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
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
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, ...

- Facebook posts, tweets, emails, cell phone records, ...

- Newspapers, magazines, news broadcasts, ...

- Foreign news sources, treaties, sermons, fatwas, ...
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?

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

Justin Grimmer (Stanford University)
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)
- 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
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

Justin Grimmer (Stanford University)
Text as Data
September 23rd, 2014 4 / 25
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:

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

- Comparing, Organizing, and Classifying Texts ⇝ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing

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
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 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
What Can Text Methods Do?

Haystack metaphor: Improve Reading
- Interpreting the meaning of a sentence or phrase ⟷ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts ⟷ Organizing haystack
  - Humans: terrible. Tiny active memories
  - Computers: amazing

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
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
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 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 ⇝ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts ⇝ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing ⇝ 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
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.
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?
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?
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→ high dimensional, not self contained
# Texts are Surprisingly Simple

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

<table>
<thead>
<tr>
<th>Word</th>
<th>No. Times Used in Press Release</th>
</tr>
</thead>
<tbody>
<tr>
<td>department</td>
<td>12</td>
</tr>
<tr>
<td>grant</td>
<td>9</td>
</tr>
<tr>
<td>program</td>
<td>7</td>
</tr>
<tr>
<td>firefight</td>
<td>7</td>
</tr>
<tr>
<td>secure</td>
<td>5</td>
</tr>
<tr>
<td>homeland</td>
<td>4</td>
</tr>
<tr>
<td>fund</td>
<td>3</td>
</tr>
<tr>
<td>award</td>
<td>2</td>
</tr>
<tr>
<td>safety</td>
<td>2</td>
</tr>
<tr>
<td>service</td>
<td>2</td>
</tr>
<tr>
<td>AFGP</td>
<td>2</td>
</tr>
<tr>
<td>support</td>
<td>2</td>
</tr>
<tr>
<td>equip</td>
<td>2</td>
</tr>
<tr>
<td>applaud</td>
<td>2</td>
</tr>
<tr>
<td>assist</td>
<td>2</td>
</tr>
</tbody>
</table>
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

- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100) ≈ 4.75 \times 10^{115}

Big Number: 7 Billion RAs
Impossibly Fast (enumerate one clustering every millisecond)
Working around the clock (24/7/365)
≈ 1.54 \times 10^{84} \times (14,000,000,000,000) years

Automated methods can help with even small problems
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)

Big Number: 7 Billion RAs
Impossibly Fast (enumerate one clustering every millisecond)
Working around the clock (24/7/365)
≈ 1.54 × 10^84 × (14,000,000,000)

Automated methods can help with even small problems
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) ≈ 4.75 × 10^{115} partitions
Not just for “big data”

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

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

Automated methods can help with even small problems
Statistical and Computational tools for working with texts
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
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 \texttt{R} programming language
- Experience with:
  - Programming functions
  - Writing for loops
  - Using standard \texttt{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 to data
- Statistical component
  - Applying algorithms, statistics to analyze texts

Our workspace

1) RStudio \(\leadsto\) lowers startup costs of R
2) R Markdown \(\leadsto\) integrates write up and code
3) Enthought Python Distribution (academic license) \(\leadsto\) 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:

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 ⇝ citations

2) Alison McQueen (Stanford): Characterizing Hobbes' context ⇝ political theory

3) Robert Gulotty (Stanford) and Judith Goldstein (Stanford) Examine trade speeches in the 19th century Congress

Talk to me about your ideas!

Justin Grimmer (Stanford University)
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 ⇝ citations

2) Alison McQueen (Stanford): Characterizing Hobbes' context ⇝ political theory

3) Robert Gulotty (Stanford) and Judith Goldstein (Stanford) Examine trade speeches in the 19th century Congress

Talk to me about your ideas!

Justin Grimmer (Stanford University)
Evaluation

Final Project:

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

2) Contributing to ongoing research projects
   1) Michael Crespin (U of Oklahoma, Congressional Scholar): Categorizing floor speeches ⇝ citations
   2) Alison McQueen (Stanford): Characterizing Hobbes’ context ⇝ political theory
   3) Robert Gulotty (Stanford) and Judith Goldstein (Stanford): Examine trade speeches in the 19th century Congress

Talk to me about your ideas!

Justin Grimmer (Stanford University)
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 ⇝ citations

2) Alison McQueen (Stanford): Characterizing Hobbes' context ⇝ political theory

2) Robert Gulotty (Stanford) and Judith Goldstein (Stanford) Examine trade speeches in the 19th century Congress

Talk to me about your ideas!

Justin Grimmer (Stanford University)
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 ⇝ citations
   2) Alison McQueen (Stanford): Characterizing Hobbes' context ⇝ political theory

Talk to me about your ideas!

