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?

R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R and Your Home Work

Thoughts on home work?

Text Analysis requires hacking (Schrodt Talk)

R: (A Pep Talk)

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
- 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).

R: A Refresher

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:

```
> 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
```

R : A Refresher

For loops:

- `for` loops allow the repetition of a function (set of functions) repeatedly
- Example 1: printing a number

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

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

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

R : A Refresher

Reading text into 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,

$$W_i = \frac{\mu_{i,\text{Hamilton}} - \mu_{i,\text{Madison}}}{\sigma^2_{i,\text{Hamilton}} + \sigma^2_{i,\text{Madison}}}$$

where,

$$\begin{aligned}
\mu_{i,\text{Hamilton}} &\equiv \text{Rate Hamilton used word } i \text{ (per word rate)} \\
\mu_{i,\text{Madison}} &\equiv \text{Rate Madison used word } i \\
\sigma^2_{i,\text{Hamilton}} &\equiv \text{Variance in rate Hamilton used word } i \\
\sigma^2_{i,\text{Madison}} &\equiv \text{Variance in rate Madison used word } i
\end{aligned}$$

- 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,

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

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

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

- Classes (known) $\Rightarrow$ {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) $\Rightarrow$ {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?

Types of Classification Problems

Topic: What is this text about?

- Policy area of legislation
  $\Rightarrow$ {Agriculture, Crime, Environment, ...}
- Campaign agendas
  $\Rightarrow$ {Abortion, Campaign, Finance, Taxing, ... }

Types of Classification Problems

Topic: What is this text about?

- Policy area of legislation
  $\Rightarrow$ {Agriculture, Crime, Environment, ...}
- Campaign agendas
  $\Rightarrow$ {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
  $\Rightarrow$ {Agriculture, Crime, Environment, ...}
- Campaign agendas
  $\Rightarrow$ {Abortion, Campaign, Finance, Taxing, ... }

Sentiment: What is said in this text? [Public Opinion]

- Positions on legislation
  $\Rightarrow$ { Support, Ambiguous, Oppose }
- Positions on Court Cases
  $\Rightarrow$ { Agree with Court, Disagree with Court }
- Liberal/Conservative Blog Posts
  $\Rightarrow$ { Liberal, Middle, Conservative, No Ideology Expressed }

Types of Classification Problems

Topic: What is this text about?

- Policy area of legislation
  $\Rightarrow$ {Agriculture, Crime, Environment, ...}
- Campaign agendas
  $\Rightarrow$ {Abortion, Campaign, Finance, Taxing, ... }

Sentiment: What is said in this text? [Public Opinion]

- Positions on legislation
  $\Rightarrow$ { Support, Ambiguous, Oppose }
- Positions on Court Cases
  $\Rightarrow$ { Agree with Court, Disagree with Court }
- Liberal/Conservative Blog Posts
  $\Rightarrow$ { 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
  $\Rightarrow$ {Agriculture, Crime, Environment, ...}
- Campaign agendas
  $\Rightarrow$ {Abortion, Campaign, Finance, Taxing, ... }

Sentiment: What is said in this text? [Public Opinion]

- Positions on legislation
  $\Rightarrow$ { Support, Ambiguous, Oppose }
- Positions on Court Cases
  $\Rightarrow$ { Agree with Court, Disagree with Court }
- Liberal/Conservative Blog Posts
  $\Rightarrow$ { Liberal, Middle, Conservative, No Ideology Expressed }

Style/Tone: How is it said?

- Taunting in floor statements
  $\Rightarrow$ { Partisan Taunt, Intra party taunt, Agency taunt, ... }
- Negative campaigning
  $\Rightarrow$ { 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

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

$$Y_d = \mathbf{W}' \mathbf{x}_d$$

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:

Validation

Classification Validity:

- Training: build dictionary on subset of documents with known labels

Validation

Classification Validity:

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

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

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

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?

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?

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

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 your 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 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

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

Hand Coding: A Brief Digression

Humans should be able to classify documents into categories

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

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

	Actual Label	
Guess	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

Assessing Classification

Measures of classification performance

	Actual Label	
Guess	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

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

Assessing Classification

Measures of classification performance

	Actual Label	
Guess	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

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

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

Assessing Classification

Measures of classification performance

	Actual Label	
Guess	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

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

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

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

Assessing Classification

Measures of classification performance

	Actual Label	
Guess	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

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

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

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

$$F_{\text{Liberal}} = \frac{2\text{Precision}_{\text{Liberal}}\text{Recall}_{\text{Liberal}}}{\text{Precision}_{\text{Liberal}} + \text{Recall}_{\text{Liberal}}}$$

