Web Scraping & Text Mining

Paulo Serôdio

Postdoctoral Researcher
School of Economics
Universitat de Barcelona

May 14, 2018

Introduction

Introduction

Introduction

Introduction

Introduction

Your background

Course Materials & Structure

http://www.pauloserodio.com/eui2018

()

June 2, 2017

What this class is about. . .

Text as the new frontier of. . .

data: lots of it (literally petabytes) on the web. . . not to mention archives.

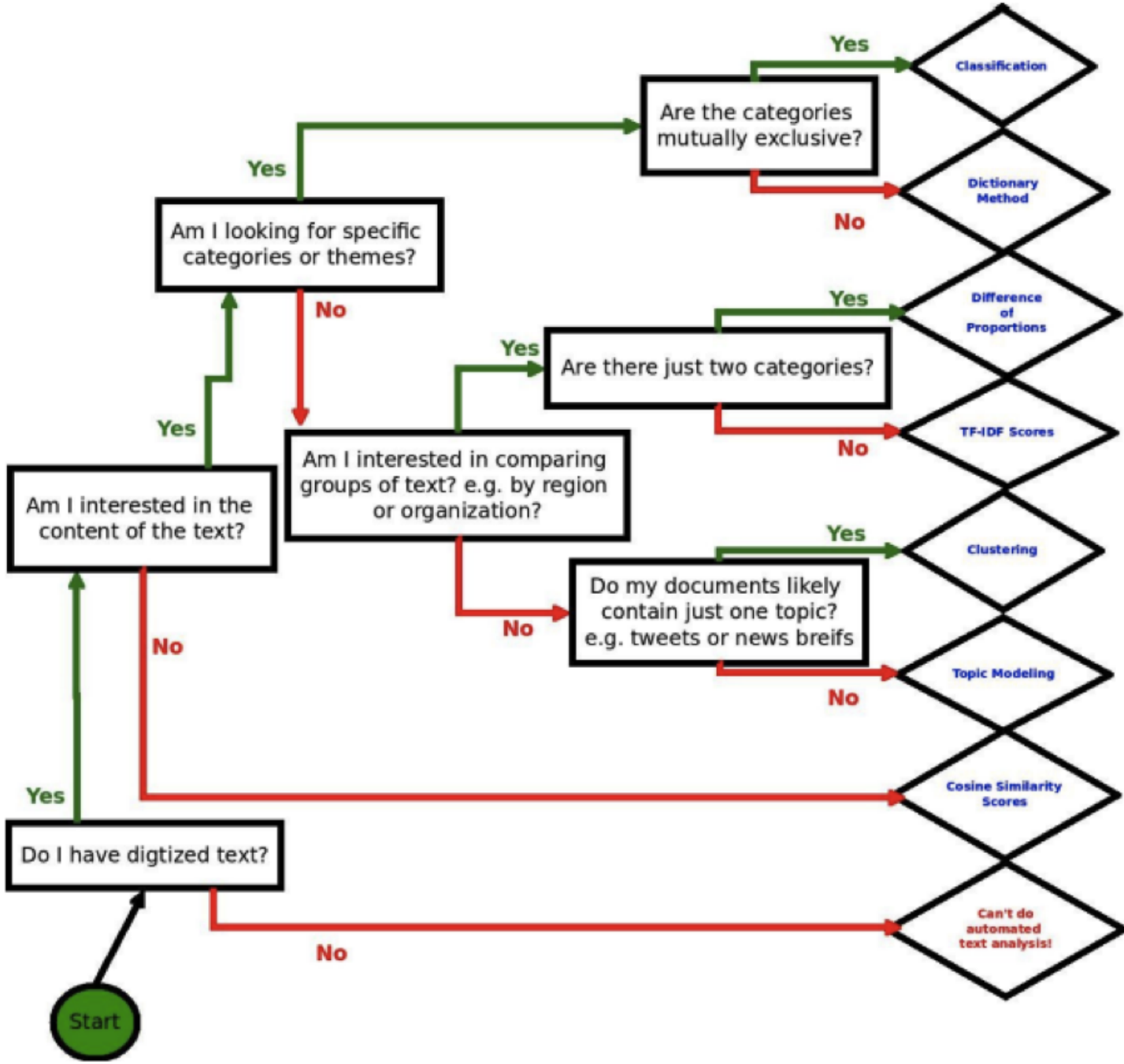
methods: unstructured data needs to be harvested and modeled.

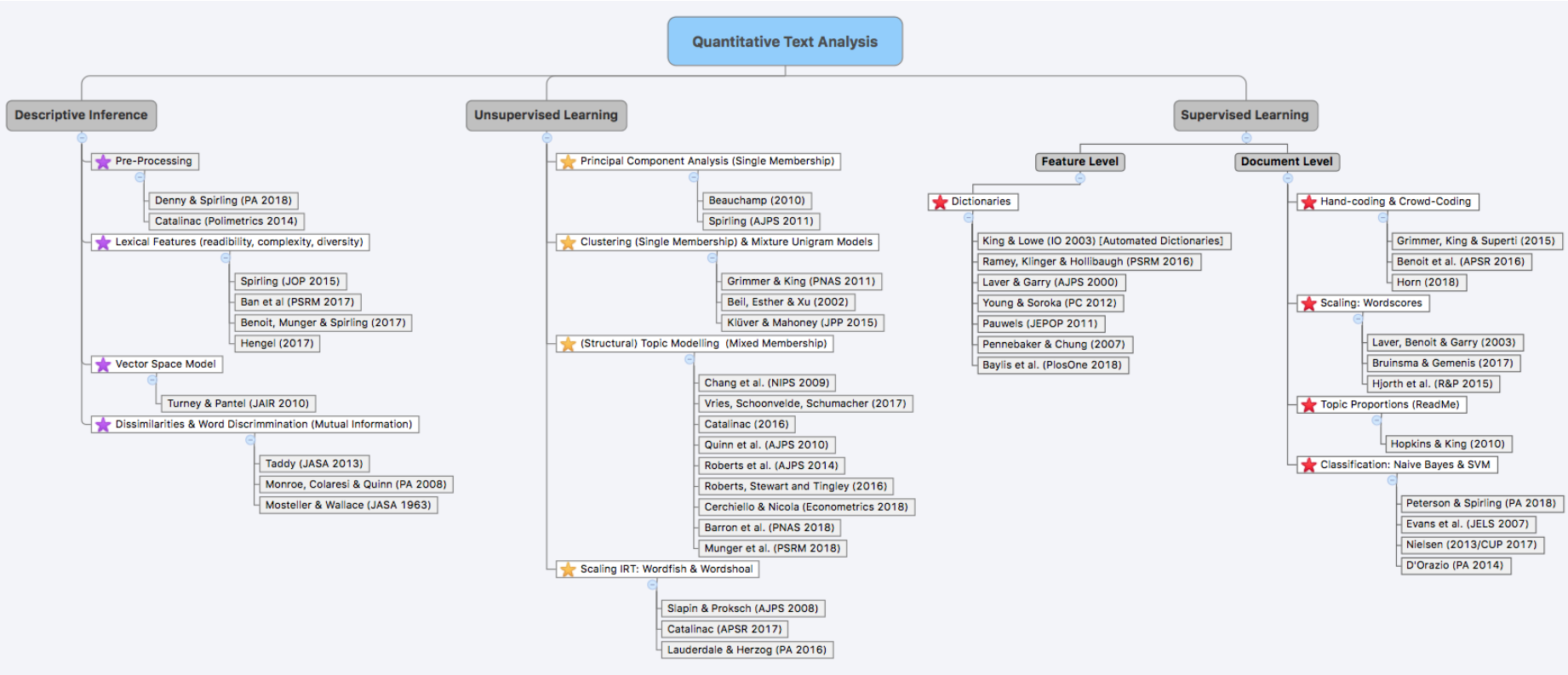
social science: politicians give speeches, thinkers write articles, nations sign treaties, users connect on Facebook etc.

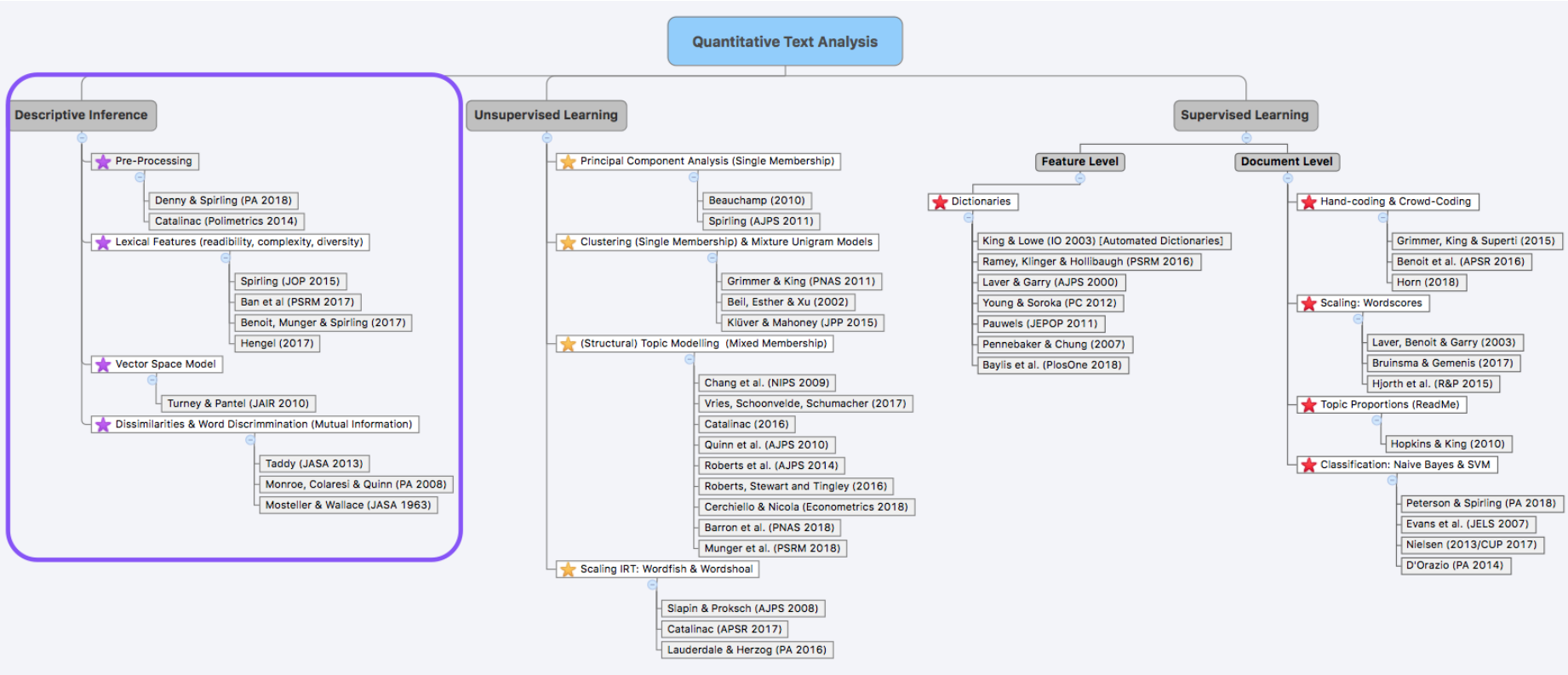
Introduction to quantitative 'text-as-data' approaches as strategies to learn more about social scientific phenomena of interest.

