Political Science 452: Text as Data

Justin Grimmer

Assistant Professor
Department of Political Science
Stanford University

April 20th, 2011
Where We’ve Been, Where We’re Going

- Class 1: Finding Text Data
- Class 2: Representing Texts Quantitatively
- Class 3: Dictionary Methods for Classification
- Class 4: Comparing Language Across Groups
- Class 5: Texts in Space
- Class 6: Clustering
- Class 7: Topic models
- Class 8: Supervised methods for classification
- Class 9: Ensemble methods for classification
- Class 10: Wild Card (What do we want to cover?)
  - Scaling speech (Ideal point estimates from text)
  - Large text collections (sparse matrices, approximate inference methods)
  - Natural Language Processing (Watson question answering)
- Applications: present your work
- Applications: The Taunting Project
Outline for Today’s Lecture

- R, A Pep Talk and some Commands/Ideas that are useful
- **Example 1**: Stylometry– Disputed Federalist papers
- General Set up of Classification Problems
- The Dictionary Solution to Classification
  - Find words that separate classes
  - Use this to infer document classes
- Separating words: off-the-shelf and proprietary
- Combination Rules
- Validation
- **Example 2**: Measuring Happiness

Can we do better with machines? (SkyNet went self aware at 8:11 pm last night, by the way)
R and Your Home Work

Thoughts on home work?

- This class assumes knowledge of basic R programming
- If you need a refresher: http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
- I'll explain the logic of any homework problem
- I'll provide tips on useful functions
- You need to connect the dots
- Learning a language: the investment is worth it!
- Rapidly developing toolkits, intelligent consumer (producer) of statistical methods
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher: http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
- I'll explain the logic of any homework problem
- I'll provide tips on useful functions
- You need to connect the dots
- Learning a language: the investment is worth it!

Rapidly developing tool kits, intelligent consumer (producer) of statistical methods

Justin Grimmer (Stanford University)
Thoughts on home work?

**Text Analysis requires hacking** (Schrodt Talk)

**R**: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher: [http://cran.r-project.org/doc/manuals/R-intro.pdf](http://cran.r-project.org/doc/manuals/R-intro.pdf)
- Policy on R: Teach you how to fish
- I'll explain the logic of any homework problem
- I'll provide tips on useful functions
- You need to connect the dots

Learning a language: the investment is worth it!

Rapidly developing tool kits, intelligent consumer (producer) of statistical methods
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher:
  http://cran.r-project.org/doc/manuals/R-intro.pdf
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher:
  http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher:
  http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
  - I’ll explain the logic of any homework problem
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher:
  http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
  - I’ll explain the logic of any homework problem
  - I’ll provide tips on useful functions
R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)
- This class assumes knowledge of basic R programming
- If you need a refresher:
  http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
  - I’ll explain the logic of any homework problem
  - I’ll provide tips on useful functions
  - You need to connect the dots
R and Your Home Work

Thoughts on home work?

**Text Analysis requires hacking** (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher:
  http://cran.r-project.org/doc/manuals/R-intro.pdf
- Policy on R: Teach you how to fish
  - I'll explain the logic of any homework problem
  - I'll provide tips on useful functions
  - You need to connect the dots
- Learning a language: the investment is worth it!
R and Your Home Work

Thoughts on home work?

**Text Analysis requires hacking** (Schrodt Talk)

R: (A Pep Talk)

- This class assumes knowledge of basic R programming
- If you need a refresher:
  - [http://cran.r-project.org/doc/manuals/R-intro.pdf](http://cran.r-project.org/doc/manuals/R-intro.pdf)
- Policy on R: Teach you how to fish
  - I’ll explain the logic of any homework problem
  - I’ll provide tips on useful functions
  - You need to connect the dots
- Learning a language: the investment is worth it!
- Rapidly developing tool kits, intelligent consumer (producer) of statistical methods
R: A Refresher

Cover three things (quickly):

1) Stemming
2) list
3) for
4) Bringing text into R
How to get stemming to work

- Fix your “class path”. Go here:
  http://weka.wikispaces.com/Stemmers
- Use the python code I posted (or modified version)
- Use different software undergrad