Assessing Classification

Measures of classification performance

	Actual Label	
Guess	Liberal	Conservative
Liberal	True Liberal	False Liberal
Conservative	False Conservative	True Conservative

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

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

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

$$F_{\text{Liberal}} = \frac{2\text{Precision}_{\text{Liberal}}\text{Recall}_{\text{Liberal}}}{\text{Precision}_{\text{Liberal}} + \text{Recall}_{\text{Liberal}}}$$

Under reported for dictionary classification

Validation, Dictionaries from other Fields

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 ($)

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

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 ($)

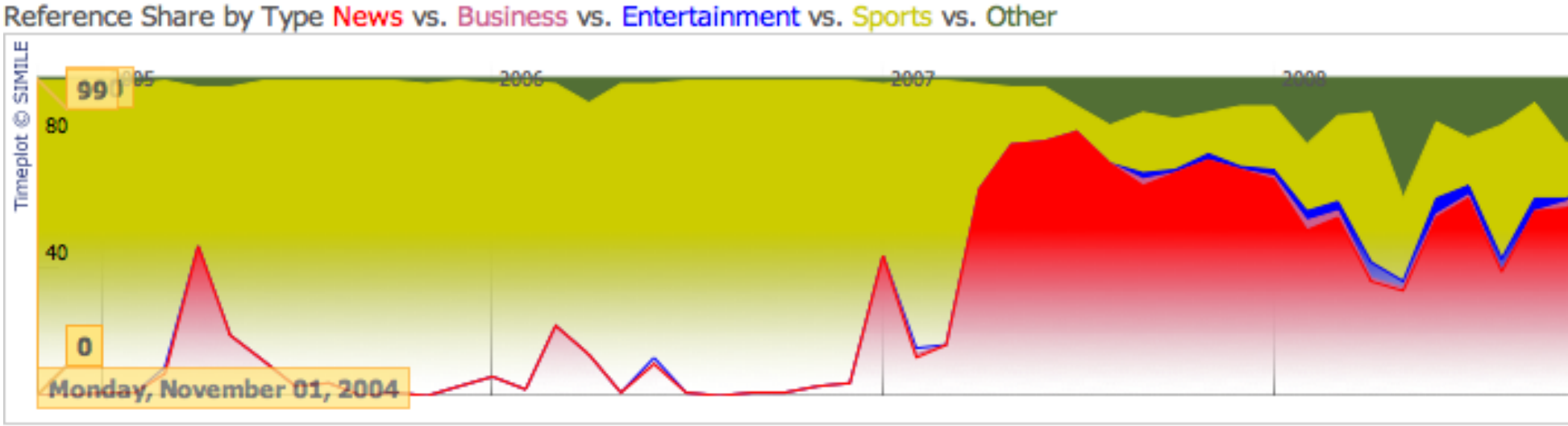
Previous state of art: Harvard-IV-4 Dictionary applied to texts
Loughran and McDonald (2011): Financial Documents are Different

- 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, andunanticipated`

Face Validity: It Can Work!