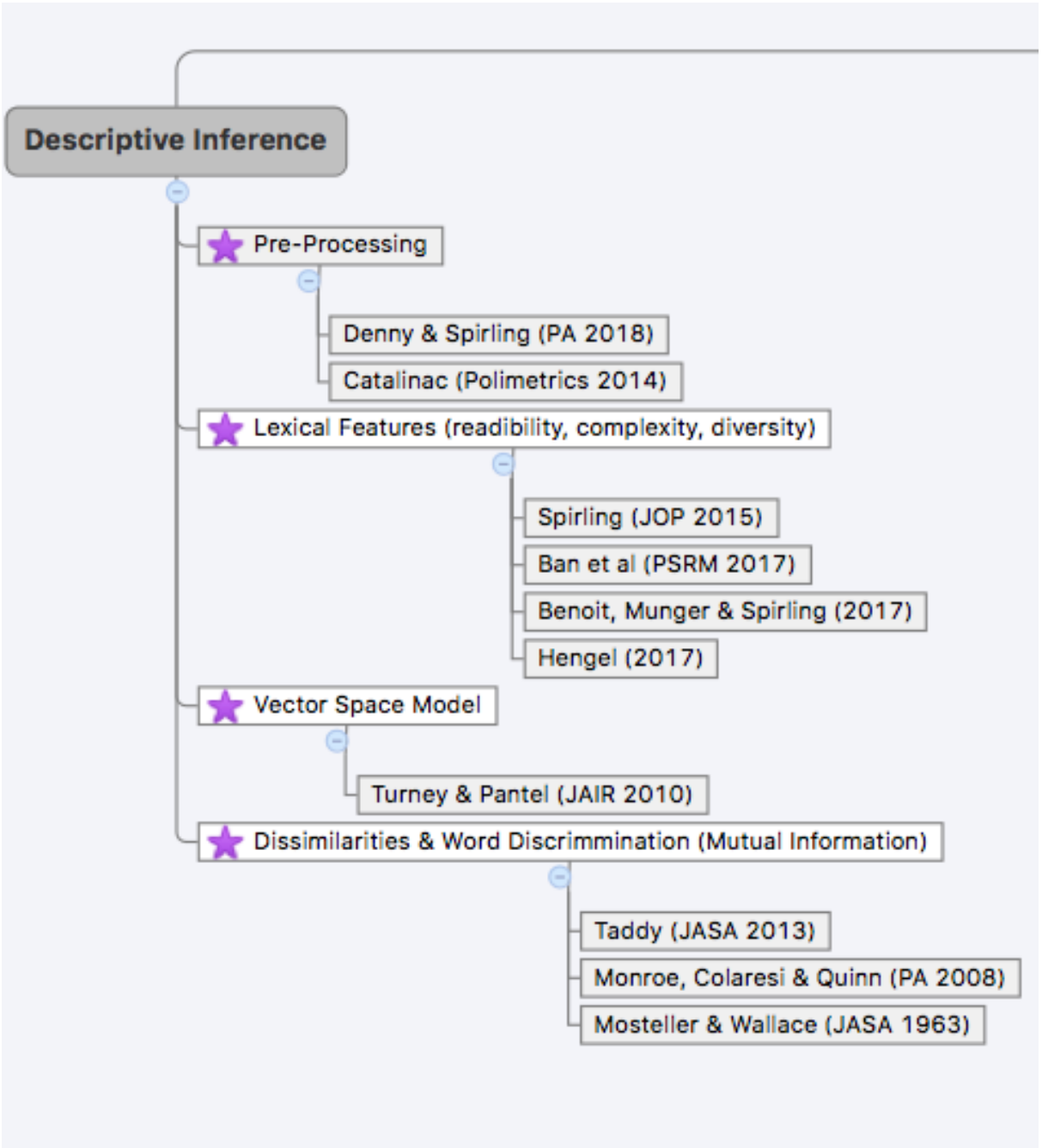
Overview

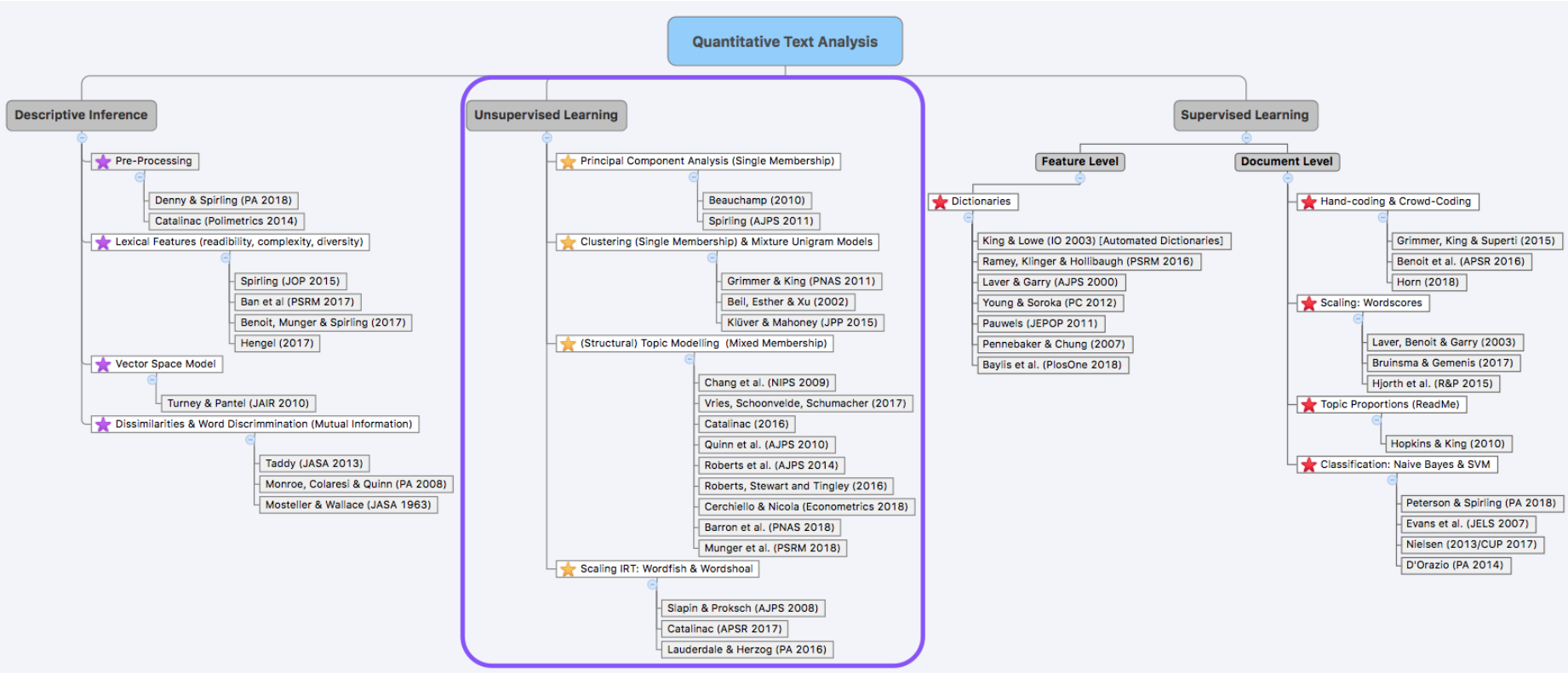
- Descriptive inference: how to characterize text, vector space model, collocations, bag-of-words, (dis)similarity measures, diversity, complexity, style, bursts.
- Basic supervised techniques: dictionaries, sentiment, events, scaling.
- Basic unsupervised techniques: clusters, scaling, topics.