A brief demo (code is posted).
Lists:

- list An object in R that facilitates storage of different object types and objects of same type but different sizes
- a list object has a hierarchical indexing
- For example:

```r
> example<- list(c('a', 'b', 'c'), 1, c(1:100))
> example[[1]]  # First element of list:
[1] "a" "b" "c"
> example[[3]][10]  # 10th element of third list element
[1] 10
```
For loops:

- for loops allow the repetition of a function (set of functions) repeatedly

- Example 1: printing a number

```r
> for(j in 1:100) {
+ print(j)
}
```

- Suppose I want to sequentially add numbers 1 to 100 (without formula)

```r
> sum <- c()
> sum[1]<- 1
>for(j in 2:100){
+ sum[j]<- sum[j-1] + j }
```
Reading text into R:

```r
> files<- dir('~/Directory')  ## Directory containing text files
> example<- list()
> for(j in 1:length(files)){
+ opens = file(files[j], 'r')
+ lines = readLines(opens)
+ example[[j]]<- lines
+ close(opens)
+ }
```
Who Wrote Disputed Federalist Papers?

Federalist papers
- Persuade citizens of New York State to adopt constitution
- Canonical texts in study of American politics
- 77 essays
  - Published from 1787-1788 in Newspapers
  - And under the name Publius, anonymously

Who Wrote the Federalist papers?
- Jay wrote essays 2, 3, 4, 5, and 64
- Hamilton: wrote 43 papers
- Madison: wrote 12 papers

Disputed: Hamilton or Madison?
- Essays: 49-58, 62, and 63
- Joint Essays: 18-20

Problem: identify authors of the disputed papers.
Classify: Hamilton/Madison
How to Identify the Authors?

Mosteller and Wallace: use word counts

A strategy:
- Focus on filler (stop) words
  - Invariant to topic
  - Author’s style
- Training set: Use 1/2 of undisputed essays to identify discriminating words
- Test set: demonstrate words discriminate authorship in new texts
- Apply to disputed texts

**Discrimination function**: set of words that separate Madison + Hamilton texts.
Creating a Dictionary

Use training set to create dictionary

- Weights
  - For each word $i$, construct weight $W_i$,

\[ W_i = \frac{\mu_{i,\text{Hamilton}} - \mu_{i,\text{Madison}}}{\sigma_{i,\text{Hamilton}}^2 + \sigma_{i,\text{Madison}}^2} \]

where,

- $\mu_{i,\text{Hamilton}} \equiv$ Rate Hamilton used word $i$ (per word rate)
- $\mu_{i,\text{Madison}} \equiv$ Rate Madison used word $i$
- $\sigma_{i,\text{Hamilton}}^2 \equiv$ Variance in rate Hamilton used word $i$
- $\sigma_{i,\text{Madison}}^2 \equiv$ Variance in rate Madison used word $i$

- Trimming weights: Focus on discriminating words
- Cut off: For all $|W_i| < 0.025$ set $W_i = 0.$
Determining Authorship

For each disputed document $d$, compute discrimination statistic

$$Y_d = \sum_{i=1}^{N} W_i x_{i,d}$$

where,

- $W_i \equiv$ Weights (positive, negative) assigned to word $i$
- $x_{i,d} \equiv$ Stop Words in document $d$

$Y_d$ allows for classification (linear discriminator)

- Above midpoint in training set $\rightarrow$ Hamilton text
- Below midpoint in training set $\rightarrow$ Madison text

Findings: Madison is the author of the disputed Federalist papers.
Classification

Classification problem
Classification

Classification problem $\Rightarrow$ Authorship problem.
Classification

Classification problem $\Rightarrow$ Authorship problem.

General Features:
Classification

Classification problem $\Rightarrow$ Authorship problem.

General Features:

- Classes (known)
Classification problem ⇒ Authorship problem.

General Features:

- Classes (known) ⇒ \{ Hamilton Essay, Madison Essay \}
Classification

Classification problem $\Rightarrow$ Authorship problem.