Key, Huddy, Lebo, and Skiena (2011): LYDIA System

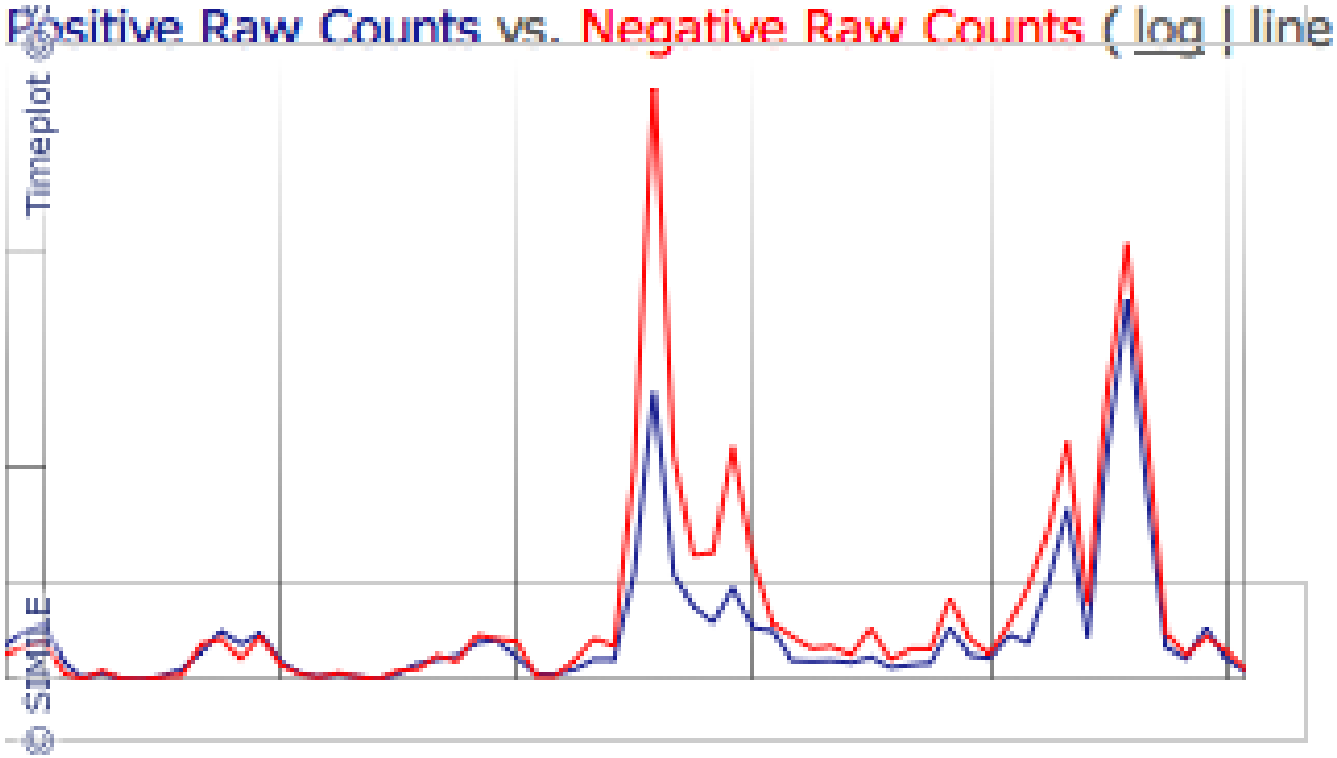
`http://www.textmap.com`



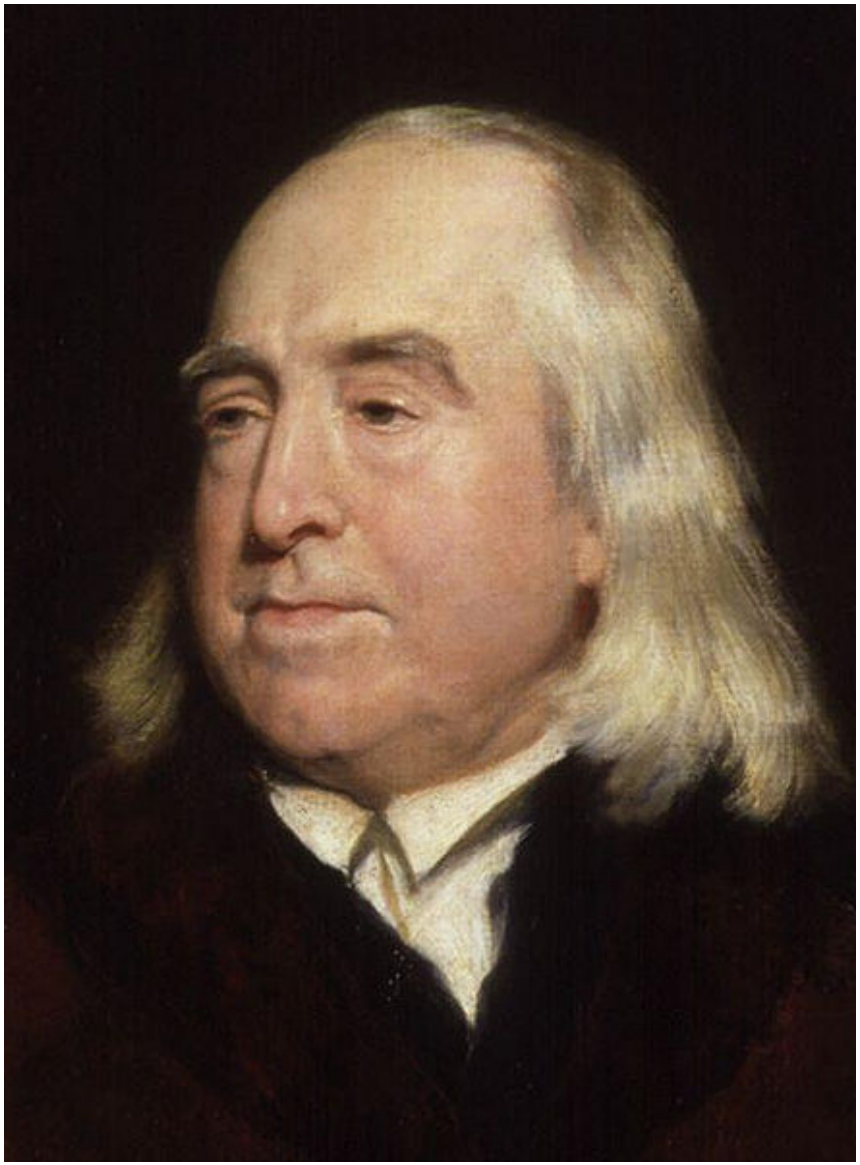
Face Validity: It Can Work!