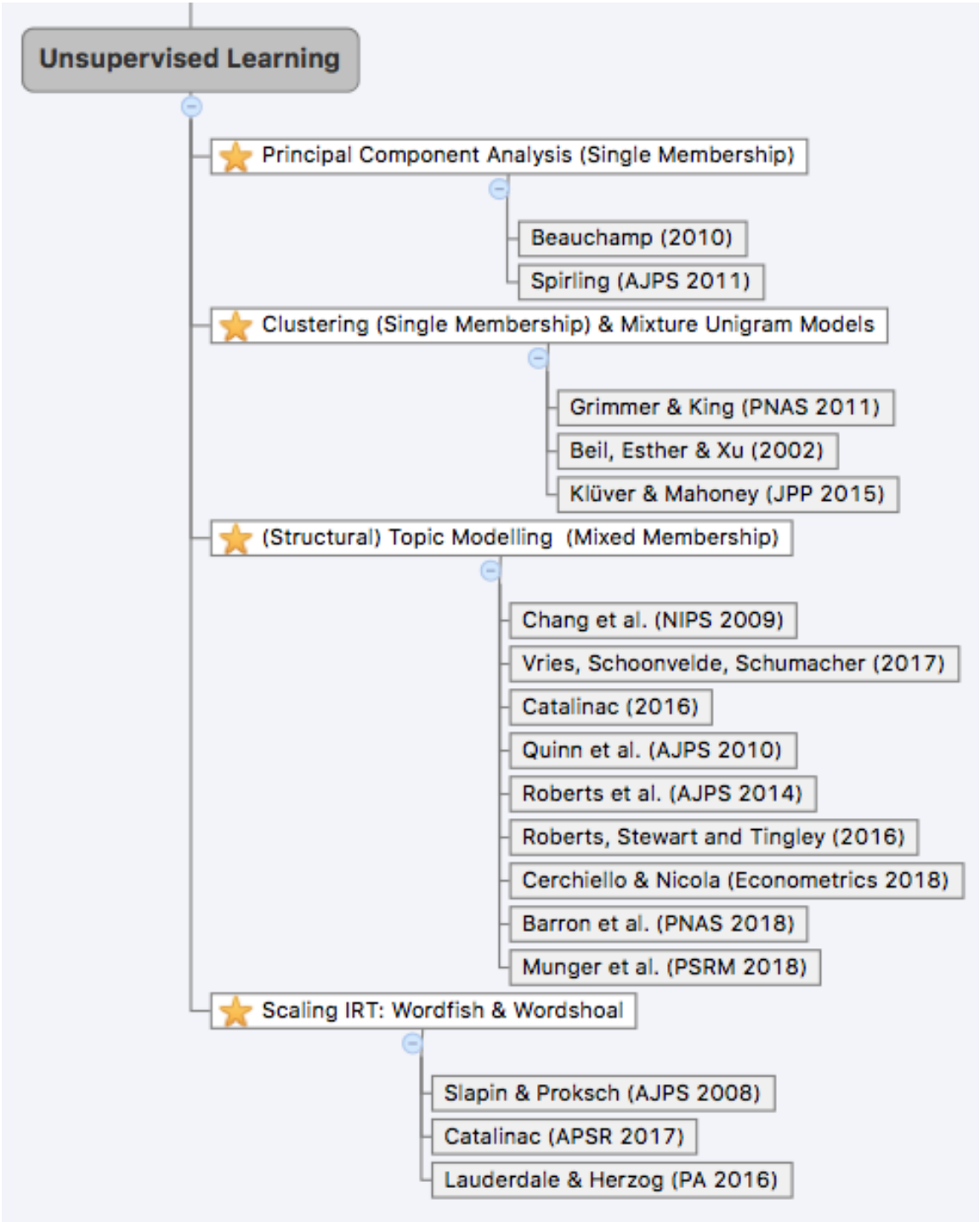

Start
Do I have digtized text?
Yes
No
Am I interested in the content of the text?
Yes
No
Am I looking for specific categories or themes?
Yes
No
Are the categories mutually exclusive?
Yes
No
Classification
Dictionary Method
Am I interested in comparing groups of text? e.g. by region or organization?
Yes
No
Are there just two categories?
Yes
No
Difference of Proportions
TF-IDF Scores
Do my documents likely contain just one topic? e.g. tweets or news breifs
Yes
No
Clustering
Topic Modeling
Cosine Similarity Scores
Can't do automated text analysis!

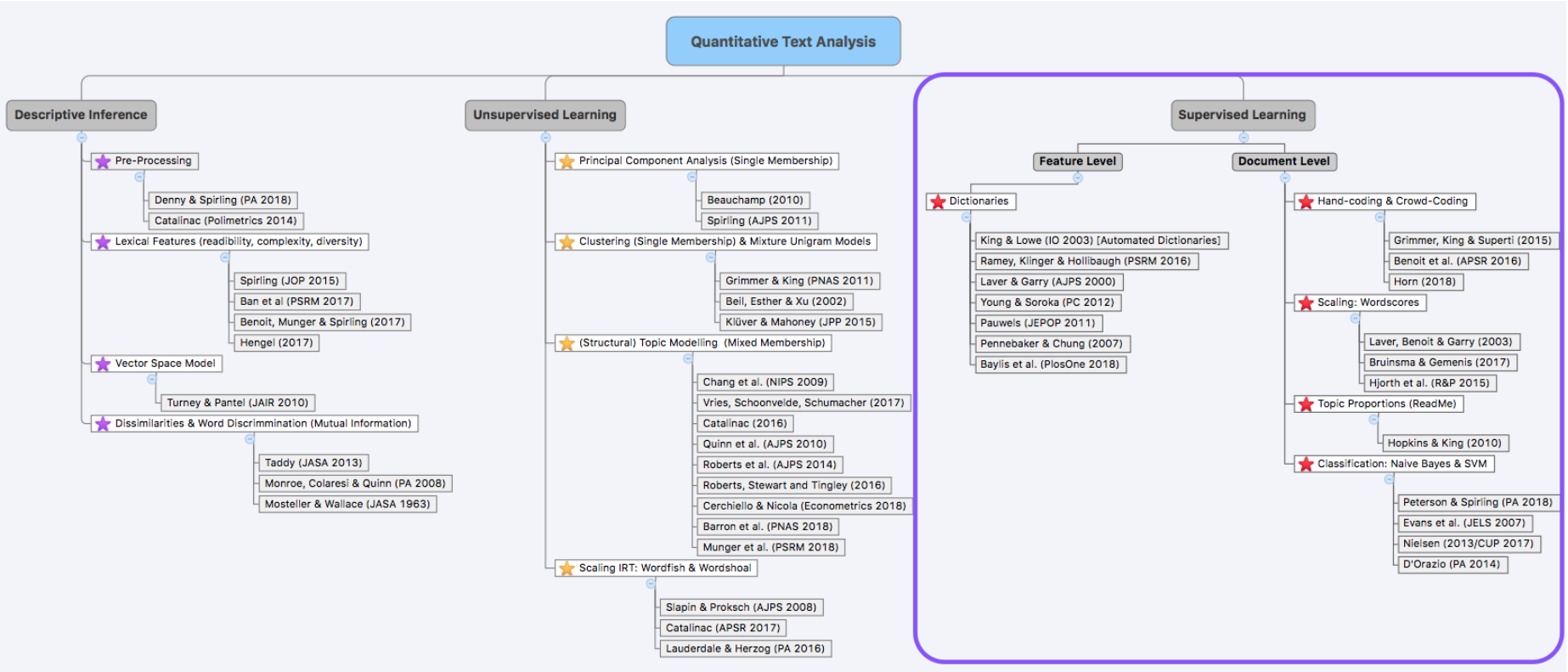

Quantitative Text Analysis
Descriptive Inference
Pre-Processing
Denny & Spirling (PA 2018)
Catalinac (Polimetrics 2014)
Lexical Features (readibility, complexity, diversity)
Spirling (JOP 2015)
Ban et al (PSRM 2017)
Benoit, Munger & Spirling (2017)
Hengel (2017)
Vector Space Model
Turney & Pantel (JAIR 2010)
Dissimilarities & Word Discrimmination (Mutual Information)
Taddy (JASA 2013)
Monroe, Colaresi & Quinn (PA 2008)
Mosteller & Wallace (JASA 1963)
Unsupervised Learning
Principal Component Analysis (Single Membership)
Beauchamp (2010)
Spirling (AJPS 2011)
Clustering (Single Membership) & Mixture Unigram Models
Grimmer & King (PNAS 2011)
Beil, Esther & Xu (2002)
Klüver & Mahoney (JPP 2015)
(Structural) Topic Modelling (Mixed Membership)
Chang et al. (NIPS 2009)
Vries, Schoonvelde, Schumacher (2017)
Catalinac (2016)
Quinn et al. (AJPS 2010)
Roberts et al. (AJPS 2014)
Roberts, Stewart and Tingley (2016)
Cerchiello & Nicola (Econometrics 2018)
Barron et al. (PNAS 2018)
Munger et al. (PSRM 2018)
Scaling IRT: Wordfish & Wordshoal
Slapin & Proksch (AJPS 2008)
Catalinac (APSR 2017)
Lauderdale & Herzog (PA 2016)
Supervised Learning
Feature Level
Dictionaries
King & Lowe (IO 2003) [Automated Dictionaries]
Ramey, Klinger & Hollibaugh (PSRM 2016)
Laver & Garry (AJPS 2000)
Young & Soroka (PC 2012)
Pauwels (JEPOP 2011)
Pennebaker & Chung (2007)
Baylis et al. (PlosOne 2018)
Document Level
Hand-coding & Crowd-Coding
Grimmer, King & Superti (2015)
Benoit et al. (APSR 2016)
Horn (2018)
Scaling: Wordscores
Laver, Benoit & Garry (2003)
Bruinsma & Gemenis (2017)
Hjorth et al. (R&P 2015)
Topic Proportions (ReadMe)
Hopkins & King (2010)
Classification: Naive Bayes & SVM
Peterson & Spirling (PA 2018)
Evans et al. (JELS 2007)
Nielsen (2013/CUP 2017)
D'Orazio (PA 2014)