General Features:

- Classes (known) $\Rightarrow$ \{Hamilton Essay, Madison Essay\}
- Observations $\Rightarrow$ Disputed essays
Classification

Classification problem $\Rightarrow$ Authorship problem.
General Features:

- Classes (known) $\Rightarrow$ \{Hamilton Essay, Madison Essay\}
- Observations $\Rightarrow$ Disputed essays
- Features (data) $\Rightarrow$ Word counts of stop words
Classification

Classification problem $\Rightarrow$ Authorship problem.

General Features:

- Classes (known) $\Rightarrow \{\text{Hamilton Essay, Madison Essay}\}$
- Observations $\Rightarrow$ Disputed essays
- Features (data) $\Rightarrow$ Word counts of stop words

General structure across problems
Types of Classification Problems

**Topic**: What is this text about?

- Policy area of legislation
  - Agriculture, Crime, Environment, ...
- Campaign agendas
  - Abortion, Campaign, Finance, Taxing, ...
- Sentiment on legislation
  - Positions on legislation
    - Support, Ambiguous, Oppose
  - Positions on Court Cases
    - Agree with Court, Disagree with Court
- Style/Tone
  - Taunting in floor statements
    - Partisan Taunt, Intra party taunt, Agency taunt, ...
  - Negative campaigning
    - Negative ad, Positive ad
Types of Classification Problems

**Topic**: What is this text about?
- Policy area of legislation
  ⇒ \{Agriculture, Crime, Environment, \ldots\}
- Campaign agendas
  ⇒ \{Abortion, Campaign, Finance, Taxing, \ldots\}
Types of Classification Problems

**Topic:** What is this text about?
- Policy area of legislation
  ⇒ \{Agriculture, Crime, Environment, ...\}
- Campaign agendas
  ⇒ \{Abortion, Campaign, Finance, Taxing, ...\}

**Sentiment:** What is said in this text? [Public Opinion]
Types of Classification Problems

**Topic:** What is this text about?
- Policy area of legislation
  ⇒ \{Agriculture, Crime, Environment, ...\}
- Campaign agendas
  ⇒ \{Abortion, Campaign, Finance, Taxing, ...\}

**Sentiment:** What is said in this text? [Public Opinion]
- Positions on legislation
  ⇒ \{ Support, Ambiguous, Oppose \}
- Positions on Court Cases
  ⇒ \{ Agree with Court, Disagree with Court \}
- Liberal/Conservative Blog Posts
  ⇒ \{ Liberal, Middle, Conservative, No Ideology Expressed \}
Types of Classification Problems

**Topic:** What is this text about?
- Policy area of legislation
  ⇒ {Agriculture, Crime, Environment, ...}
- Campaign agendas
  ⇒ {Abortion, Campaign, Finance, Taxing, ...}

**Sentiment:** What is said in this text? [Public Opinion]
- Positions on legislation
  ⇒ {Support, Ambiguous, Oppose}
- Positions on Court Cases
  ⇒ {Agree with Court, Disagree with Court}
- Liberal/Conservative Blog Posts
  ⇒ {Liberal, Middle, Conservative, No Ideology Expressed}

**Style/Tone:** How is it said?
Types of Classification Problems

**Topic:** What is this text about?
- Policy area of legislation
  ⇒ \{Agriculture, Crime, Environment, ...\}
- Campaign agendas
  ⇒ \{Abortion, Campaign, Finance, Taxing, ...\}

**Sentiment:** What is said in this text? [Public Opinion]
- Positions on legislation
  ⇒ \{Support, Ambiguous, Oppose\}
- Positions on Court Cases
  ⇒ \{Agree with Court, Disagree with Court\}
- Liberal/Conservative Blog Posts
  ⇒ \{Liberal, Middle, Conservative, No Ideology Expressed\}

**Style/Tone:** How is it said?
- Taunting in floor statements
  ⇒ \{Partisan Taunt, Intra party taunt, Agency taunt, ...\}
- Negative campaigning
  ⇒ \{Negative ad, Positive ad\}
Classification Using Identifying Words

Dictionary Approach to Classification: Begin with:

