Text as Data

Justin Grimmer

Associate Professor
Department of Political Science
Stanford University

September 23rd, 2014

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
    - Hard to find
    - Time Consuming
    - Not generalizable (each new data set...new coding scheme)

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
    - Hard to find
    - Time Consuming
    - Not generalizable (each new data set...new coding scheme)
    - Difficult to store/search

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)
  - Difficult to store/search
  - Idiosyncratic to coders/researcher

Text and Political Science

A pre-2000's view of text in social science

- Social interaction often occurs in texts
- Social Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)
  - Difficult to store/search
  - Idiosyncratic to coders/researcher
  - Statistical methods/algorithms, computationally intensive

A post-2000's view of text in social science:

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...
- Facebook posts, tweets, emails, cell phone records, ...

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...
- Facebook posts, tweets, emails, cell phone records, ...
- Newspapers, magazines, news broadcasts, ...

A post-2000's view of text in social science:

Massive collections of texts are increasingly used as a data source in social science:

- Congressional speeches, press releases, newsletters, ...
- Facebook posts, tweets, emails, cell phone records, ...
- Newspapers, magazines, news broadcasts, ...
- Foreign news sources, treaties, sermons, fatwas, ...

Why?

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $=$ 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $= 1$ LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: $10,000 megabyte. 2014: $<<<<<$ $0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
    - Generalizable: one method can be used across many methods and to unify collections of texts
    - Systematic: parameters/statistics demonstrate how models make coding decisions
    - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
    - Laws

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email $=$ 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: $10,000 megabyte. 2014: $<<<<<$ $0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: $10,000 megabyte. 2014: $<<<<<$ $0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions
  - Press Releases

Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: \$10,000 megabyte. 2014: $<<<<<$ \$0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic: parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts, computing power is inexpensive
- Unchanged Demand: Social life (politics, economic exchanges, social interactions) occurs in texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions
  - Press Releases

What Can Text Methods Do?

Haystack metaphor:

What Can Text Methods Do?

Haystack metaphor: Improve Reading

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\rightsquigarrow$ Organizing hay stack

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing $\rightsquigarrow$ largely what we'll discuss today

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\leadsto$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\leadsto$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing $\leadsto$ largely what we'll discuss today

What automated text methods don't do:

What Can Text Methods Do?

Haystack metaphor: Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts $\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing $\rightsquigarrow$ largely what we'll discuss today

What automated text methods don't do:

- Develop a comprehensive statistical model of language
- Replace the need to read
- Develop a single tool + evaluation for all tasks

Texts are Deceptively Complex

We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?
- Who is the `We`?

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?
- Who is the `We`?
- What is the `mountaintop` (literal?)

Texts are Deceptively Complex

`We've got some difficult days ahead. But it doesn't matter with me now. Because I've been to the mountaintop. And I don't mind. Like anybody, I would like to live a long life. Longevity has its place. But I'm not concerned about that now.`

- Who is the `I` ?
- Who is the `We`?
- What is the `mountaintop` (literal?)

Texts $\rightsquigarrow$ high dimensional, not self contained

Texts are Surprisingly Simple

(Lamar Alexander (R-TN) Feb 10, 2005)

Word	No. Times Used in Press Release
department	12
grant	9
program	7
firefight	7
secure	5
homeland	4
fund	3
award	2
safety	2
service	2
AFGP	2
support	2
equip	2
applaud	2
assist	2

Texts are Surprisingly Simple (?)

`US Senators Bill Frist (R-TN) and Lamar Alexander (R-TN) today applauded the U S Department of Homeland Security for awarding a $8,190 grant to the Tracy City Volunteer Fire Department under the 2004 Assistance to Firefighters Grant Program's (AFGP) FirePrevention and Safety Program...`

Not just for "big data"

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) $=$ number of ways of partitioning n objects
- Bell(2) $= 2$ (AB, A B)
- Bell(3) $= 5$ (ABC, AB C, A BC, AC B, A B C)

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100) $\approx 4.75 \times 10^{115}$ partitions

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100) $\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100) $\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times$

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) $=$ number of ways of partitioning n objects
- Bell(2) $= 2$ (AB, A B)
- Bell(3) $= 5$ (ABC, AB C, A BC, AC B, A B C)
- Bell(5) $= 52$
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$ years

Not just for "big data"

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$ years

Automated methods can help with even small problems

What We'll Do:

Statistical and Computational tools for working with texts

Prerequisites

Statistics:

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with `R` programming language

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with `R` programming language
- Experience with:

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with R programming language
- Experience with:
    - Programming functions

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with `R` programming language
- Experience with:
  - Programming functions
  - Writing for loops

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with R programming language
- Experience with:
  - Programming functions
  - Writing for loops
  - Using standard R packages

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with R programming language
- Experience with:
  - Programming functions
  - Writing for loops
  - Using standard R packages
  - Creating plots

Prerequisites

Statistics:

- Probability Theory/Univariate Inference (Old 350a)
- Linear Regression (Old 350b)
- (Ideally) Model Based Inference (Old 350c)
- Willingness to learn new statistical models(!!)