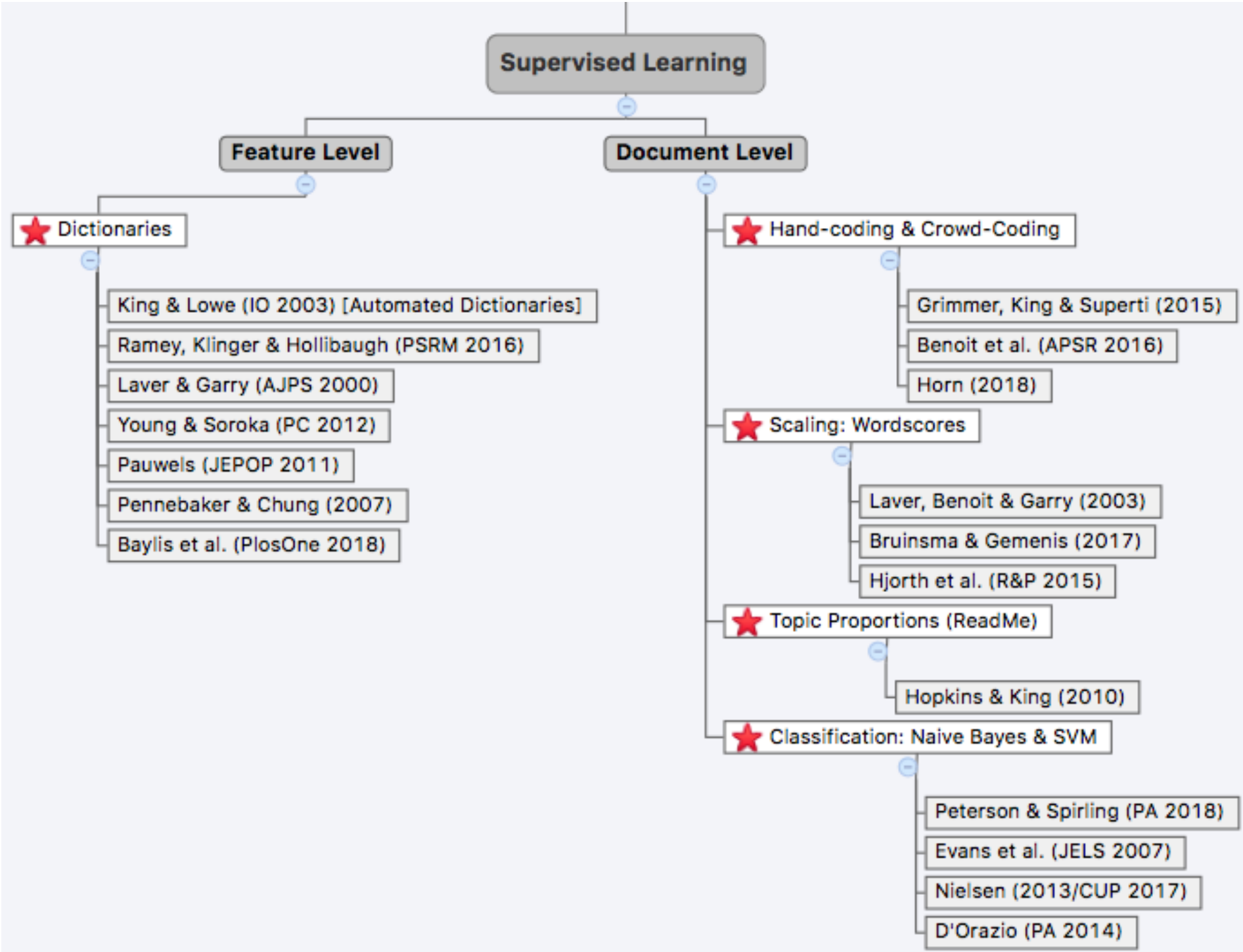

Quantitative Text Analysis
Descriptive Inference
Pre-Processing
Denny & Spirling (PA 2018)
Catalinac (Polimetrics 2014)
Lexical Features (readibility, complexity, diversity)
Spirling (JOP 2015)
Ban et al (PSRM 2017)
Benoit, Munger & Spirling (2017)
Hengel (2017)
Vector Space Model
Turney & Pantel (JAIR 2010)
Dissimilarities & Word Discrimmination (Mutual Information)
Taddy (JASA 2013)
Monroe, Colaresi & Quinn (PA 2008)
Mosteller & Wallace (JASA 1963)
Unsupervised Learning
Principal Component Analysis (Single Membership)
Beauchamp (2010)
Spirling (AJPS 2011)
Clustering (Single Membership) & Mixture Unigram Models
Grimmer & King (PNAS 2011)
Beil, Esther & Xu (2002)
Klüver & Mahoney (JPP 2015)
(Structural) Topic Modelling (Mixed Membership)
Chang et al. (NIPS 2009)
Vries, Schoonvelde, Schumacher (2017)
Catalinac (2016)
Quinn et al. (AJPS 2010)
Roberts et al. (AJPS 2014)
Roberts, Stewart and Tingley (2016)
Cerchiello & Nicola (Econometrics 2018)
Barron et al. (PNAS 2018)
Munger et al. (PSRM 2018)
Scaling IRT: Wordfish & Wordshoal
Slapin & Proksch (AJPS 2008)
Catalinac (APSR 2017)
Lauderdale & Herzog (PA 2016)
Supervised Learning
Feature Level
Dictionaries
King & Lowe (IO 2003) [Automated Dictionaries]
Ramey, Klinger & Hollibaugh (PSRM 2016)
Laver & Garry (AJPS 2000)
Young & Soroka (PC 2012)
Pauwels (JEPOP 2011)
Pennebaker & Chung (2007)
Baylis et al. (PlosOne 2018)
Document Level
Hand-coding & Crowd-Coding
Grimmer, King & Superti (2015)
Benoit et al. (APSR 2016)
Horn (2018)
Scaling: Wordscores
Laver, Benoit & Garry (2003)
Bruinsma & Gemenis (2017)
Hjorth et al. (R&P 2015)
Topic Proportions (ReadMe)
Hopkins & King (2010)
Classification: Naive Bayes & SVM
Peterson & Spirling (PA 2018)
Evans et al. (JELS 2007)
Nielsen (2013/CUP 2017)
D'Orazio (PA 2014)


Descriptive Inference
Pre-Processing
Denny & Spirling (PA 2018)
Catalinac (Polimetrics 2014)
Lexical Features (readibility, complexity, diversity)
Spirling (JOP 2015)
Ban et al (PSRM 2017)
Benoit, Munger & Spirling (2017)
Hengel (2017)
Vector Space Model
Turney & Pantel (JAIR 2010)
Dissimilarities & Word Discrimmination (Mutual Information)
Taddy (JASA 2013)
Monroe, Colaresi & Quinn (PA 2008)
Mosteller & Wallace (JASA 1963)


Quantitative Text Analysis
Descriptive Inference
Pre-Processing
Denny & Spirling (PA 2018)
Catalinac (Polimetrics 2014)
Lexical Features (readibility, complexity, diversity)
Spirling (JOP 2015)
Ban et al (PSRM 2017)
Benoit, Munger & Spirling (2017)
Hengel (2017)
Vector Space Model
Turney & Pantel (JAIR 2010)
Dissimilarities & Word Discrimmination (Mutual Information)
Taddy (JASA 2013)
Monroe, Colaresi & Quinn (PA 2008)
Mosteller & Wallace (JASA 1963)
Unsupervised Learning
Principal Component Analysis (Single Membership)
Beauchamp (2010)
Spirling (AJPS 2011)
Clustering (Single Membership) & Mixture Unigram Models
Grimmer & King (PNAS 2011)
Beil, Esther & Xu (2002)
Klüver & Mahoney (JPP 2015)
(Structural) Topic Modelling (Mixed Membership)
Chang et al. (NIPS 2009)
Vries, Schoonvelde, Schumacher (2017)
Catalinac (2016)
Quinn et al. (AJPS 2010)
Roberts et al. (AJPS 2014)
Roberts, Stewart and Tingley (2016)
Cerchiello & Nicola (Econometrics 2018)
Barron et al. (PNAS 2018)
Munger et al. (PSRM 2018)
Scaling IRT: Wordfish & Wordshoal
Slapin & Proksch (AJPS 2008)
Catalinac (APSR 2017)
Lauderdale & Herzog (PA 2016)
Supervised Learning
Feature Level
Dictionaries
King & Lowe (IO 2003) [Automated Dictionaries]
Ramey, Klinger & Hollibaugh (PSRM 2016)
Laver & Garry (AJPS 2000)
Young & Soroka (PC 2012)
Pauwels (JEPOP 2011)
Pennebaker & Chung (2007)
Baylis et al. (PlosOne 2018)
Document Level
Hand-coding & Crowd-Coding
Grimmer, King & Superti (2015)
Benoit et al. (APSR 2016)
Horn (2018)
Scaling: Wordscores
Laver, Benoit & Garry (2003)
Bruinsma & Gemenis (2017)
Hjorth et al. (R&P 2015)
Topic Proportions (ReadMe)
Hopkins & King (2010)
Classification: Naive Bayes & SVM
Peterson & Spirling (PA 2018)
Evans et al. (JELS 2007)
Nielsen (2013/CUP 2017)
D'Orazio (PA 2014)


Unsupervised Learning
Principal Component Analysis (Single Membership)
Beauchamp (2010)
Spirling (AJPS 2011)
Clustering (Single Membership) & Mixture Unigram Models
Grimmer & King (PNAS 2011)
Beil, Esther & Xu (2002)
Klüver & Mahoney (JPP 2015)
(Structural) Topic Modelling (Mixed Membership)
Chang et al. (NIPS 2009)
Vries, Schoonvelde, Schumacher (2017)
Catalinac (2016)
Quinn et al. (AJPS 2010)
Roberts et al. (AJPS 2014)
Roberts, Stewart and Tingley (2016)
Cerchiello & Nicola (Econometrics 2018)
Barron et al. (PNAS 2018)
Munger et al. (PSRM 2018)
Scaling IRT: Wordfish & Wordshoal
Slapin & Proksch (AJPS 2008)
Catalinac (APSR 2017)
Lauderdale & Herzog (PA 2016)


Quantitative Text Analysis
Descriptive Inference
Pre-Processing
Denny & Spirling (PA 2018)
Catalinac (Polimetrics 2014)
Lexical Features (readibility, complexity, diversity)
Spirling (JOP 2015)
Ban et al (PSRM 2017)
Benoit, Munger & Spirling (2017)
Hengel (2017)
Vector Space Model
Turney & Pantel (JAIR 2010)
Dissimilarities & Word Discrimmination (Mutual Information)
Taddy (JASA 2013)
Monroe, Colaresi & Quinn (PA 2008)
Mosteller & Wallace (JASA 1963)
Unsupervised Learning
Principal Component Analysis (Single Membership)
Beauchamp (2010)
Spirling (AJPS 2011)
Clustering (Single Membership) & Mixture Unigram Models
Grimmer & King (PNAS 2011)
Beil, Esther & Xu (2002)
Klüver & Mahoney (JPP 2015)
(Structural) Topic Modelling (Mixed Membership)
Chang et al. (NIPS 2009)
Vries, Schoonvelde, Schumacher (2017)
Catalinac (2016)
Quinn et al. (AJPS 2010)
Roberts et al. (AJPS 2014)
Roberts, Stewart and Tingley (2016)
Cerchiello & Nicola (Econometrics 2018)
Barron et al. (PNAS 2018)
Munger et al. (PSRM 2018)
Scaling IRT: Wordfish & Wordshoal
Slapin & Proksch (AJPS 2008)
Catalinac (APSR 2017)
Lauderdale & Herzog (PA 2016)
Supervised Learning
Feature Level
Dictionaries
King & Lowe (IO 2003) [Automated Dictionaries]
Ramey, Klinger & Hollibaugh (PSRM 2016)
Laver & Garry (AJPS 2000)
Young & Soroka (PC 2012)
Pauwels (JEPOP 2011)
Pennebaker & Chung (2007)
Baylis et al. (PlosOne 2018)
Document Level
Hand-coding & Crowd-Coding
Grimmer, King & Superti (2015)
Benoit et al. (APSR 2016)
Horn (2018)
Scaling: Wordscores
Laver, Benoit & Garry (2003)
Bruinsma & Gemenis (2017)
Hjorth et al. (R&P 2015)
Topic Proportions (ReadMe)
Hopkins & King (2010)
Classification: Naive Bayes & SVM
Peterson & Spirling (PA 2018)
Evans et al. (JELS 2007)
Nielsen (2013/CUP 2017)
D'Orazio (PA 2014)


Supervised Learning
Feature Level
Document Level
Dictionaries
King & Lowe (IO 2003) [Automated Dictionaries]
Ramey, Klinger & Hollibaugh (PSRM 2016)
Laver & Garry (AJPS 2000)
Young & Soroka (PC 2012)
Pauwels (JEPOP 2011)
Pennebaker & Chung (2007)
Baylis et al. (PlosOne 2018)
Hand-coding & Crowd-Coding
Grimmer, King & Superti (2015)
Benoit et al. (APSR 2016)
Horn (2018)
Scaling: Wordscores
Laver, Benoit & Garry (2003)
Bruinsma & Gemenis (2017)
Hjorth et al. (R&P 2015)
Topic Proportions (ReadMe)
Hopkins & King (2010)
Classification: Naive Bayes & SVM
Peterson & Spirling (PA 2018)
Evans et al. (JELS 2007)
Nielsen (2013/CUP 2017)
D'Orazio (PA 2014)

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is: need to make qualitative judgements about what the text reveals, and validation requires substantive knowledge
- 'Distant reading' instead of 'close reading'—not focussed on interpretation in light of norms or belief systems.
- Important: quantitative work is reliable and replicable (easily) and can cope with large volume of material.

