - Parsing
 - Multinomial Models
 - Topic Models
- 4 Cluster Model
 - Algorithm
- 5 Application
 - Congressional Speech Data

The basics

The basics

- A document is a collection of words or phrases.

The basics

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

The basics

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

Table: What did homework consist of?

The basics

- 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

Parsing

Parsing

- Greatest, Greatly, and Greatliest....

Parsing

- Greatest, Greatly, and Greatliest....

It ain't that easy...

Parsing

- Greatest, Greatly, and Greatliest....

It ain't that easy...

Crystial rosey yeah I poe that
We connected with Cali we back door that

Parsing

- Greatest, Greatly, and Greatliest....

It ain't that easy...

Crystial rosey yeah I poe that
We connected with Cali we back door that
You see my wrist man keep your pink wrist bands
She can't believe I'm in a chevy even though I'm rich man

Parsing

- Greatest, Greatly, and Greatliest....

It ain't that easy...

Crystial rosey yeah I poe that
We connected with Cali we back door that
You see my wrist man keep your pink wrist bands
She can't believe I'm in a chevy even though I'm rich man
Chevy Ridin' High - Dre (of Cool and Dre) f/ Rick Ross

Mo'(Multinomial) Models

Mo'(Multinomial) Models

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

Mo'(Multinomial) Models

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

Mo'(Multinomial) Models

- If word 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

Topic Models

A topic is a distribution of words.

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

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text

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

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text

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

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text Analysis

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

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text Analysis

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 triathlon training $\sim \theta_1$ Running + θ_2 Biking + θ_3 Swimming

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text Analysis

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 triathlon training $\sim \theta_1$ Running + θ_2 Biking + θ_3 Swimming
- Issue: We can no longer collapse observations, must use all n observations

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text Analysis

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 triathlon training $\sim \theta_1$ Running + θ_2 Biking + θ_3 Swimming
- Issue: We can no longer collapse observations, must use all n observations
- Workarounds: See Ryan's paper¹ or mine²

¹Wang, 2012. Sparse Coding and an Application to Topic Modeling.

²Auner, 2014. Combining Latent Topics with Document Attributes in Text Analysis

Outline

- 1 Motivation
- 2 Goals
- 3 Text as Data
 - Overview
 - Parsing
 - Multinomial Models
 - Topic Models
- 4 Cluster Model
 - Algorithm
- 5 Application
 - Congressional Speech Data

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 \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 $i = 1, \dots, n$

Algorithm for Cluster Membership Model with Gamma Lasso Penalty

- ① Initialize cluster membership u_i for $i = 1, \dots, n$
- 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 cluster membership u_i for $i = 1, \dots, n$
- 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 $\operatorname{argmax}_{k=1, \dots, K} [\ell(u_i | \alpha, \phi, \Gamma)]$

Algorithm for Cluster Membership Model with Gamma Lasso Penalty

- 1 Initialize cluster membership u_i for $i = 1, \dots, n$
- 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 $\operatorname{argmax}_{k=1,\dots,K} [\ell(u_i | \alpha, \phi, \Gamma)]$
- 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.

Outline

- 1 Motivation
- 2 Goals
- 3 Text as Data
 - Overview
 - Parsing
 - Multinomial Models
 - Topic Models
- 4 Cluster Model
 - Algorithm
- 5 Application
 - Congressional Speech Data

Comparison with the Topic Model

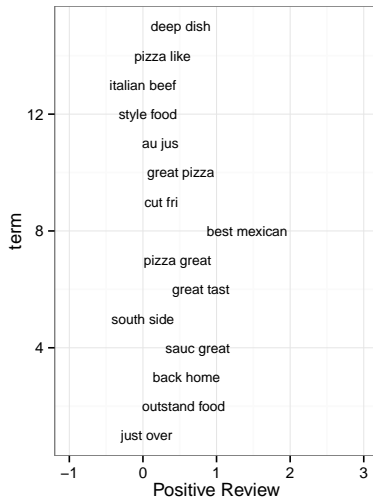
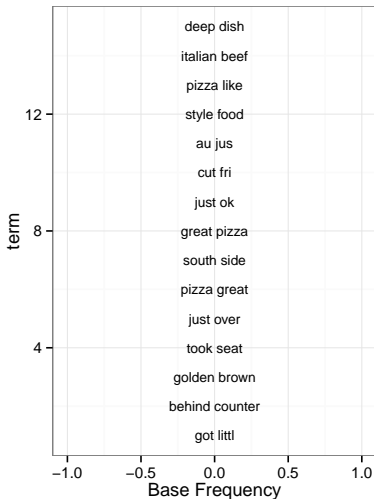
Good news: We are able to recover similar topics with our model:

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: Restaurant Review



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!

The screenshot shows the Genius website interface. At the top, there's a search bar with the text "Search: rapper, song title, or lyrics" and a magnifying glass icon. Below the search bar are navigation tabs: "ADD NEW SONG", "FORUMS", "VERIFIED ARTISTS", and "RAP STATS". The main content area displays the title "Kanye West – Stronger Lyrics" in large blue text. Below the title, it says "Produced By: Kanye West, Mike Dean & Timbaland". Further down, it indicates "Track 3 on Graduation". There are statistics showing "216,684 views", "2 viewing", "31 annotations", and a "Locked" status. Below these are social media sharing buttons for "PYONG" (25), "Like" (150), and "Tweet" (29), along with "Embed" and "Follow" buttons. A link "How do I create annotations?" is also present. The lyrics section begins with "[Produced by Kanye West, Mike Dean, and Timbaland]" and "[Hook]". The visible lyrics are: "N-now th-that that don't kill me", "Can only make me stronger", "I need you to hurry up now", "Cause I can't wait much longer", and "I know I got to be right now".

Results

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

Dip your feet in

Dip your feet in

- *Textir* or *Gamlr* package by Matt Taddy
- Currently only for R
- Python coming soon!

Thank You

Thank You