Justin Grimmer (Stanford University)
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 $\Rightarrow$ citations
   2) Alison McQueen (Stanford): Characterizing Hobbes' context $\Rightarrow$ political theory
   3) Robert Gulotty (Stanford $\Rightarrow$ 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\(\rightarrow\) citations
   2) Alison McQueen (Stanford): Characterizing Hobbes' context\(\rightarrow\) political theory
   3) Robert Gulotty (Stanford\(\rightarrow\) U of Chicago) and Judith Goldstein (Stanford) Examine trade speeches in the 19th century Congress

Talk to me about your ideas!
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 ⇝ work in two-person teams
- We will not adjudicate disputes (frankly, unimportant)
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\textarrow 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 → work in two-person teams
   - We will not adjudicate disputes (frankly, unimportant)
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
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

- Dictionary Methods
  - Assume $\Rightarrow$ known categories
  - Measure prevalence of categories
  - Discriminating Words
    - Assume $\Rightarrow$ known categories
    - Statistical methods/algorithms to measure word discrimination
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 $\leadsto$ Document Term Matrices

- Dictionary Methods
  - Assume $\leadsto$ known categories
  - Assume $\leadsto$ 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 \(\rightarrow\) Document Term Matrices

- **Dictionary Methods**
  - **Assume** \(\rightarrow\) known categories
  - **Assume** \(\rightarrow\) 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 $\leadsto$ Document Term Matrices

- Dictionary Methods
  - Assume $\leadsto$ known categories
  - Assume $\leadsto$ known how words relate to groups
  - Measure prevalence of categories

- Discriminating Words
  - Assume $\leadsto$ known categories
  - Statistical methods/algorithms to measure word discrimination
Plan for the Course

- Geometry of Texts
Plan for the Course

- **Geometry of Texts**
  - Assume relationship between texts
  - Statistical methods/algorithms to project (scale) texts in lower dimension
Plan for the Course

- Geometry of Texts
  - Assume 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 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 $\Rightarrow$ 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 $\Rightarrow$ Known distance
    - Assume $\Rightarrow$ Known objective
    - Assume $\Rightarrow$ Known method for optimization
    - Statistical model to partition documents
Plan for the Course

- Geometry of Texts
  - Assume 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 Known distance
    - Assume Known objective
    - Assume Known method for optimization
    - Statistical model to partition documents
- Computer Assisted Clustering
Plan for the Course

- Geometry of Texts
  - **Assume** 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** Known distance
    - **Assume** Known objective
    - **Assume** Known method for optimization
    - Statistical model to partition documents
  - Computer Assisted Clustering
    - **Assume** 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 documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics

- Structural Topic Models
  - Assume condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language

- Supervised Methods
  - Assume 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
  - Wordscores

- Unsupervised
  - Item Response Theory (IRT) Models
Plan for the Course

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

- Structural Topic Models
  - Assume 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 documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics
- Structural Topic Models
  - Assume 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 documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics

- Structural Topic Models
  - Assume condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language

- Supervised Methods
  - Assume 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→ documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics

- Structural Topic Models
  - Assume→ condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language

- Supervised Methods
  - Assume→ 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 documents are mixture of topics
  - Statistical method for measuring topics and document attention to topics

- Structural Topic Models
  - Assume condition on characteristics
  - Measure topics, prevalence of topics across characteristics, distinctiveness of language

- Supervised Methods
  - Assume 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 Wordscores
  - Unsupervised 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 $\Rightarrow$ 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 \( \mapsto \) 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
Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text $\leadsto$ 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 $\rightarrow$ 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 $\rightarrow$ 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 \( \rightarrow \) 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 $\Rightarrow$ 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 $\Rightarrow$ useful for specific tasks
Principle 1: All Quantitative Models of Language are Wrong—But Some are Useful

- Data generation process for text \(\rightarrow\) 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 \(\rightarrow\) useful for specific tasks
- Validation \(\rightarrow\) 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
Principle 3: There is no Globally Best Method for Automated Text Analysis
- Supervised methods\(\Rightarrow\) known categories
Four Principles of Automated Text Analysis

Principle 3: There is no Globally Best Method for Automated Text Analysis
- Supervised methods ⇒ known categories
- Unsupervised methods ⇒ discover categories
Four Principles of Automated Text Analysis

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

- Supervised methods $\rightarrow$ known categories
- Unsupervised methods $\rightarrow$ discover categories
- Debate $\rightarrow$ 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 ⇒ variable performance across tasks
Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods \( \Rightarrow \) variable performance across tasks
- Few theorems to guarantee performance
Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods $\Rightarrow$ variable performance across tasks
- Few theorems to guarantee performance
- Apply methods $\Rightarrow$ validate
Principle 4: Validate, Validate, Validate

- Quantitative methods $\leadsto$ variable performance across tasks
- Few theorems to guarantee performance
- Apply methods $\leadsto$ 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!