Goal of Text Analysis

In many (most?) social science applications of text as data, we are trying to make an inference about a *latent variable*.

→ something which we cannot observe directly but which we can make inferences about from things we can observe. Examples include ideology, ambition, narcissism, propensity to vote etc.

In traditional social science research, we might observe roll call votes, donation decisions, responses to survey questions, etc.

Here, the thing we can observe are the words spoken, the passages written, the issues debated or whatever.

And. . .

- the latent variable of interest may pertain to the. . .

author ‘what does this Senator prioritize?’, ‘where is this party in ideological space?’

doc ‘does this treaty represent a fair deal for American Indians?’, ‘how did the discussion of lasers change over time?’

both ‘how does the way Japanese politicians talk about national defence change in response to electoral system shift?’

What Can Text Methods Do?

Haystack metaphor:

What Can Text Methods Do?

Haystack metaphor: Improve Reading

What Can Text Methods Do?

Haystack metaphor: Improve Reading

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

What Can Text Methods Do?

Haystack metaphor: Improve Reading

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

What Can Text Methods Do?

Haystack metaphor: Improve Reading

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

What Can Text Methods Do?

Haystack metaphor: Improve Reading

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

What Can Text Methods Do?

Haystack metaphor: Improve Reading

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

What automated text methods don't do:

What Can Text Methods Do?

Haystack metaphor: Improve Reading

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

What automated text methods don't do:

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

Texts are Deceptively Complex

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

Texts are Deceptively Complex

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

- Who is the I ?

Texts are Deceptively Complex

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

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

Texts are Deceptively Complex

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

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

Texts are Deceptively Complex

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

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

Texts$\rightsquigarrow$ high dimensional, not self contained

Texts are Surprisingly Simple

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

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

Texts are Surprisingly Simple (?)

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

Not just for “big data”

Not just for “big data”

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Not just for "big data"

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

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

Automated methods can help with even small problems

Why text? Why not text?

Why text? Why not text?

- Text data is bad quantitative data: if what we care about is not the text but the latent concept, the signal to noise ratio is seldom good! The text to a reader may scream out the latent concept, but there is still a lot of noise in the text;

Why text? Why not text?

- Text data is bad quantitative data: if what we care about is not the text but the latent concept, the signal to noise ratio is seldom good! The text to a reader may scream out the latent concept, but there is still a lot of noise in the text;
- But there is a lot of it (plenty of useful information in text if we can find it);

Why text? Why not text?

- the difficulty is selecting throwing away the right information. Three filters:
  1. Corpus construction: what are the relevant texts? This choice is particularly important for unsupervised methods because they aim to explain variation in the corpus;
  2. Feature selection: Bag of words / n/skip-grams / collocations / word2vec / other representations; keep everything or use a dictionary?
  3. Modelling feature generation: model things as continuous dimensions, simplexes, single membership? this choice is less important than people often think; we have control about how we describe variation in the data, but we've already determined the r x c of the matrix; whatever summary of the matrix we choose (modelling selection), they'll typically give us similar things.

Why text? Why not text?

- the difficulty is selecting throwing away the right information. Three filters:
  1. Corpus construction: what are the relevant texts? This choice is particularly important for unsupervised methods because they aim to explain variation in the corpus;
  2. Feature selection: Bag of words / n/skip-grams / collocations / word2vec / other representations; keep everything or use a dictionary?
  3. Modelling feature generation: model things as continuous dimensions, simplexes, single membership? this choice is less important than people often think; we have control about how we describe variation in the data, but we've already determined the r x c of the matrix; whatever summary of the matrix we choose (modelling selection), they'll typically give us similar things.
- At the end of the day we can measure some things, somewhat reliably.

Four Principles of Automated Text Analysis

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

Four Principles of Automated Text Analysis

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

- Data generation process for text$\rightsquigarrow$ unknown

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

- Computer-Assisted Reading

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

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

Four Principles of Automated Text Analysis

Principle 2: Quantitative Methods Augment Humans, Not Replace Them

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

Four Principles of Automated Text Analysis

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

Four Principles of Automated Text Analysis

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

- Supervised methods$\rightsquigarrow$ known categories

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

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

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

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

- Quantitative methods$\rightsquigarrow$ variable performance across tasks

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

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

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

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

Four Principles of Automated Text Analysis

Principle 4: Validate, Validate, Validate

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

We need to think carefully about...

- the appropriate population and sample

→ document selection, stochastic view of text

- what we actually care about in the observed data, how to get at it, how to characterize it.

→ feature selection, feature representation, description

- exactly how to aggregate/mine/model the observed data—the texts with their relevant features measured/coded—that we have.

→ statistical choices

- what we can infer about the latent variables.

→ comparing, testing, validating.

Elements of textual data

Key concepts

Elements of textual data

Key concepts

- (text) corpus a large and structured set of texts for analysis

Elements of textual data

Key concepts

- (text) corpus a large and structured set of texts for analysis
- types for our purposes, a unique word

Elements of textual data

Key concepts

- (text) corpus a large and structured set of texts for analysis
- types for our purposes, a unique word
- tokens any word – so token count is total words

Elements of textual data

Key concepts

- (text) corpus a large and structured set of texts for analysis
- types for our purposes, a unique word
- tokens any word – so token count is total words
- stems words with suffixes removed

Elements of textual data

Key concepts

- (text) corpus a large and structured set of texts for analysis
- types for our purposes, a unique word
- tokens any word – so token count is total words
- stems words with suffixes removed
- lemmas canonical word form (the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached)

Defining "documents"

selecting units of textual analysis

- words
- n-word sequences
- pages
- paragraphs
- themes
- natural units (speech, poem, manifesto)
- key: depends on the research design;

In general, we will...

Get Texts

> An expert hospital consultant has written to my hon. Friend…

> Order. The Minister must be allowed to reply without interruption.

> I am grateful to my hon. Friend for her question. I pay tribute to her work with the International Myeloma Foundation…

> My constituent, Brian Jago, was fortunate enough to receive a course of Velcade, as a result of which he does not have to…

→ Document Term Matrix

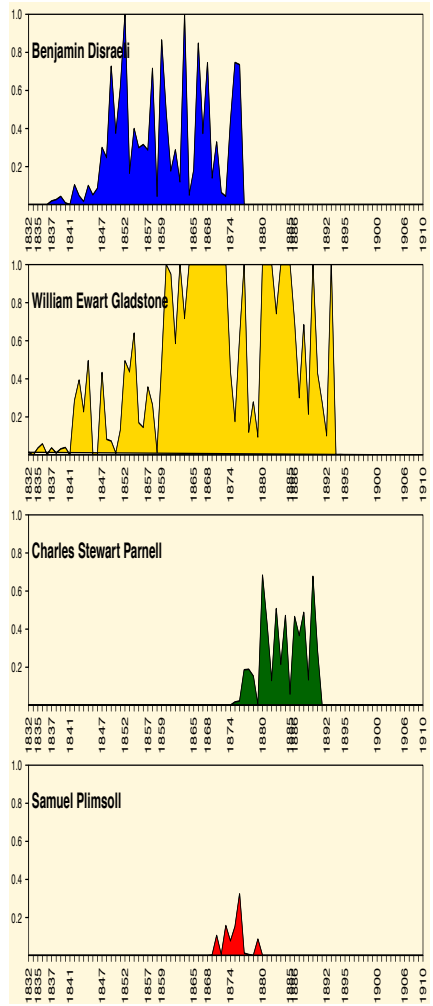
$$
\begin{array}{c} \\ MP_{001} \\ MP_{002} \\ \\ MP_{i} \\ \\ MP_{654} \\ MP_{655} \end{array}
\begin{array}{c}
\begin{array}{cccc} a & an & \dots & ze \end{array} \\
\begin{pmatrix}
2 & 0 & \dots & 1 \\
0 & 3 & \dots & 0 \\
\vdots & \vdots & \dots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 2
\end{pmatrix}
\end{array}
$$

→ Operate

- (dis)similarity
- diversity
- readability
- scale
- classify
- topic model
- burstiness
- sentiment

...

→ Inference



I. Defining the Corpus

defn (typically) large set of texts or documents which we wish to analyze.

→ how large? if small enough to read in reasonable time, you should probably just do that.

'structured', in the sense that you know what the documents are, where they begin and end, who authored them etc.

'unstructured data' in sense that what is wanted (e.g. ideological position) may not be directly observable.

may be annotated in sense that metadata —data that is not part of the document itself—is available. Examples include markup, authorship and date information, linguistic tagging (more below)

e.g. court transcripts, legislative records, Twitter feeds, Brown Corpus etc.

Sampling

The corpus is made up of the documents within it, but these may be a sample of the total population of documents available.

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

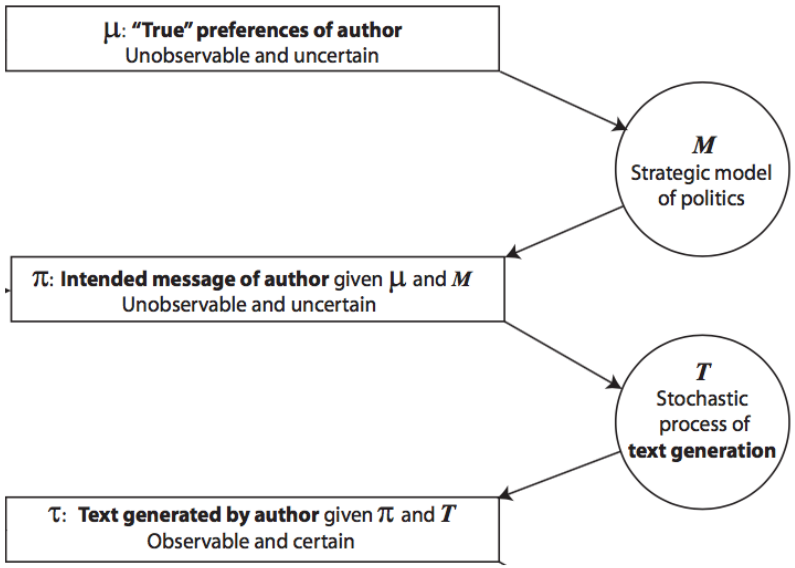
Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

$\rightarrow$ depending on your philosophical position, you still need to think about sampling error. This is because there exists a superpopulation of populations from which the universe you observed came from.