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 frequence $x_{i,d}$

$$\text{Happiness}_d = \frac{\sum_{i=1}^{N} W_i x_{i,d}}{\sum_{i=1}^{N} x_{i,d}}$$

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.
⋮

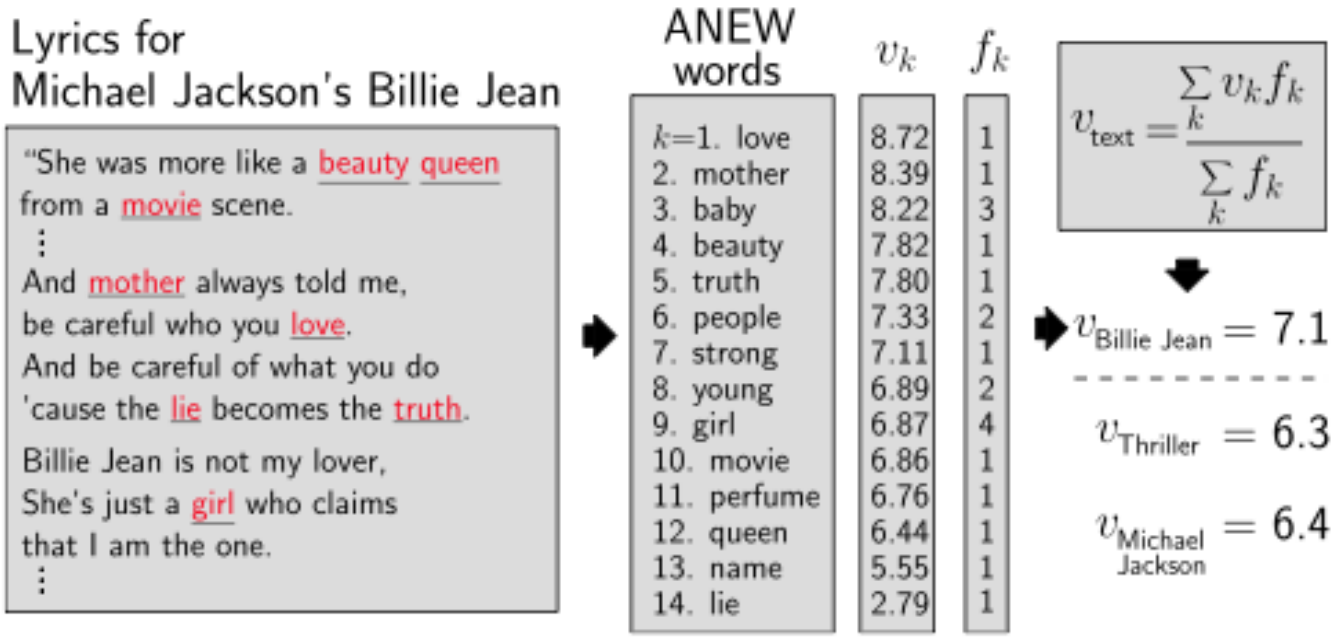
ANEW words	v_k	f_k
k=1. love	8.72	1
2. mother	8.39	1
3. baby	8.22	3
4. beauty	7.82	1
5. truth	7.80	1
6. people	7.33	2
7. strong	7.11	1
8. young	6.89	2
9. girl	6.87	4
10. movie	6.86	1
11. perfume	6.76	1
12. queen	6.44	1
13. name	5.55	1
14. lie	2.79	1

