Text as Data: What you need to know

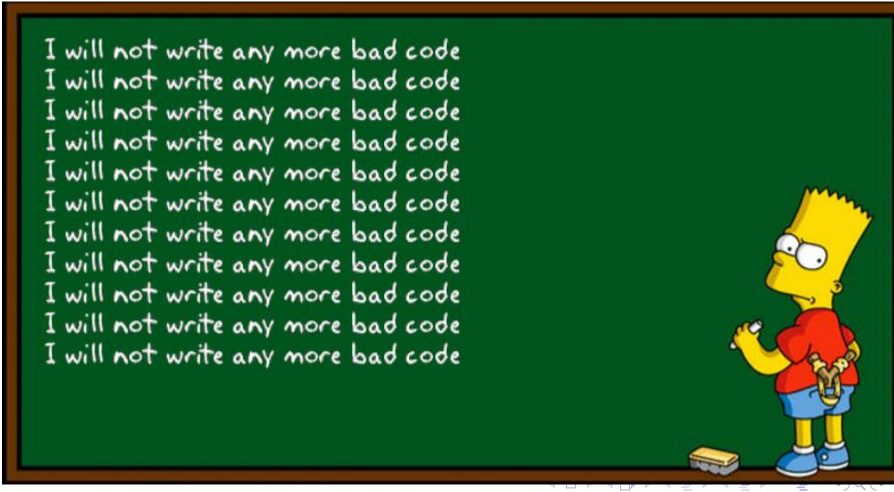
Nelson Auner

Prepared for TGG

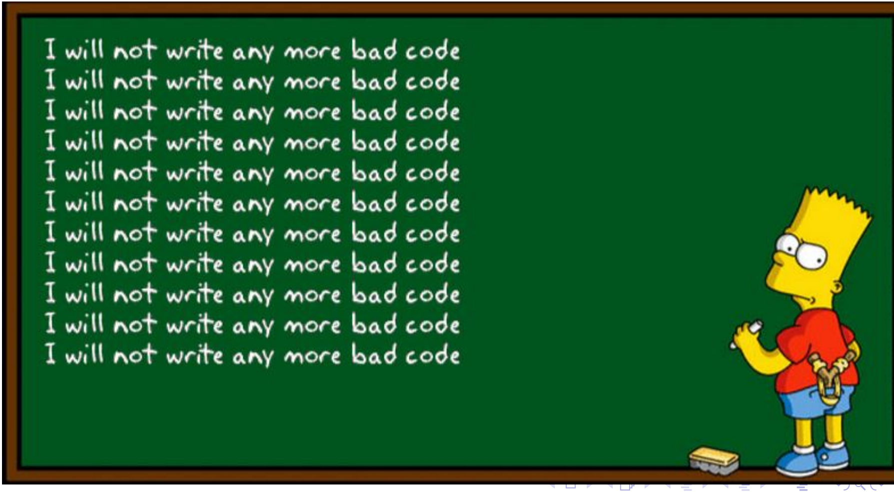
October 16, 2014

A quick aside...

A quick aside...



A quick aside...



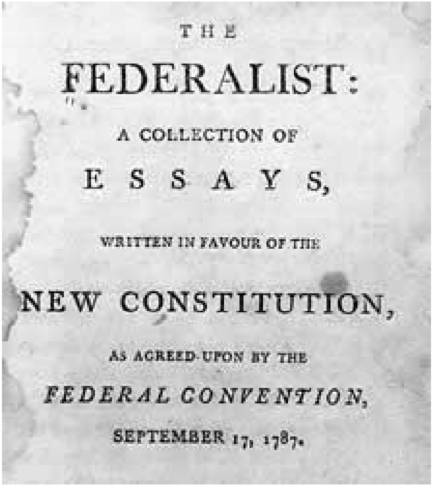
Outline

Drowning in text

Drowning in text

Motivation: Historical

Motivation: Historical


THE
FEDERALIST:
A COLLECTION OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE
FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

Motivation: In the News

Motivation: In the News

BloombergBusinessweek
Technology

Global Economics | Companies & Industries | Politics & Policy | Technology | Markets & Finance | Innovation & Design | Lifestyle

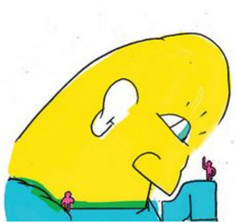
Focus On Big Data

What's the Dollar Value of Online Patient Chatter?

By Caroline Chen | October 09, 2014

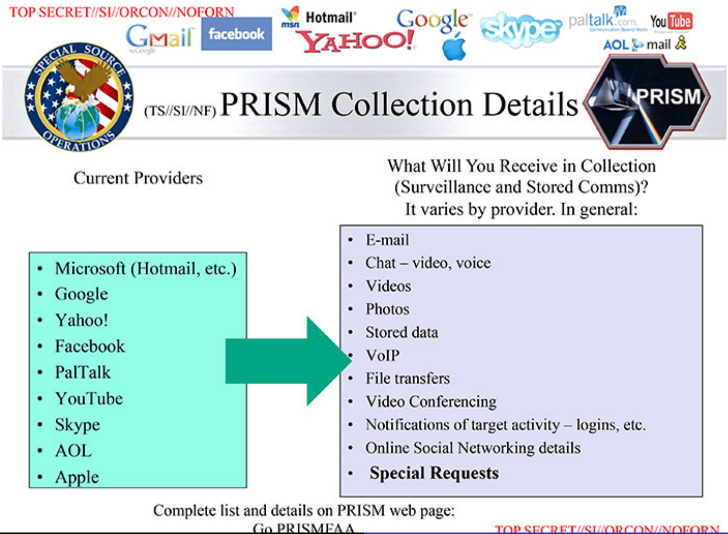
SEND TO kindle

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.