Computational:

- Familiarity with `R` programming language
- Experience with:
  - Programming functions
  - Writing for loops
  - Using standard `R` packages
  - Creating plots
- Willingness to learn `Python`

Course Staff

Me: Justin Grimmer

Office: Encina West 414 (last door on left)

Office Hours: I'm usually in during business hours. Set up an appointment if you must meet with me

Contact: Gchat: justin.grimmer@gmail.com; Cell phone (617) 710-6803

Course Staff

Me: Justin Grimmer

Office: Encina West 414 (last door on left)

Office Hours: I'm usually in during business hours. Set up an appointment if you must meet with me

Contact: Gchat: justin.grimmer@gmail.com; Cell phone (617) 710-6803

Programming TA

Python/R/Programming: Frances Zlotnick

Office/Programming Section: Encina Hall West, Room 417

Office Hours: 230-430 and by appointment

Contact: Zlotnick@stanford.edu

Evaluation

Homework:

- Weekly homework assignments
- Computational Component
  - Preprocessing texts
  - Moving from texts $\rightsquigarrow$ data
- Statistical component
  - Applying algorithms, statistics to analyze texts

Our workspace

1) `RStudio` $\rightsquigarrow$ lowers startup costs of `R`
2) `R Markdown` $\rightsquigarrow$ integrates write up and code
3) `Enthought Python Distribution` (academic license) $\rightsquigarrow$ python distribution that ships with most packages

Writeup can also occur in LaTeX

Evaluation

Homework:

1) Will be distributed on Tuesday
2) Due on Tuesday, 5pm
3) `Email:` Frannie and me

Collaborate!

1) Work together in groups
2) Individual write ups

Evaluation

Final Project:

Evaluation

Final Project:

1) An original research paper

Evaluation

Final Project:

1) An original research paper
   - Part of a dissertation
   - Field paper
   - Paper for publication

Evaluation

Final Project:

1) An original research paper
   - Part of a dissertation
   - Field paper
   - Paper for publication
2) Contributing to ongoing research project

Evaluation

Final Project:

1) An original research paper
   - Part of a dissertation
   - Field paper
   - Paper for publication
2) Contributing to ongoing research project
   1) Michael Crespin (U of Oklahoma, Congressional Scholar): Categorizing floor speeches $\rightsquigarrow$ citations

Evaluation

Final Project:

1) An original research paper
   - Part of a dissertation
   - Field paper
   - Paper for publication
2) Contributing to ongoing research project
   1) Michael Crespin (U of Oklahoma, Congressional Scholar): Categorizing floor speeches $\rightsquigarrow$ citations
   2) Alison McQueen (Stanford): Characterizing Hobbes' context $\rightsquigarrow$ political theory

Evaluation

Final Project:

1) An original research paper
   - Part of a dissertation
   - Field paper
   - Paper for publication
2) Contributing to ongoing research project
   1) Michael Crespin (U of Oklahoma, Congressional Scholar): Categorizing floor speeches $\rightsquigarrow$ citations
   2) Alison McQueen (Stanford): Characterizing Hobbes' context $\rightsquigarrow$ political theory
   3) Robert Gulotty (Stanford $\rightsquigarrow$ U of Chicago) and Judith Goldstein (Stanford) Examine trade speeches in the 19th century Congress

Evaluation

Final Project:

1) An original research paper
   - Part of a dissertation
   - Field paper
   - Paper for publication
2) Contributing to ongoing research project
   1) Michael Crespin (U of Oklahoma, Congressional Scholar): Categorizing floor speeches $\rightsquigarrow$ citations
   2) Alison McQueen (Stanford): Characterizing Hobbes' context $\rightsquigarrow$ political theory
   3) Robert Gulotty (Stanford $\rightsquigarrow$ U of Chicago) and Judith Goldstein (Stanford) Examine trade speeches in the 19th century Congress

Talk to me about your ideas!