$$v_{\text{text}} = \frac{\sum_k v_k f_k}{\sum_k f_k}$$

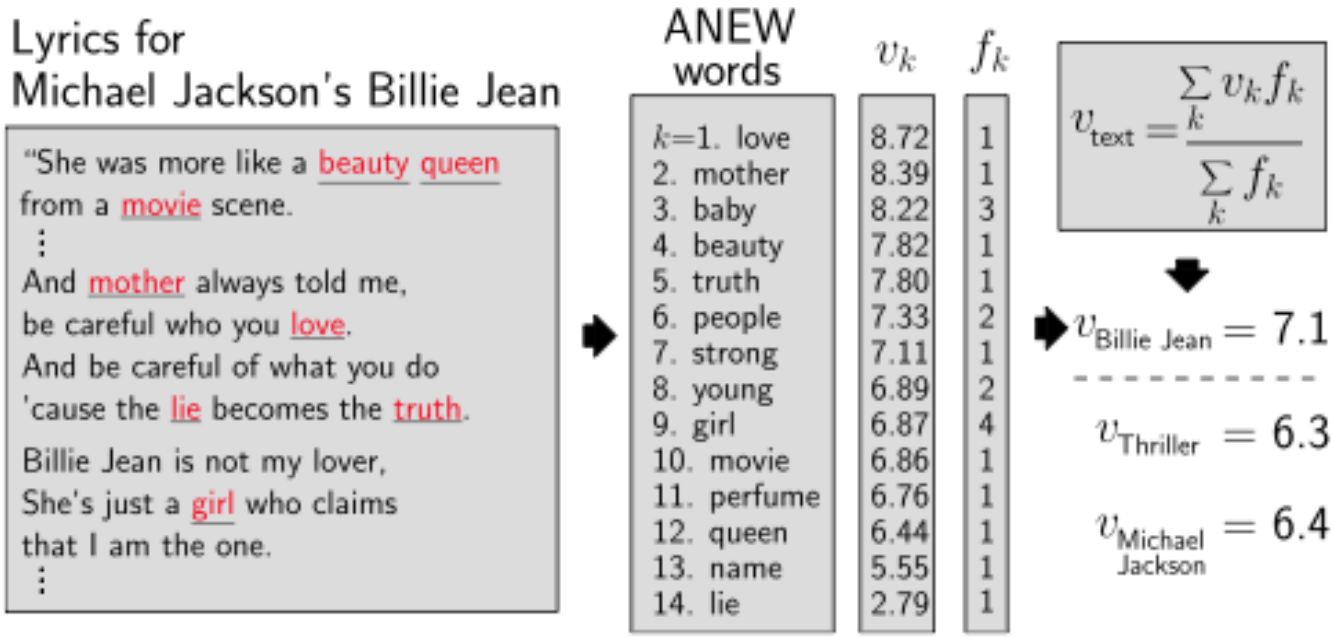
$v_{\text{Billie Jean}} = 7.1$

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

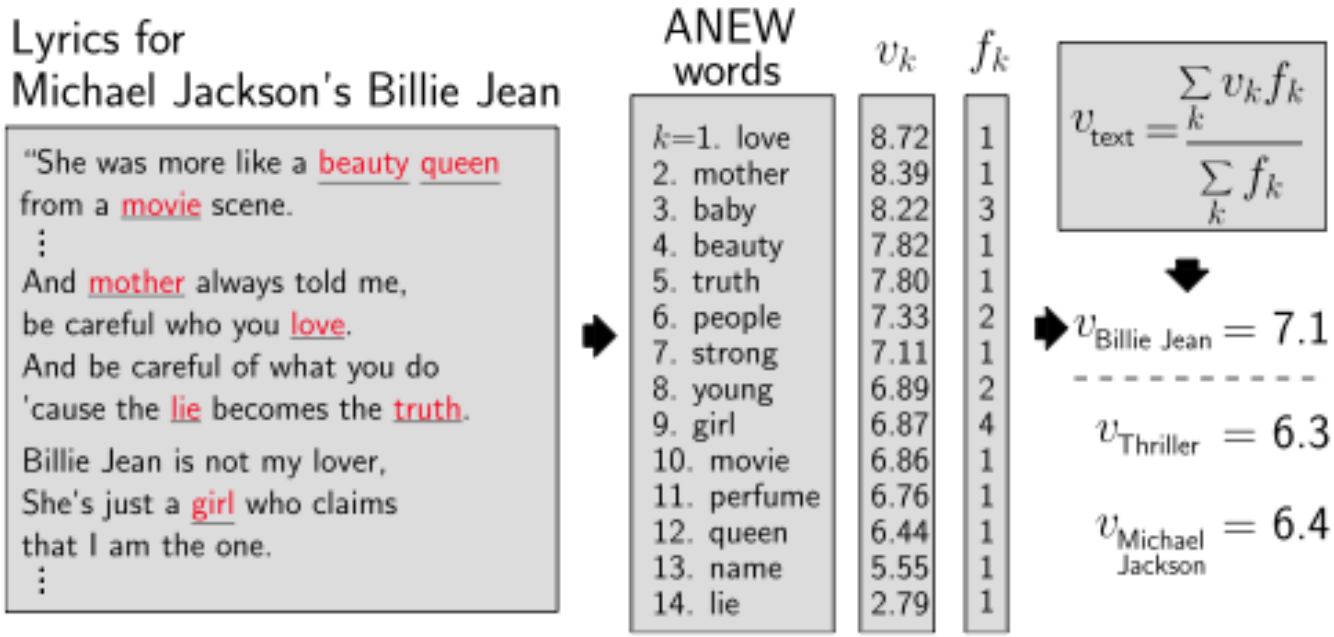