- Classification scheme
- Documents, Some Labeled According to Classification Scheme
  - Training Set: used to develop dictionary
  - Test Set: used to test dictionary
  - Classification Set: unlabeled documents classified using dictionary
- Quantitative Representation of Text

Steps to produce classification:

1) Identification of Separating Words
   a) Preexisting Dictionary (we will detail many today)
   b) Create own dictionary (using techniques developed next week)
   c) Create own dictionary using Mechanical Turk

2) Method to apply word measures to texts
   a) Separating plane (avoid geometry today, background)
   b) Simplest: presence/absence of terms

3) Validation
   a) Demonstrate that weights/application perform well
   b) Critical role of “test” set (calibration set)

4) Classify unlabeled documents
Word Weights: Separating Classes

General Classification Goal: Identify Features that Separate Classes

How To Find Features?

- Dictionaries:
  - Rely on Humans
  - Use humans to identify words that associate with classes
  - Measure how well words separate (positive/negative, emotional, ...)

- Supervised Classification Methods (Week 8/9):
  - Rely on statistical models
  - Given set of coded documents, statistical relationship between classes/words
  - Statistical measures of separation

Key point: this is the same task
Pre-existing dictionaries

Most common way to use dictionary: Already created sets of words

Warning: Dictionaries May Not Transfer Well Across Domains

Most common dictionaries:

- General Inquirer Database
  (http://www.wjh.harvard.edu/~inquirer/)
  - { Positive, Negative }
  - 3627 negative and positive word strings
  - Workhorse for classification across many domains/papers

- DICTION
  - { Certain, Uncertain }, { Optimistic, Pessimistic }
  - $\approx 10,000$ words

- Linguistic Inquiry Word Count
  - { Positive emotion, Negative emotion }
  - 2300 words grouped into 70 classes

- Harvard-IV-4
- Affective Norms for English Words
- ...
Generating New Words

Three methods (non-exhaustive):
- Methods next week (identify separating words)
- Manual generation
  - Careful thought (by you or group) about useful words
- Mechanical Turk
  - Example (Dodds and Danforth): { Happy, Unhappy }
  - Ask turkers: how happy is
    elevator, car, pretty, young
  Output as dictionary

Validation: does this classify well (out of sample)?
Applying Methods to Documents

After creating/selecting dictionary:
- Set of separating words $x_i$, ($i = 1, \ldots, N$)
- Weights attached to words $W_i$
  - Liberal words: -1
  - Conservative words: +1

Focus on binary classification \{liberal, conservative\}

For each document $d$ calculate score for document

$$Y_d = \sum_{i=1}^{N} W_i x_{i,d}$$

$$Y_d = W' x_d$$

Classify:

$Y_d > 0 \Rightarrow$ Conservative statement
$Y_d < 0 \Rightarrow$ Liberal statement
$Y_d \approx 0$ Ambiguous

$Y_d$ often used as measurement of categories
Validation

Classification Validity:

- Training: build dictionary on subset of documents with known labels
- Test: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
- Is the classification scheme well defined for your texts?
- Can humans accomplish the coding task?
- Is the dictionary using appropriate?

Replicate classification exercise

- How well does our method perform on held out documents?
- Why held out?

Over fitting

- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:

- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels

- Requires hand coded documents
- Hand coded documents useful for other reasons
- Is the classification scheme well defined for your texts?
- Can humans accomplish the coding task?
- Is the dictionary you are using appropriate?

Replicate classification exercise

- How well does our method perform on held out documents?
- Why held out?

---

Over fitting

- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels

- Requires hand coded documents
- Hand coded documents useful for other reasons
- Is the classification scheme well defined for your texts?
- Can humans accomplish the coding task?
- Is the dictionary you're using appropriate?

Replicate classification exercise
- How well does our method perform on held out documents?
- Why held out?

Over fitting
- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:

- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents

- Hand coded documents useful for other reasons

- Is the classification scheme well defined for your texts?
- Can humans accomplish the coding task?
- Is the dictionary your using appropriate?

Replicate classification exercise

- How well does our method perform on held out documents?
- Why held out?