Public Sector Use

Public Sector Use



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..."

Outline

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 opportunities to utilize text analysis?

Are you comfortable talking about text?

Can you...

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

Outline

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.

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

[2]Auner, 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

[1]Wang, 2012. Sparse Coding and an Application to Topic Modeling.
[2]Auner, 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

[1] Wang, 2012. Sparse Coding and an Application to Topic Modeling.
[2] Auner, 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

[1] Wang, 2012. Sparse Coding and an Application to Topic Modeling.
[2] Auner, 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 triathalon training $\sim \theta_1$ Running $+ \theta_2$ Biking $+ \theta_3$ Swimming

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

[2] Auner, 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 triathalon training $\sim \theta_1$ Running $+ \, \theta_2$ Biking $+ \, \theta_3$ Swimming
- Issue: We can no longer collapse observations, must use all n observations

[1] Wang, 2012. Sparse Coding and an Application to Topic Modeling.
[2] Auner, 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

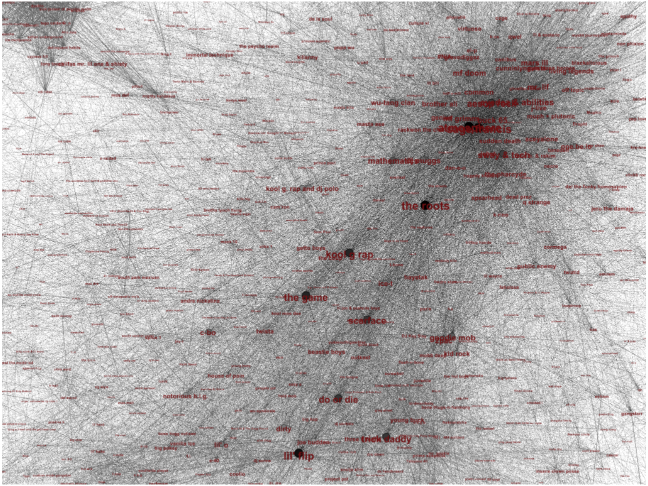
- A book about triathalon training $\sim \theta_1$ Running $+ \theta_2$ Biking $+ \theta_3$ Swimming
- Issue: We can no longer collapse observations, must use all n observations
- Workarounds: See Ryan's paper[1] or mine [2]

[1] Wang, 2012. Sparse Coding and an Application to Topic Modeling.
[2] Auner, 2014. Combining Latent Topics with Document Attributes in Text Analysis

Outline

Uncovering relationships

Uncovering relationships



Uncovering topics

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

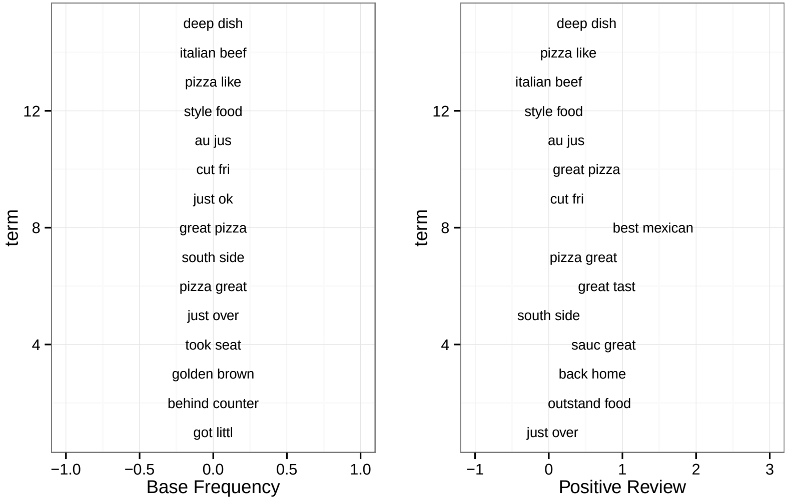
Uncovering topics

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)

Topics on Yelp



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	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

Dip your feet in

Dip your feet in

- *Textir* or *Gamlr*
- Currently only for R
- Python coming soon!

Thank You

- Math: $x_i \sim MN(q_{ij}, m_{ij})$; $\quad q_{ij} = \frac{exp(\alpha_j + y_i\phi_j + u_i\Gamma_{kj})}{\sum_{l=1}^{p} exp(\alpha_l + y_i\phi_l + u_i\Gamma_{kl})}$
- Talk based on
  Combining Latent Topics with Document Attributes in Text Analysis
- Advisors: Prof. Matt Taddy[3], Prof. Stephen Stigler[4]

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

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