$v_{\text{Michael Jackson}} = 6.4$



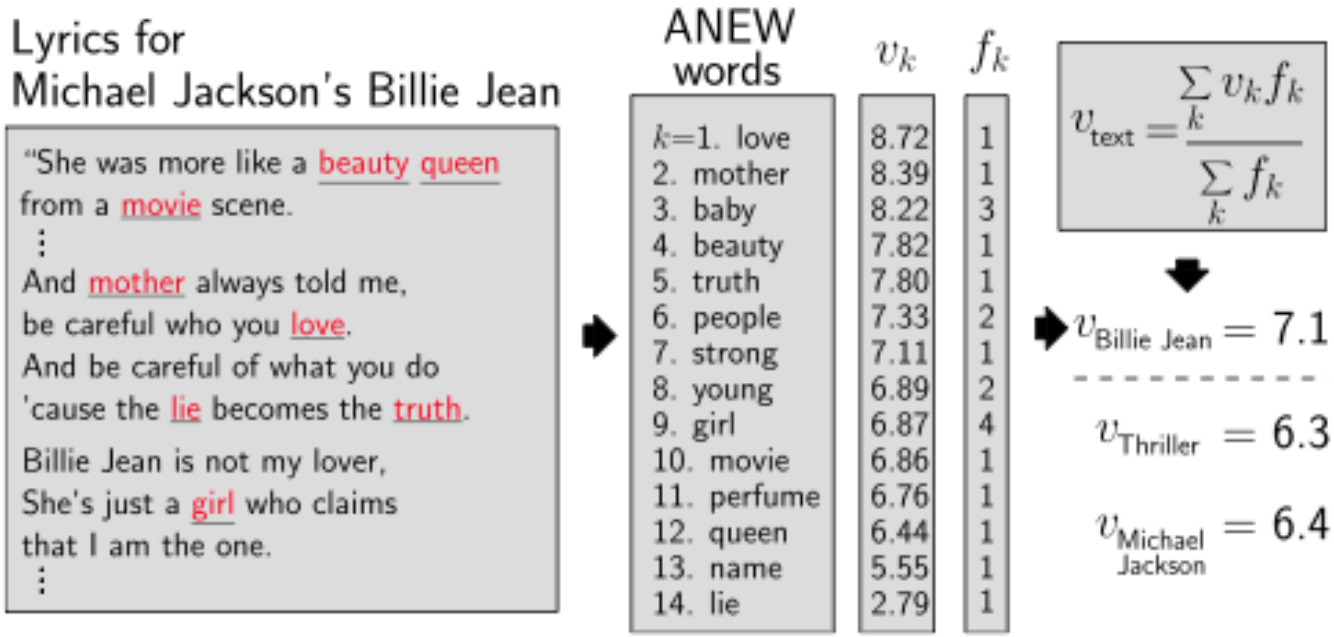
Homework Hints: One approach: write a `for` loop searching for words in dictionary (caution: is dictionary stemmed?)



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?)