Over fitting

- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons

Replicate classification exercise
- How well does our method perform on held out documents?
- Why held out?

Over fitting
- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?

Replicate classification exercise
- How well does our method perform on held out documents?
  - Why held out?

Over fitting
- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
Validation

Classification Validity:
- Training: build dictionary on subset of documents with known labels
- Test: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary your using appropriate?

Replicate classification exercise
- How well does our method perform on held out documents?
- Why held out?

Over fitting
- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: Cross-validation
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary you're using appropriate?

Replicate classification exercise
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary your using appropriate?

**Replicate** classification exercise
- How well does our method perform on **held out** documents?
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary you're using appropriate?

**Replicate** classification exercise
- How well does our method perform on **held out** documents?
- Why held out?
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary your using appropriate?

**Replicate** classification exercise
- How well does our method perform on held out documents?
- Why held out? **Over fitting**
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary you're using appropriate?

Replicate classification exercise
- How well does our method perform on **held out** documents?
- Why held out? **Over fitting**
- Using off-the-shelf dictionary: all labeled documents to test
Validation

Classification Validity:
- **Training**: build dictionary on subset of documents with known labels
- **Test**: apply dictionary method to other documents with known labels
- Requires hand coded documents
- Hand coded documents useful for other reasons
  - Is the classification scheme well defined for your texts?
  - Can humans accomplish the coding task?
  - Is the dictionary you're using appropriate?

**Replicate** classification exercise
- How well does our method perform on held out documents?
- Why held out? **Overfitting**
- Using off-the-shelf dictionary: all labeled documents to test
- Supervised learning classification: **Cross-validation**
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

- This is hard
  - Why?
    - Ambiguity in language
    - Limited working memory
    - Ambiguity in classification rules

A procedure for training coders:
1) Coding rules
2) Apply to new texts
3) Assess coder agreement (statistics coming in Week 8!)
4) Using this information, revise coding rules
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

- This is hard
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

- This is hard
- Why?
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

- This is hard
- Why?
  - Ambiguity in language

Justin Grimmer (Stanford University)
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

- This is hard
- Why?
  - Ambiguity in language
  - Limited working memory
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories
- This is hard
- Why?
  - Ambiguity in language
  - Limited working memory
  - Ambiguity in classification rules
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories
- This is hard
- Why?
  - Ambiguity in language
  - Limited working memory
  - Ambiguity in classification rules
- A procedure for training coders:
Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

- This is hard
- Why?
  - Ambiguity in language
  - Limited working memory
  - Ambiguity in classification rules
- A procedure for training coders:
  1) Coding rules
  2) Apply to new texts
  3) Assess coder agreement (statistics coming in Week 8!)
  4) Using information this information, revise coding rules
Assessing Classification

Measures of classification performance

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>True Liberal</td>
</tr>
<tr>
<td>Conservative</td>
<td>False Conservative</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{\text{True Liberal} + \text{True Conservative}}{\text{True Liberal} + \text{True Conservative} + \text{False Liberal} + \text{False Conservative}} \)

Precision (Liberal) = \( \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}} \)

Recall (Liberal) = \( \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}} \)

\( F_\text{Liberal} = 2 \cdot \frac{\text{Precision (Liberal)} \cdot \text{Recall (Liberal)}}{\text{Precision (Liberal)} + \text{Recall (Liberal)}} \)
### Assessing Classification

#### Measures of classification performance

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>True Liberal</td>
</tr>
<tr>
<td>Liberal</td>
<td>False Liberal</td>
</tr>
<tr>
<td>Conservative</td>
<td>False Conservative</td>
</tr>
<tr>
<td>Conservative</td>
<td>True Conservative</td>
</tr>
</tbody>
</table>

\[
\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}
\]
### Assessing Classification

Measures of classification performance

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>Conservative</td>
</tr>
<tr>
<td>True Liberal</td>
<td>False Liberal</td>
</tr>
<tr>
<td>False Liberal</td>
<td>True Conservative</td>
</tr>
</tbody>
</table>

**Accuracy**

\[
\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}
\]