Random error may not be the only concern: corpus should be representative in some well defined sense for inferences to be meaningful.

Sample v. "population"

- Basic Idea: Observed text is a stochastic realization
- Systematic features shape most of observed verbal content
- Non-systematic, random features also shape verbal content



Implications of a stochastic view of text

Implications of a stochastic view of text

- Observed text is not the only text that could have been generated

Implications of a stochastic view of text

- Observed text is not the only text that could have been generated
- Very different if you are trying to monitor something like hate speech, where what you actually say matters, not the value of your "expected statement"

Implications of a stochastic view of text

- Observed text is not the only text that could have been generated
- Very different if you are trying to monitor something like hate speech, where what you actually say matters, not the value of your "expected statement"
- Means that having "all the text" is still not a "population"

Sampling strategies for selecting texts

Sampling strategies for selecting texts

- Difference between a sample and a population

Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling

Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling
- May not be necessary to perform any sampling

Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling
- May not be necessary to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"

Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling
- May not be necessary to perform any sampling
- Be wary of sampling that is a feature of the social system: "social bookkeeping"
- Different types of sampling vary from random to purposive: random sampling; non-random sampling

Sampling biases

Sampling biases

- Resource bias: online data (what pop excluded?), historical data (press presence bias?), archives (texts losts, not stored?), government data (transparency?);

Sampling biases

- Resource bias: online data (what pop excluded?), historical data (press presence bias?), archives (texts losts, not stored?), government data (transparency?);
- Incentive bias: records of negative v positive information equally likely? online posts reflection of most (un)successful moments? political censorship (e.g. Spanish Government's censoring public broadcaster TVE's coverage of government officials – both ommission & manipulation of news?

Sampling biases

- Resource bias: online data (what pop excluded?), historical data (press presence bias?), archives (texts losts, not stored?), government data (transparency?);
- Incentive bias: records of negative v positive information equally likely? online posts reflection of most (un)successful moments? political censorship (e.g. Spanish Government's censoring public broadcaster TVE's coverage of government officials – both ommission & manipulation of news?
- Medium bias: text is constrained by platform (twitter 140 characters forces abbreviation – not for Chinese users; increasing usage of metadata – emoticons);

Sampling biases

- Resource bias: online data (what pop excluded?), historical data (press presence bias?), archives (texts losts, not stored?), government data (transparency?);
- Incentive bias: records of negative v positive information equally likely? online posts reflection of most (un)successful moments? political censorship (e.g. Spanish Government's censoring public broadcaster TVE's coverage of government officials – both ommission & manipulation of news?
- Medium bias: text is constrained by platform (twitter 140 characters forces abbreviation – not for Chinese users; increasing usage of metadata – emoticons);
- Algorithm bias: how would you select ghost stories from the library?

Document-Term Matrices

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & \dots & 3 \\ 0 & 2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 5 \end{pmatrix}$$

$\boldsymbol{X} = N \times K$ matrix

Document-Term Matrices

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & \dots & 3 \\ 0 & 2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 5 \end{pmatrix}$$

$\boldsymbol{X} = N \times K$ matrix

- $N =$ Number of documents

Document-Term Matrices

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & \dots & 3 \\ 0 & 2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 5 \end{pmatrix}$$

$\boldsymbol{X} = N \times K$ matrix

- $N =$ Number of documents
- $K =$ Number of features

Document Term Matrices

Regular expressions and search are useful

Document Term Matrices

Regular expressions and search are useful
We want to use statistics/algorithms to characterize text

Document Term Matrices

Regular expressions and search are useful
We want to use statistics/algorithms to characterize text
We'll put it in a document-term matrix

Document Term Matrices

Preprocessing$\rightsquigarrow$Simplify text, make it useful

Document Term Matrices

Preprocessing$\rightsquigarrow$Simplify text, make it useful

Lower dimensionality

Document Term Matrices

Preprocessing$\rightsquigarrow$Simplify text, make it useful

Lower dimensionality

- For our purposes

Document Term Matrices

Preprocessing$\rightsquigarrow$Simplify text, make it useful
Lower dimensionality

- For our purposes

Remember: characterize the Hay stack

Document Term Matrices

Preprocessing$\rightsquigarrow$Simplify text, make it useful
Lower dimensionality

- For our purposes

Remember: characterize the Hay stack

- If you want to analyze a straw of hay, these methods are unlikely to work

Document Term Matrices

Preprocessing$\rightsquigarrow$Simplify text, make it useful
Lower dimensionality

- For our purposes

Remember: characterize the Hay stack

- If you want to analyze a straw of hay, these methods are unlikely to work
- But even if you want to closely read texts, characterizing hay stack can be useful

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words
4) Create Equivalence Class: Stem, Lemmatize, or synonym

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words
4) Create Equivalence Class: Stem, Lemmatize, or synonym
5) Discard less useful features$\rightsquigarrow$ depends on application

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words
4) Create Equivalence Class: Stem, Lemmatize, or synonym
5) Discard less useful features$\rightsquigarrow$ depends on application
6) Other reduction, specialization

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words
4) Create Equivalence Class: Stem, Lemmatize, or synonym
5) Discard less useful features$\rightsquigarrow$ depends on application
6) Other reduction, specialization

Output: Count vector, each element counts occurrence of stems

Preprocessing for Quantitative Text Analysis

One (of many) recipe for preprocessing: retain useful information

1) Remove capitalization, punctuation
2) Discard Word Order (Bag of Words Assumption)
3) Discard stop words
4) Create Equivalence Class: Stem, Lemmatize, or synonym
5) Discard less useful features$\rightsquigarrow$ depends on application
6) Other reduction, specialization

Output: Count vector, each element counts occurrence of stems

Provide tools to preprocess via this recipe

The Bag of Words Assumption

Assumption: Discard Word Order

```
Now we are engaged in a great civil war, testing whether
that nation, or any nation
```

The Bag of Words Assumption

Assumption: Discard Word Order

```
now we are engaged in a great civil war testing whether
that nation or any nation
```

The Bag of Words Assumption

Assumption: Discard Word Order

Unigrams

Unigram	Count
a	1
any	1
are	1
civil	1
engaged	1
great	1
in	1
nation	2
now	1
or	1
testing	1
that	1
war	1
we	1
whether	1

The Bag of Words Assumption

Assumption: Discard Word Order

Bigrams

Bigram	Count
now we	1
we are	1
are engaged	1
engaged in	1
in a	1
a great	1
great civil	1
civil war	1
war testing	1
testing whether	1
whether that	1
that nation	1
nation or	1
or any	1
any nation	1

The Bag of Words Assumption

Assumption: Discard Word Order

Trigrams

Trigram	Count
now we are	1
we are engaged	1
are engaged in	1
engaged in a	1
in a great	1
a great civil	1
great civil war	1
civil war testing	1
war testing whether	1
whether that nation	1
that nation or	1
nation or any	1
or any nation	1

How Could This Possibly Work?

Speech is:

- Ironic

  `The Raiders make very good personnel decisions`

- Subtle Negation (Source: Janyce Wiebe) :

  `They have not succeeded, and will never succeed, in breaking the will of this valiant people`

- Order Dependent (Source: Arthur Spirling):

  `Peace, no more war`

  `War, no more peace`

How Could This Possibly Work?

Three answers

1) It might not: Validation is critical (task specific)
2) Central Tendency in Text: Words often imply what a text is about `war, civil, union` or tone `consecrate, dead, died, lives`.

   Likely to be used repeatedly: create a theme for an article
3) Human supervision: Inject human judgement (coders): helps methods identify subtle relationships between words and outcomes of interest

   Dictionaries

   Training Sets

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.

2. strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.
4. add descriptive annotations that preserve context: tagging.
5. map tokens back to common form: lemmatization, stemming.

6. operate/model.

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.

“PREPROCESSING”

6. operate/model.

Well. . .

what to do depends on what language features you are most interested in.

if the grammatical structure of sentences matters, makes sense to keep most, if not all, punctuation.

e.g. social media: does use of `!` differ by age group?

but mostly just interested in coarse features (such as word frequencies); converting most punctuation to whitespace is quick and better than keeping it.

NB ‘dictionaries’ can be used to map contractions back to their component parts

e.g. tell us that `won't` could be `will not`

but may not be as important as you think.

‘superfluous’ material: capitalization

Federalist 1

```
The subject speaks its own importance; comprehending in its
consequences nothing less than the existence of the UNION, the
safety and welfare of the parts of which it is composed, the fate
of an empire in many respects the most interesting in the world.
```

is the one use of ‘The’ the same word as the seven uses of ‘the’?

is ‘UNION’ the same word as ‘union’ and ‘Union’ as used elsewhere in this essay?

yes → lowercase (uppercase) everything

or keep lists (dictionary) of proper nouns, lowercase everything else

or lowercase words at the beginning of a sentence (how do we know where a sentence begins?) leave everything else as is

Quick Note on Terminology

a type is a unique sequence of characters that are grouped together in some meaningful way. Mostly a word (for us), but might also be a word plus punctuation, or a number etc.

e.g. 'France', 'American Revolution', '1981'

a token is a particular *instance* of type.

e.g. "Dog eat dog world", contains three types, but four tokens (for most purposes).

a term is a type that is part of the system's 'dictionary' (i.e. what the quantitative analysis technique recognizes as a type to be recorded etc). Could be different from the tokens, but often closely related.

e.g. stemmed word like 'treasuri', which doesn't appear in the document itself.

Tokens and tokenization

The text is now 'clean', and we want to pull out the meaningful subunits—the tokens. We will use a tokenizer.

→ usually the tokens are words, but might include numbers or punctuation too.

Common rule for a tokenizer is to use whitespace as the marker.

but given application might require something more subtle

e.g. "`Brown vs Board of Education`" may not be usefully tokenized as '`Brown`', '`vs`', '`Board`', '`of`', '`Education`'

Exceptions and Other Ideas

In some languages, tokenizing is a non-trivial problem because whitespace may not be used:

问世间情是何物，直教生死相许。
天南地北双飞客，老翅几回寒暑。

We may want to deal directly with multiword expressions in some contexts.There are rules which help us identify them relatively quickly and accurately.

e.g. ‘`White House`’ , ‘`traffic light`’

NB these words mean something ‘special’ (and slightly opaque) when combined. Related to idea of collocations: words that appear together more often than we’d predict based on random sampling.