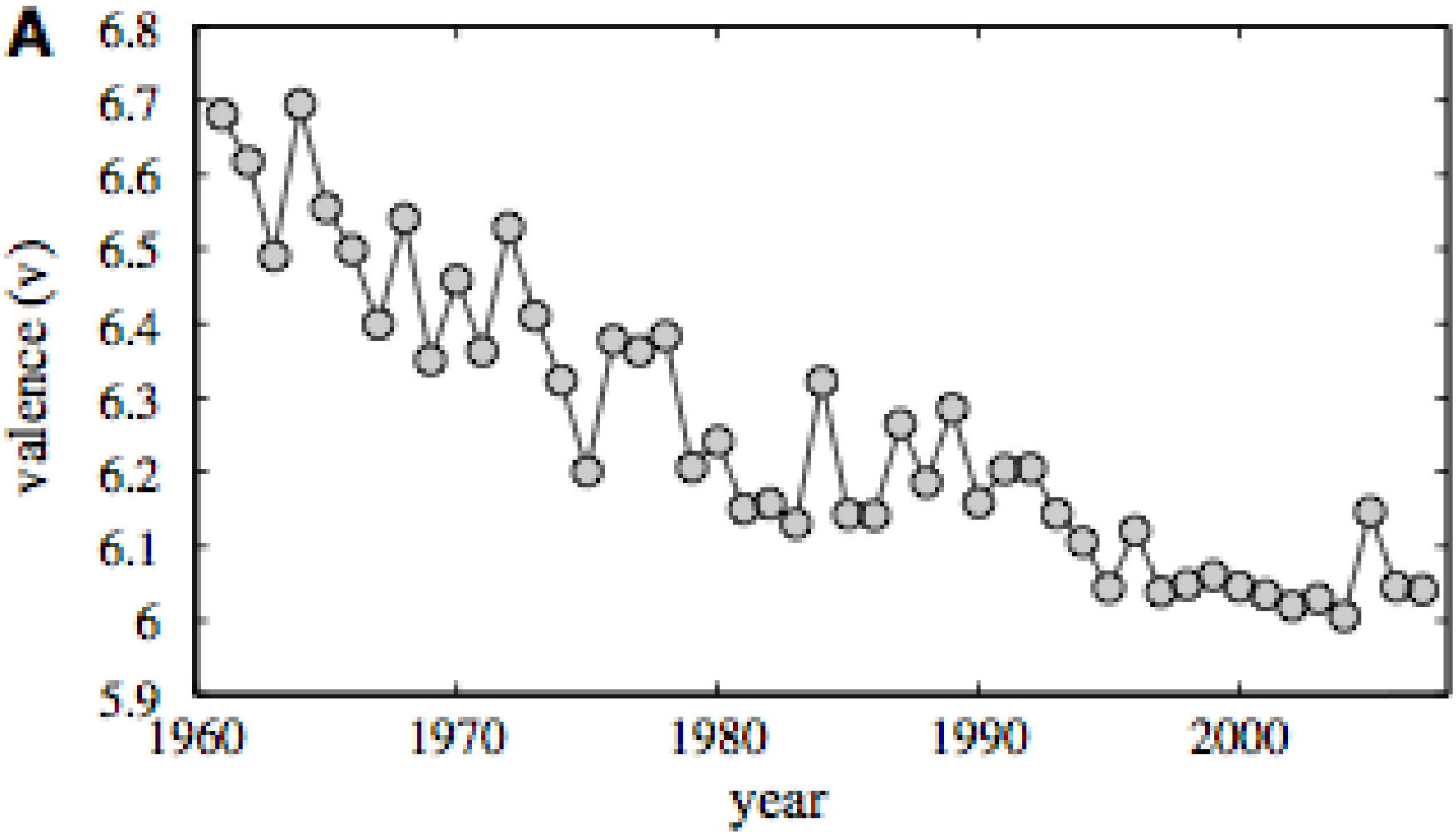


Homework Hints: One approach: write a `for` loop searching for words in dictionary (caution: is dictionary stemmed?)
Second approach: use TDM, match with dictionary list (caution: is dictionary stemmed?)
Happiest Song on Thriller?