**Precision}_{Liberal**

\[
\text{Precision}_{Liberal} = \frac{\text{TrueLib}}{\text{TrueLib} + \text{False Liberal}}
\]
Assessing Classification

Measures of classification performance

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>True Liberal</td>
</tr>
<tr>
<td>Conservative</td>
<td>False Conservative</td>
</tr>
</tbody>
</table>

**Accuracy** = \[
\frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}
\]

**Precision\text{Liberal}** = \[
\frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}}
\]

**Recall\text{Liberal}** = \[
\frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}}
\]

Under reported for dictionary classification
Assessing Classification

Measures of classification performance

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>True Liberal</td>
</tr>
<tr>
<td>Conservative</td>
<td>False Conservative</td>
</tr>
</tbody>
</table>

- **Accuracy**
  \[
  \text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}
  \]

- **Precision** \(_{\text{Liberal}}\)
  \[
  \text{Precision}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}}
  \]

- **Recall** \(_{\text{Liberal}}\)
  \[
  \text{Recall}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}}
  \]

- **F** \(_{\text{Liberal}}\)
  \[
  F_{\text{Liberal}} = \frac{2\text{Precision}_{\text{Liberal}} \times \text{Recall}_{\text{Liberal}}}{\text{Precision}_{\text{Liberal}} + \text{Recall}_{\text{Liberal}}}
  \]
# Assessing Classification

## Measures of classification performance

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liberal</td>
<td>Conservative</td>
</tr>
<tr>
<td>True Liberal</td>
<td>False Liberal</td>
</tr>
<tr>
<td>Conservative</td>
<td>False Conservative</td>
</tr>
<tr>
<td>True Conservative</td>
<td></td>
</tr>
</tbody>
</table>

### Accuracy

\[
\text{Accuracy} = \frac{\text{TrueLib} + \text{TrueCons}}{\text{TrueLib} + \text{TrueCons} + \text{FalseLib} + \text{FalseCons}}
\]

### Precision

\[
\text{Precision}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Liberal}}
\]

### Recall

\[
\text{Recall}_{\text{Liberal}} = \frac{\text{True Liberal}}{\text{True Liberal} + \text{False Conservative}}
\]

### F

\[
F_{\text{Liberal}} = \frac{2 \times \text{Precision}_{\text{Liberal}} \times \text{Recall}_{\text{Liberal}}}{\text{Precision}_{\text{Liberal}} + \text{Recall}_{\text{Liberal}}}
\]

Under reported for dictionary classification
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports

Previous state of art: Harvard-IV-4 Dictionary applied to texts

Loughran and McDonald (2011): Financial Documents are Different
- Negative words in Harvard, Not Negative in Accounting:
  - tax, cost, capital, board, liability, foreign, cancer, crude (oil), tire
- 73% of Harvard negative words in this set (!!!!!)
- Not Negative Harvard, Negative in Accounting:
  - felony, litigation, restated, misstatement, and unanticipated

Context Matters

Justin Grimmer (Stanford University)

Text as Data

April 20th, 2011 24 / 30
Validation, Dictionaries from other Fields

Accounting Research: measure **tone** of 10-K reports
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
  - tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts

Loughran and McDonald (2011): Financial Documents are Different
  - Negative words in Harvard, Not Negative in Accounting:
    - tax, cost, capital, board, liability, foreign, cancer,
    - crude (oil), tire
  - 73% of Harvard negative words in this set (!!!!!)

- Not Negative Harvard, Negative in Accounting:
  - felony, litigation, restated, misstatement, and unanticipated

Context Matters
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
  - tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts
  - Negative words in Harvard, Not Negative in Accounting:
    - tax, cost, capital, board, liability, foreign, cancer, crude (oil), tire
  - 73% of Harvard negative words in this set (!!!)
  - Not Negative Harvard, Negative in Accounting:
    - felony, litigation, restated, misstatement, and unanticipated