Evaluation

Final Project:

Evaluation

Final Project:

1) Poster Session

Evaluation

Final Project:

1) Poster Session
   - Opportunity to receive feedback on your projects

Evaluation

Final Project:

1) Poster Session
    - Opportunity to receive feedback on your projects
2) Final paper

Evaluation

Final Project:

1) Poster Session
   - Opportunity to receive feedback on your projects
2) Final paper
   - Research length (25-30 pages)

Evaluation

Final Project:

1) Poster Session
   - Opportunity to receive feedback on your projects
2) Final paper
   - Research length (25-30 pages)
   - Format appropriate for your field

Evaluation

Final Project:

1) Poster Session
   - Opportunity to receive feedback on your projects
2) Final paper
   - Research length (25-30 pages)
   - Format appropriate for your field

- Collaborative$\rightsquigarrow$ work in two-person teams

Evaluation

Final Project:

1) Poster Session
   - Opportunity to receive feedback on your projects
2) Final paper
   - Research length (25-30 pages)
   - Format appropriate for your field

- Collaborative$\rightsquigarrow$ work in two-person teams
- We will not adjudicate disputes (frankly, unimportant)

Evaluation

Participation:

Evaluation

Participation:

- Attend class

Evaluation

Participation:

- Attend class
- Ask questions (!!!)

Evaluation

Participation:

- Attend class
- Ask questions (!!!)
- Enroll in Piazza course site

Evaluation

Participation:

- Attend class
- Ask questions (!!!)
- Enroll in Piazza course site
  - piazza.com/stanford/fall2014/polsci452

Evaluation

Participation:

- Attend class
- Ask questions (!!!)
- Enroll in Piazza course site
  - piazza.com/stanford/fall2014/polsci452
  - I'll post lecture slides there and readings (ensures auditors/guests have access)

Evaluation

Participation:

- Attend class
- Ask questions (!!!)
- Enroll in Piazza course site
  - piazza.com/stanford/fall2014/polsci452
  - I'll post lecture slides there and readings (ensures auditors/guests have access)
  - Post Questions/Answer Questions/Course Announcements

Plan for the Course

Computational and Statistical tools

Plan for the Course

Computational and Statistical tools

- Acquiring and Preprocessing Text data

Plan for the Course

Computational and Statistical tools

- Acquiring and Preprocessing Text data
  - Basics of webscraping
  - Regular expressions
  - Text $\rightsquigarrow$ Document Term Matrices

Plan for the Course

Computational and Statistical tools

- Acquiring and Preprocessing Text data
  - Basics of webscraping
  - Regular expressions
  - Text $\rightsquigarrow$ Document Term Matrices
- Dictionary Methods

Plan for the Course

Computational and Statistical tools

- Acquiring and Preprocessing Text data
  - Basics of webscraping
  - Regular expressions
  - Text $\rightsquigarrow$ Document Term Matrices
- Dictionary Methods
  - Assume$\rightsquigarrow$ known categories
  - Assume$\rightsquigarrow$ known how words relate to groups
  - Measure prevalence of categories

Plan for the Course

Computational and Statistical tools

- Acquiring and Preprocessing Text data
  - Basics of webscraping
  - Regular expressions
  - Text $\rightsquigarrow$ Document Term Matrices
- Dictionary Methods
  - Assume$\rightsquigarrow$ known categories
  - Assume$\rightsquigarrow$ known how words relate to groups
  - Measure prevalence of categories
- Discriminating Words

Plan for the Course

Computational and Statistical tools

- Acquiring and Preprocessing Text data
  - Basics of webscraping
  - Regular expressions
  - Text $\rightsquigarrow$ Document Term Matrices
- Dictionary Methods
  - Assume$\rightsquigarrow$ known categories
  - Assume$\rightsquigarrow$ known how words relate to groups
  - Measure prevalence of categories
- Discriminating Words
  - Assume$\rightsquigarrow$ known categories
  - Statistical methods/algorithms to measure word discrimination

Plan for the Course

- Geometry of Texts

Plan for the Course

- Geometry of Texts
  - Assume$\rightsquigarrow$ relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension

Plan for the Course

- Geometry of Texts
  - Assume$\rightsquigarrow$ relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)

Plan for the Course

- Geometry of Texts
  - Assume $\rightsquigarrow$ relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
  - Fully Automated Clustering Methods

Plan for the Course

- Geometry of Texts
  - Assume$\rightsquigarrow$ relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
  - Fully Automated Clustering Methods
    - Assume$\rightsquigarrow$ Known distance
    - Assume$\rightsquigarrow$ Known objective
    - Assume$\rightsquigarrow$ Known method for optimization
    - Statistical model to partition documents