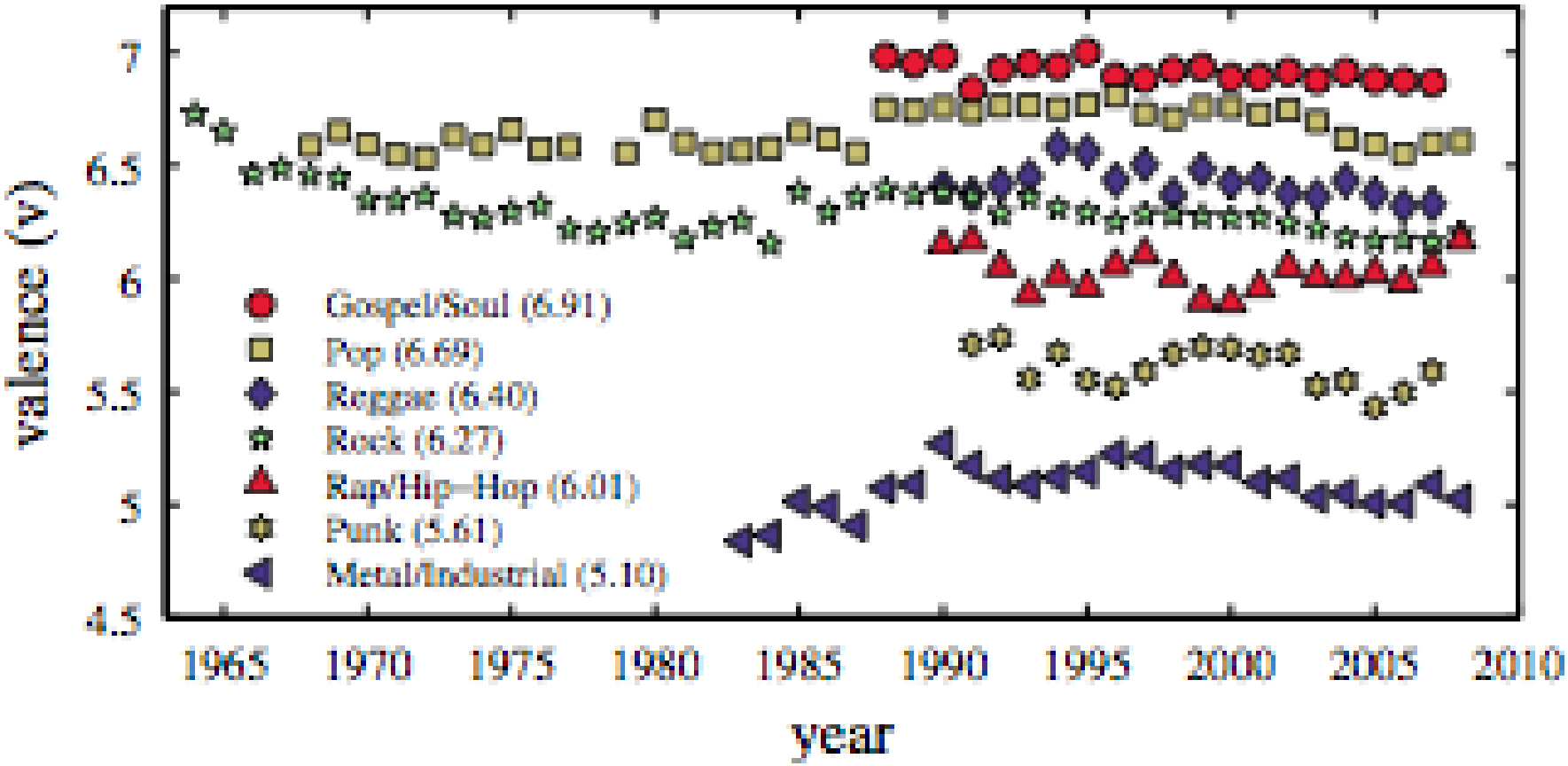


Homework Hints: One approach: write a `for` loop searching for words in dictionary (caution: is dictionary stemmed?)
Second approach: use TDM, match with dictionary list (caution: is dictionary stemmed?)
Happiest Song on Thriller?
P.Y.T. (Pretty Young Thing) (This is the right answer!)

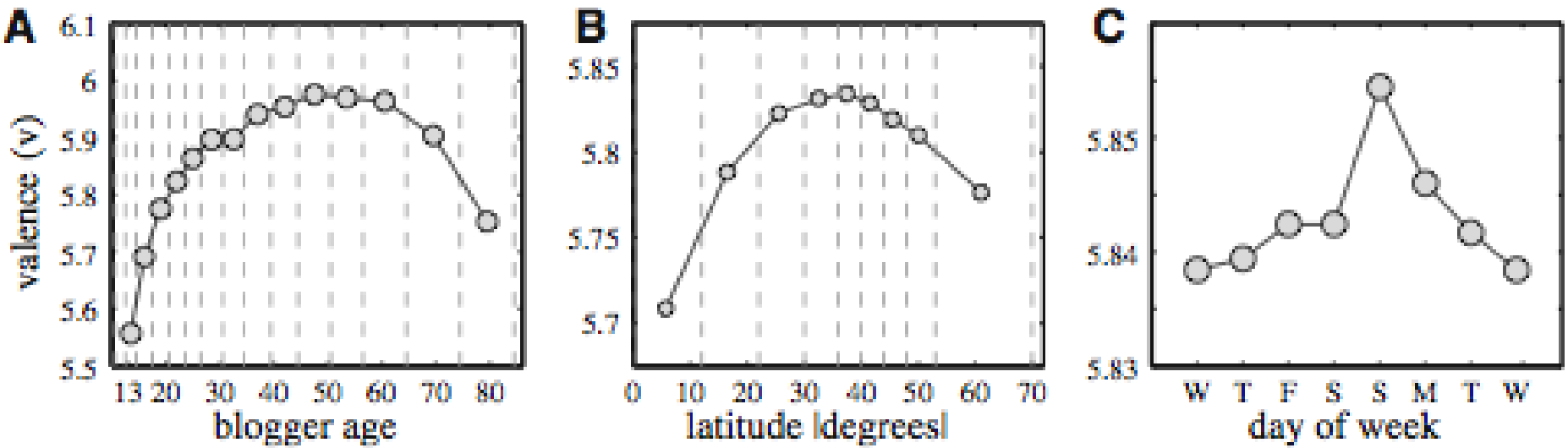
Happiness in Society



Happiness in Society



Happiness in Society



This week: introduction to classification, dictionaries
Next week: Know classes, infer differences in language
Then: Geometry of texts