Context Matters
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
  - tone matters ($)
Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
- tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different
- Negative words in Harvard, Not Negative in Accounting:
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
  - tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different
  - Negative words in Harvard, Not Negative in Accounting:
    tax, cost, capital, board, liability, foreign, cancer, crude(oil), tire
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
  - tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different
  - Negative words in Harvard, Not Negative in Accounting:
    tax, cost, capital, board, liability, foreign, cancer, crude(oil), tire
  - 73% of Harvard negative words in this set(!!!!!!)
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
- tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different
- Negative words in Harvard, Not Negative in Accounting:
  tax, cost, capital, board, liability, foreign, cancer, crude(oil), tire
- 73% of Harvard negative words in this set(!!!!!!)
- Not Negative Harvard, Negative in Accounting:
Validation, Dictionaries from other Fields

Accounting Research: measure tone of 10-K reports
  - tone matters ($)

Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different
  - Negative words in Harvard, Not Negative in Accounting:
    tax, cost, capital, board, liability, foreign, cancer, crude(oil), tire
  - 73% of Harvard negative words in this set(!!!!!!)
  - Not Negative Harvard, Negative in Accounting:
    felony, litigation, restated, misstatement, and unanticipated
Face Validity: It Can Work!

Key, Huddy, Lebo, and Skiena (2011): LYDIA System
http://www.textmap.com
Face Validity: It Can Work!

Key, Huddy, Lebo, and Skiena (2011): LYDIA System
http://www.textmap.com
Measuring Happiness

- Quantifying Happiness: How happy is society?
- How Happy is a Song?
- Blog posts?
- Facebook posts? (Gross National Happiness)

Use Dictionary Methods
Measuring Happiness

Dodds and Danforth (2009): Use a dictionary method to measure happiness

- Original study: Affective Norms for English Words (Now Using mturk)
- Bradley and Lang 1999: 1034 words, Affective reaction to words
  - On a scale of 1-9 how happy does this word make you?
    Happy: triumphant (8.82)/paradise (8.72)/ love (8.72)
    Neutral: street (5.22)/ paper (5.20)/ engine (5.20)
    Unhappy: funeral (1.39)/ rape (1.25)/ suicide (1.25)

- Happiness for text \( d \) (with word \( i \) having happiness \( W_i \) and document frequency \( x_{i,d} \))

\[
\text{Happiness}_d = \frac{\sum_{i=1}^{N} W_i x_{i,d}}{\sum_{i=1}^{N} x_{i,d}}
\]
Homework Hints:
One approach: write a `for` loop searching for words in dictionary (caution: is dictionary stemmed?)
Second approach: use TDM, match with dictionary list (caution: is dictionary stemmed?)

Happiest Song on Thriller?
P.Y.T. (Pretty Young Thing) (This is the right answer!)

```
Justin Grimmer (Stanford University)
Text as Data
April 20th, 2011
28 / 30
```
Lyrics for Michael Jackson’s Billie Jean

“She was more like a **beauty queen**
from a **movie** scene.

And **mother** always told me,
be careful who you **love**.
And be careful of what you do
’cause the **lie** becomes the **truth**.

Billie Jean is not my lover,
She’s just a **girl** who claims
that I am the one.

<table>
<thead>
<tr>
<th>ANEW words</th>
<th>$v_k$</th>
<th>$f_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=1. love</td>
<td>8.72</td>
<td>1</td>
</tr>
<tr>
<td>2. mother</td>
<td>8.39</td>
<td>1</td>
</tr>
<tr>
<td>3. baby</td>
<td>8.22</td>
<td>3</td>
</tr>
<tr>
<td>4. beauty</td>
<td>7.82</td>
<td>1</td>
</tr>
<tr>
<td>5. truth</td>
<td>7.80</td>
<td>1</td>
</tr>
<tr>
<td>6. people</td>
<td>7.33</td>
<td>2</td>
</tr>
<tr>
<td>7. strong</td>
<td>7.11</td>
<td>2</td>
</tr>
<tr>
<td>8. young</td>
<td>6.89</td>
<td>2</td>
</tr>
<tr>
<td>9. girl</td>
<td>6.87</td>
<td>4</td>
</tr>
<tr>
<td>10. movie</td>
<td>6.86</td>
<td>1</td>
</tr>
<tr>
<td>11. perfume</td>
<td>6.76</td>
<td>1</td>
</tr>
<tr>
<td>12. queen</td>
<td>6.44</td>
<td>1</td>
</tr>
<tr>
<td>13. name</td>
<td>5.55</td>
<td>1</td>
</tr>
<tr>
<td>14. lie</td>
<td>2.79</td>
<td>1</td>
</tr>
</tbody>
</table>