Removing Stop Words

There are certain words that serve as linguistic connectors ('function words') which we can remove.

→ this simplifies our document considerably, with little loss of substantive 'content'. Indeed, search engines often ignore them.

There are many lists available, and we may add to them in an application specific way.

e.g. working with Congressional speech data, '`representative`' might be a stop word; in *Hansard* data, '`honourable`' might be.

NB in some specific applications, function word usage **is** important—we'll discuss this when we deal with authorship attribution.

Some stop words

a	about	above	after	again	against	all
am	an	and	any	are	aren't	as
at	be	because	been	before	being	below
between	both	but	by	can't	cannot	could
couldn't	did	didn't	do	does	doesn't	doing
don't	down	during	each	few	for	from
further	had	hadn't	has	hasn't	have	haven't
having	he	he'd	he'll	he's	her	here
here's	hers	herself	him	himself	his	how
how's	i	i'd	i'll	i'm	i've	if
in	into	is	isn't	it	it's	its
itself	let's	me	more	most	mustn't	my
myself	no	nor	not	of	off	on
once	only	or	other	ought	our	ours
ourselves	out	over	own	same	shan't	she
she'd	she'll	she's	should	shouldn't	so	some
such	than	that	that's	the	their	theirs
them	themselves	then	there	there's	these	they
they'd	they'll	they're	they've	this	those	through
to	too	under	until	up	very	was
wasn't	we	we'd	we'll	we're	we've	were
weren't	what	what's	when	when's	where	where's
which	while	who	who's	whom	why	why's
with	won't	would	wouldn't	you	you'd	you'll
you're	you've	your	yours	yourself	yourselves	

Tagging

so far tokens are on even footing—no distinctions drawn between nouns, verbs, nouns acting as subjects, nouns acting as objects, etc.

and for many applications, this information doesn't help very much (e.g. for classification).

but in other applications we may really want to know information about the part-of-speech this word represents. We want to disambiguate in what sense a term is being used.

e.g. in 'events' studies, when we are recording who did what to whom: '`the UK bombing will force ISIS to surrender`'. Here `force` is a verb, not a noun.

→ annotating in this way is called parts-of-speech tagging.

Penn POS Tagger

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Stemming and Lemmatization

Documents may use different forms of words ('`jumped`', '`jumping`', '`jump`'), or words which are similar in concept ('`bureaucratic`', '`bureaucrat`', '`bureaucratization`') as if they are different tokens.

→ we can simplify considerably by mapping these variants (back) to the same word.

- Stemming does this using a crude (heuristic) which just 'chops off' the affixes. It returns a stem which might not be a dictionary word.
- Lemmatization does this using a vocabulary, parts of speech context and mapping rules. It returns a word in the dictionary: a lemma (which is a canonical form of a 'lexeme').

e.g. depending on context, lemmatization would return '`see`' or '`saw`' if it came across '`saw`'.

Stemming

Though technically incorrect, ‘stemming’ and ‘lemmatization’ often used interchangeably.

For small examples, one can use a ‘look up’ table: table listing what a given realization of a word should be mapped to.

btw we sometimes use ‘equivalency classes’ meaning that an internal thesaurus maps different words back to the same type of word: e.g. ‘`rightwing`’ and ‘`republican`’ to ‘`conservative`’.

In practice, need something faster (and cruder), so software implements the Porter Stemmer using algorithms like `Snowball`.

Snowball examples

Original Word		Stemmed Word
abolish	$\mapsto$	`abolish`
abolished	$\mapsto$	`abolish`
abolishing	$\mapsto$	`abolish`
abolition	$\mapsto$	`abolit`
abortion	$\mapsto$	`abort`
abortions	$\mapsto$	`abort`
abortive	$\mapsto$	`abort`
treasure	$\mapsto$	`treasure`
treasured	$\mapsto$	`treasure`
treasures	$\mapsto$	`treasure`
treasuring	$\mapsto$	`treasure`
treasury	$\mapsto$	`treasuri`

NYT

Emergency measures adopted for Beijing's first ‘‘red alert" over air pollution left millions of schoolchildren cooped up at home, forced motorists off the roads and shut down factories across the region on Tuesday, but they failed to dispel the toxic air that shrouded the Chinese capital in a soupy, metallic haze.

marked up

Emergenc y measur es adopt ed for Beij ing s first red alert over air pollut ion left million s of schoolchildren coop ed up at home, forc ed motorist s off the road s and shut down factor ies across the region on Tuesday, but they fail ed to dispel the toxic air that shroud ed the Chines e capit al in a soupy, metal lic haze.

NYT

Emergency measures adopted for Beijings first red alert over air pollution left millions of schoolchildren cooped up at home, forced motorists off the roads and shut down factories across the region on Tuesday, but they failed to dispel the toxic air that shrouded the Chinese capital in a soupy, metallic haze.

Stemmed

Emergenc measur adopt for Beij s first red alert over air pollut left million of schoolchildren coop up at home forc motorist off the road and shut down factori across the region on Tuesdai but thei fail to dispel the toxic air that shroud the Chines capit in a soupi metal haze.

We Don't Care about Word Order

We have now pre-processed our texts.

Generally, we are willing to ignore the order of the words in a document. This considerably simplifies things. And we do (almost) as well without that information as when we retain it.

NB we are treating a document as a bag-of-words (BOW).

btw, we keep multiplicity—i.e. multiple uses of same token

e.g. "`The leading Republican presidential candidate has said Muslims should be banned from entering the US.`"

→ "`lead republican presidenti candid said muslim ban enter us`"

= "`us lead said candid presidenti ban muslim republican enter`"

Could we retain Word Order?

for some applications, we might retaining word order is very important.

e.g. we have a large number of multiword expressions or named entities like '`Bill Gates`'

e.g. we think some important subtlety of expression is lost: negation perhaps—"`I want coffee, not tea`"might be interpreted very differently without word order.

→ can use *n*-grams, which are (sometimes contiguous) sequences of two (bigrams) or three (trigrams) tokens. This makes computations considerably more complex.

also can use *substrings* which are groups of *n* contiguous characters.

Using String Kernels instead...

1. `peace not war between`
2. `brothers not warfare now`
3. `be war not friendship`

documents are similar in word use terms...

Using String Kernels instead...

1. `peace not war between`
2. `brothers not warfare now`
3. `be war not friendship`

not w,

Using String Kernels instead...

1. peace |not w|ar between
2. brothers |not w|arfare now
3. be war not friendship

not w,

Using String Kernels instead...

1. peace n|ot wa|r between
2. brothers n|ot wa|rfare now
3. be war not friendship

ot wa,

Using String Kernels instead...

1. peace no|t war|between
2. brothers no|t war|fare now
3. be war not friendship

t war

original/some pre-processing

```
a military patrol boat rescued three of the kayakers on general carrera
lake and a helicopter lifted out the other three the chilean army said
```

bigrams

```
"a military" "military patrol" "patrol boat" "boat rescued" "rescued
three" "three of" "of the" "the kayakers" "kayakers on" "on general"
"general carrera" "carrera lake" "lake and" "and a" "a helicopter"
"helicopter lifted" "lifted out" "out the" "the other" "other three"
"three the" "the chilean" "chilean army" "army said"
```

trigrams

```
"a military patrol" "military patrol boat" "patrol boat rescued" "boat
rescued three" "rescued three of" "three of the" "of the kayakers" "the
kayakers on" "kayakers on general" "on general carrera" "general carrera
lake" "carrera lake and" "lake and a" "and a helicopter" "a helicopter
lifted" "helicopter lifted out" "lifted out the" "out the other" "the
other three" "other three the" "three the chilean" "the chilean army"
"chilean army said"
```

Denny & Spirling: Cautionary Tale

Political scientists often use text-as-data in an exploratory or unsupervised way. In that world, the metric isn't really 'prediction'. Yet most advice about pre-processing comes from the supervised literature.

so generally hope that our inferences are pretty much the same substantively, regardless of the (common) pre-processing steps we take.

Well is that true? Rarely (never) checked . . . and maybe not.

Denny & Spirling look at (Wordfish) scaling of four sets of UK election manifestos (1983, 1987, 1992, 1997).

Hmm...

If preprocessing makes no difference to 'results', it shouldn't matter which we do—punctuation, numbers, lowercase, stem, stops, infrequent terms, *n*-grams—in terms of manifesto estimated to be most left (or right).

P	N	L	M	S	I	G	Most Left	Most Right
T	T	T	T	T	T	T	Lab 1983	Con 1997
T	T	F	F	T	T	T	Lab 1983	Con 1983
F	T	F	F	F	T	T	Lab 1992	Con 1992
F	F	T	T	T	F	T	Lab 1997	Con 1987

$\rightarrow$ more variance than we would like!

Notation and Terminology

$d = 1, \ldots, D$ indexes documents in the corpus

$w = 1, \ldots, W$ indexes features found in documents

$\mathbf{y}_d \in \mathbb{R}^W$ is a representation of document d in a particular feature space

so each document is now a vector, with each entry representing the frequency of a particular token or feature...

$\rightarrow$ stacking those vectors on top of each other gives the document term matrix (DTM) or the document feature matrix (DFM).

$\rightarrow$ taking the transpose of the DTM gives the term document matrix (TDM) or feature document matrix (FDM).

partial DTM from Roosevelt's Inaugural Addresses

	features				
docs	american	expect	induct	presid	will
1933-Roosevelt	2	1	1	1	12
1937-Roosevelt	4	0	0	2	16
1941-Roosevelt	4	0	0	1	4
1945-Roosevelt	1	0	0	1	7

partial TDM from Roosevelt's Inaugural Addresses

	docs			
features	1933-Roosevelt	1937-Roosevelt	1941-Roosevelt	1945-Roosevelt
american	2	4	4	1
expect	1	0	0	0
induct	1	0	0	0
presid	1	2	1	1
will	12	16	4	7

IV. Weighting

To this point, we have been constructing the document vectors as counts. More formally, this is term frequency, since it simply records the number of occurrences of a given term.

but this implies that all words are of 'equal importance'. This is a problem in some domains

e.g. almost every article in political science will mention 'politics', but that suggests they are all more similar than they really are (and makes it hard to find 'different' ones).

so we may want to do something that throws certain feature relationships into starker relief.

along with term frequency, we may want to consider document frequency: the number of documents in which this word appears.

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w. When the word is common in the corpus, this will be a small number. When the word is rare, this will be a large number.

$tf_{dw} \cdot \ln \frac{|D|}{df_w}$, term frequency-inverse document frequency: tf-idf.

tf-idf

$tf_{dw} \cdot \ln \frac{|D|}{df_w}$, term frequency-inverse document frequency: tf-idf.