Plan for the Course

- Geometry of Texts
  - Assume$\rightsquigarrow$ relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
  - Fully Automated Clustering Methods
    - Assume$\rightsquigarrow$ Known distance
    - Assume$\rightsquigarrow$ Known objective
    - Assume$\rightsquigarrow$ Known method for optimization
    - Statistical model to partition documents
  - Computer Assisted Clustering

Plan for the Course

- Geometry of Texts
  - Assume$\leadsto$ relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension
- Clustering Methods (Unknown Groups, Unknown relationship of document characteristics to those groups)
  - Fully Automated Clustering Methods
    - Assume$\leadsto$ Known distance
    - Assume$\leadsto$ Known objective
    - Assume$\leadsto$ Known method for optimization
    - Statistical model to partition documents
  - Computer Assisted Clustering
    - Assume$\leadsto$ Method for organizing clusters
    - Method for generating, organizing partitions for discovery

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume$\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume $\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume$\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models
  - Assume$\rightsquigarrow$ condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume$\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models
  - Assume$\rightsquigarrow$ condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language
- Supervised Methods

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume$\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models
  - Assume$\rightsquigarrow$ condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language
- Supervised Methods
  - Assume$\rightsquigarrow$ Known categories (training documents)
  - Statistical model: learn relationship between labels, words categorize remaining documents
  - Ensembles of methods

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume$\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models
  - Assume$\rightsquigarrow$ condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language
- Supervised Methods
  - Assume$\rightsquigarrow$ Known categories (training documents)
  - Statistical model: learn relationship between labels, words categorize remaining documents
  - Ensembles of methods
- Ideological Scaling

Plan for the Course

- "Vanilla" Latent Dirichlet Allocation (Topic Models)
  - Unknown categories
  - Assume$\rightsquigarrow$ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models
  - Assume$\rightsquigarrow$ condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language
- Supervised Methods
  - Assume$\rightsquigarrow$ Known categories (training documents)
  - Statistical model: learn relationship between labels, words categorize remaining documents
  - Ensembles of methods
- Ideological Scaling
  - Application of methods, measuring political positions
  - Supervised$\rightsquigarrow$ Wordscores
  - Unsupervised $\rightsquigarrow$ Item Response Theory (IRT) Models

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text$\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
    - Time flies like an arrow, fruit flies like a banana

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace
  - "Years from now, you'll look back and you'll say that this was the moment, this was the place where America remembered what it means to hope. "

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace
  - "Years from now, you'll look back and you'll say that this was the moment, this was the place where America remembered what it means to hope. "
- Models necessarily fail to capture language $\rightsquigarrow$ useful for specific tasks

Four Principles of Automated Text Analysis

Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\rightsquigarrow$ unknown
- Complexity of language:
  - Time flies like an arrow, fruit flies like a banana
  - Make peace, not war , Make war not peace
  - "Years from now, you'll look back and you'll say that this was the moment, this was the place where America remembered what it means to hope. "
- Models necessarily fail to capture language $\rightsquigarrow$ useful for specific tasks
- Validation $\rightsquigarrow$ demonstrate methods perform task

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading
- Quantitative methods organize, direct, and suggest

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading
- Quantitative methods organize, direct, and suggest
- Humans: read and interpret

Four Principles of Automated Text Analysis

Principle 3: There is no Globally Best Method for Automated Text Analysis

Four Principles of Automated Text Analysis

Principle 3: There is no Globally Best Method for Automated Text Analysis

- Supervised methods $\rightsquigarrow$ known categories

Four Principles of Automated Text Analysis

Principle 3: There is no Globally Best Method for Automated Text Analysis

- Supervised methods $\rightsquigarrow$ known categories
- Unsupervised methods $\rightsquigarrow$ discover categories

Four Principles of Automated Text Analysis

Principle 3: There is no Globally Best Method for Automated Text Analysis

- Supervised methods $\rightsquigarrow$ known categories
- Unsupervised methods $\rightsquigarrow$ discover categories
- Debate $\rightsquigarrow$ acknowledge differences, resolved

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods $\rightsquigarrow$ variable performance across tasks

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods $\rightsquigarrow$ variable performance across tasks
- Few theorems to guarantee performance

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods$\rightsquigarrow$ variable performance across tasks
- Few theorems to guarantee performance
- Apply methods $\rightsquigarrow$ validate

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods$\rightsquigarrow$ variable performance across tasks
- Few theorems to guarantee performance
- Apply methods $\rightsquigarrow$ validate
- Avoid: blind application of methods

Going Forward

1) Assignment distributed tonight
2) Install `R` and `Python`
3) Thursday: The Statistical/Computational Background for Text as Data!