Homework Hints: One approach: write a for loop searching for words in dictionary (caution: is dictionary stemmed?)
Homework Hints: One approach: write a for loop searching for words in dictionary (caution: is dictionary stemmed?)
Second approach: use TDM, match with dictionary list (caution: is dictionary stemmed?)
Homework Hints: One approach: write a `for` loop searching for words in dictionary (caution: is dictionary stemmed?)
Second approach: use TDM, match with dictionary list (caution: is dictionary stemmed?)
Happiest Song on Thriller?

### Lyrics for Michael Jackson’s Billie Jean

“She was more like a **beauty queen** from a **movie** scene.

And **mother** always told me,
be careful who you **love**.

And be careful of what you do 'cause the **lie** becomes the **truth**.

Billie Jean is not my lover,
She's just a **girl** who claims
that I am the one.

---

**ANEW words**

<table>
<thead>
<tr>
<th>$k$</th>
<th>$v_k$</th>
<th>$f_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>love</td>
<td>8.72</td>
</tr>
<tr>
<td>2.</td>
<td>mother</td>
<td>8.39</td>
</tr>
<tr>
<td>3.</td>
<td>baby</td>
<td>8.22</td>
</tr>
<tr>
<td>4.</td>
<td>beauty</td>
<td>7.82</td>
</tr>
<tr>
<td>5.</td>
<td>truth</td>
<td>7.80</td>
</tr>
<tr>
<td>6.</td>
<td>people</td>
<td>7.33</td>
</tr>
<tr>
<td>7.</td>
<td>strong</td>
<td>7.11</td>
</tr>
<tr>
<td>8.</td>
<td>young</td>
<td>6.89</td>
</tr>
<tr>
<td>9.</td>
<td>girl</td>
<td>6.87</td>
</tr>
<tr>
<td>10.</td>
<td>movie</td>
<td>6.86</td>
</tr>
<tr>
<td>11.</td>
<td>perfume</td>
<td>6.76</td>
</tr>
<tr>
<td>12.</td>
<td>queen</td>
<td>6.44</td>
</tr>
<tr>
<td>13.</td>
<td>name</td>
<td>5.55</td>
</tr>
<tr>
<td>14.</td>
<td>lie</td>
<td>2.79</td>
</tr>
</tbody>
</table>

$$u_{\text{Billie Jean}} = 7.1$$

$$u_{\text{Thriller}} = 6.3$$

$$u_{\text{Michael Jackson}} = 6.4$$
Homework Hints: One approach: write a for loop searching for words in dictionary (caution: is dictionary stemmed?)
Second approach: use TDM, match with dictionary list (caution: is dictionary stemmed?)
Happiest Song on Thriller?
P.Y.T. (Pretty Young Thing) (This is the right answer!)
Happiness in Society

![Graph showing the trend of happiness over years from 1960 to 2000. The x-axis represents the years, and the y-axis represents the valence (v). The graph shows a downward trend with fluctuations. ]
Happiness in Society

![Graph showing the valence of different music genres over time. The x-axis represents the year from 1965 to 2010, and the y-axis represents valence. Different genres are represented by different symbols and colors, with labels for each genre and their average valence score.](image-url)
Happiness in Society

A

valence (V)

6.1
6
5.9
5.8
5.7
5.6
5.5

13 20 30 40 50 60 70 80

blogger age

B

latitude (degrees)

5.85
5.8
5.75
5.7
5.6

0 10 20 30 40 50 60 70

C

day of week

5.85
5.84
5.83

W T F S S M T W
This week: introduction to classification, dictionaries
Next week: Know classes, infer differences in language
Then: Geometry of texts