$\rightarrow$ when a word is common in a given document, but rare in the corpus as whole, this means tf is high and idf is high. So presence of that word is indicative of difference, and it is weighted up.

but if word is common in a given document, and common in the corpus, tf is high, but idf are low. So term is weighted down, and filtered out.

and very low for words occurring in every document: least discriminative words.

Example: FDR corpus

FDR used 'will' 12 times in his 1933 speech. So, tf=12.

and in his 4 speeches (our corpus), he used it (at least once) in *every* speech. So, $|D| = 4$ and $df = 4$

so the *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{4}{4}\right) = 0$

→ tf-idf=0 for 'will' in 1933.

but he used 'expect' once in 1933, and he didn't use it any other speech.

so *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{4}{1}\right) = 1.38$

→ tf-idf=1.38 for 'expect' in 1933.

→ 'expect' helps us discriminate better than 'will'.

Notes on a DTM

the way we construct the DTM—including order/nature of pre-processing—is application specific.

$\rightarrow$ in some cases, we won't need a DTM at all.

NB DTM tends to be sparse: contains lots of (mostly) zeros.

- partly a consequence of language itself: people say things in idiosyncratic ways.
- partly a consequence of reweighting: taking log(1).

in some applications, we might remove sparse terms—tokens that occur in very few docs.

NB there are efficient ways to store and manipulate sparse matrices.

"Time flies like an arrow. Fruit flies like a banana."

http://textanalysisonline.com/

Basic descriptive summaries of text

Three answers

- Readability statistics: Use a combination of syllables and sentence length to indicate "readability" in terms of complexity
- Vocabulary diversity: (At its simplest) involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens
- Word (relative) frequency
- Theme (relative) frequency
- Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Basic descriptive summaries of text

KWIC *Key words in context* refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

lime (14)

79[C.10] 4 /Which was builded of **lime** and sand;/Until they came to
247A.6 4 /That was well biggit with **lime** and stane.
303A.1 2 bower,/Well built wi **lime** and stane,/And Willie came
247A.9 2 /That was well biggit wi **lime** and stane,/Nor has he stoln
305A.2 1 a castell biggit with **lime** and stane,/O gin it stands not
305A.71 2 is my awin,/I biggit it wi **lime** and stane;/The Tinnies and
79[C.10] 6 /Which was builded with **lime** and stone.
305A.30 1 a prittie castell of **lime** and stone,/O gif it stands not
108.15 2 /W*hi*ch was made both of **lime** and stone,/Shee tooke him by
175A.33 2 castle then,/Was made of **lime** and stone;/The vttermost
178[H.2] 2 near by,/Well built with **lime** and stone;/There is a lady
178F.18 2 built with stone and **lime**!/But far mair pittie on Lady
178G.35 2 was biggit wi stane and **lime**!/But far mair pity o Lady
2D.16 1 big a cart o stane and **lime**,/Gar Robin Redbreast trail it

Key Words in Context

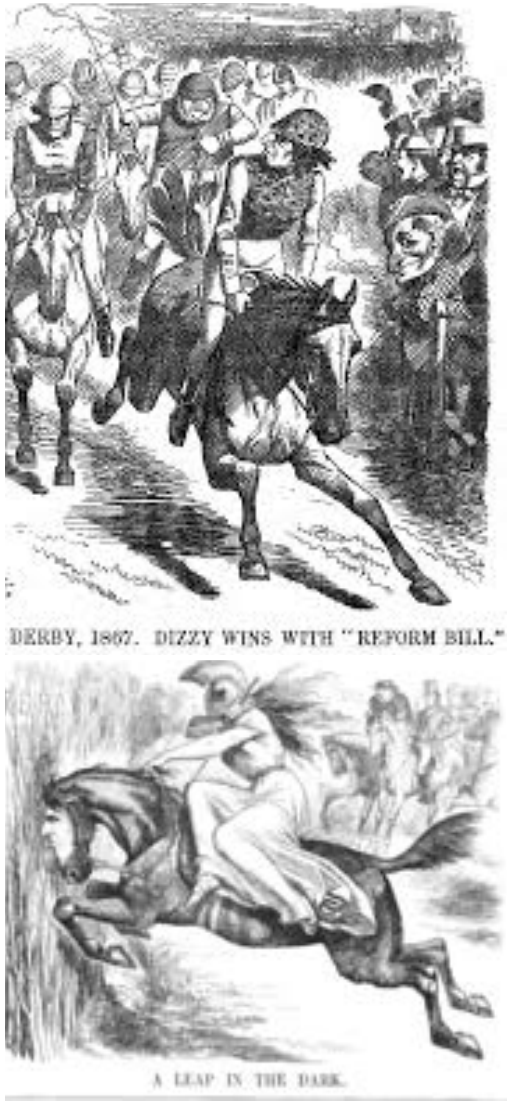
In Information Retrieval it is often extremely helpful to know how and where a particular token of interest appears, in terms of the words around it.

→ quick overview of general use, and allows for easy, follow up inspection of the document in question.

also true in social science applications where we might want to understand how a given concept appears, or when we are looking for prototypical examples.

1. keyword of interest.
2. context —typically the sentence in which it appears.
3. location code —document details.

Example: 'democratic' and the Second Reform Act



1867 House of Commons considers extending suffrage to urban working class men, via 'Representation of the People Act'

→ represents approximate doubling of electorate.

Debates of the time are lively and long. Normative notions of extending 'rights' on one hand (and pragmatic politics) vs fear of mob rule.

q What role did 'democratic' play in the debate?

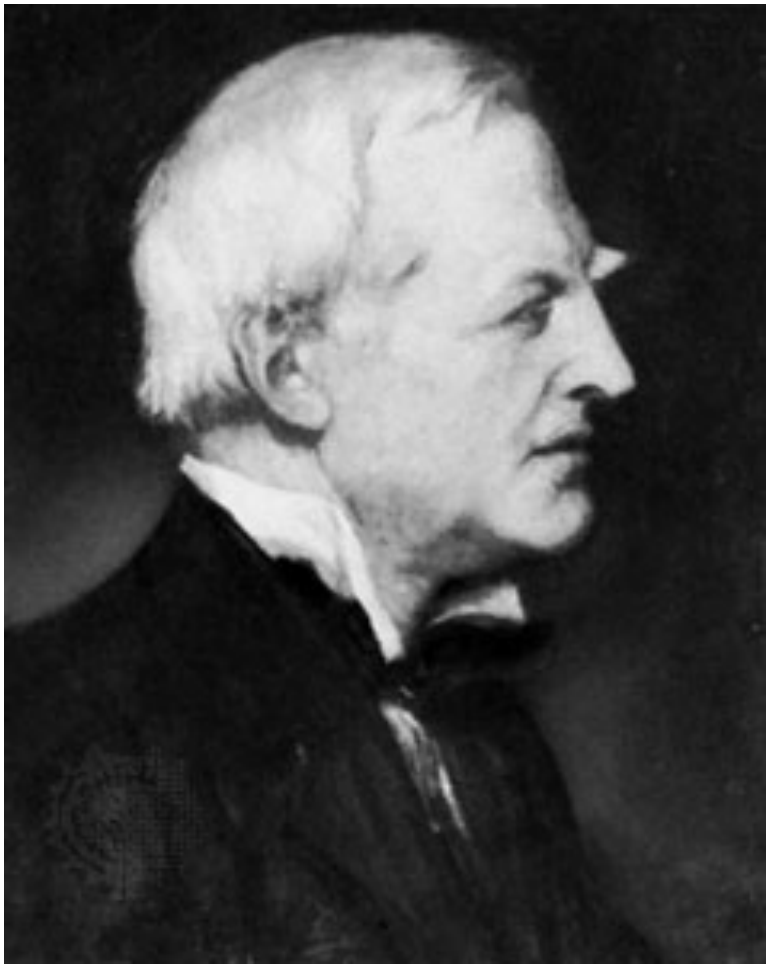
Some KWIC from the debates: `kwic()` in quanteda

	preword	word	postword
.	.	.	.
.	.	.	.
.	.	.	.
[s267549.txt, 994]	evil that attends a purely	democratic	form of Government. There could be
[s267549.txt, 1015]	here, not possibly towards a	democratic	form of government, but in
[s267738.txt, 1492]	swept away in some further	democratic	change. And it is for
[s267738.txt, 1560]	throne. When you get a	democratic	basis for your institutions, you
[s267738.txt, 1952]	differences between ourselves and other	democratic	legislatures? Where is the democratic
[s267738.txt, 1957]	democratic legislatures? Where is the	democratic	legislature which enjoys the powers
[s267738.txt, 2243]	almost utterly useless against a	democratic	Chamber, and the question to
[s267738.txt, 2286]	to the violence of the	democratic	Chamber you are creating, and,
[s267738.txt, 2294]	are creating, and, as the	democratic	principle brooks no rival, this
[s267738.txt, 2374]	spirit of democracy that the	democratic	Chamber itself would become an
[s267738.txt, 2678]	power is given to the	democratic	majority, that majority does not
[s267738.txt, 2767]	job? In accordance with the	democratic	principle the army would demand
[s267744.txt, 204]	Conservative patronage, of the most	democratic	Reform Bill ever brought in.

Detail: s267738.txt

preword	word	postword
swept away in some further	democratic	change. And it is for
throne. When you get a	democratic	basis for your institutions, you
differences between ourselves and other	democratic	legislatures? Where is the democratic
democratic legislatures? Where is the	democratic	legislature which enjoys the powers
almost utterly useless against a	democratic	Chamber, and the question to
to the violence of the	democratic	Chamber you are creating, and,
are creating, and, as the	democratic	principle brooks no rival, this
spirit of democracy that the	democratic	Chamber itself would become an
power is given to the	democratic	majority, that majority does not
job? In accordance with the	democratic	principle the army would demand

The Original Speaker and Speech



You cannot trust to a majority elected by men just above the status of paupers. The experiment has been tried; it has answered nowhere; it has failed in America, and it will not answer here.

In accordance with the democratic principle the army would demand to elect their own officers, and there would be endless change in the Constitution arising out of the present Bill, which, so far from being an end to our evils, is only the first step to them.

That was Robert Lowe, Viscount Sherbrooke, a British Liberal Party politician whose effective opposition to the Liberals electoral Reform Bill of 1866 made it possible for the Conservatives to sponsor and take credit for the Reform Act of 1867.

Lexical Diversity

Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.

The types in a document are the set of unique tokens.

thus we typically have many more tokens than types, because authors repeat tokens.

TTR we can use the type-to-token ratio as a measure of lexical diversity. This is:

$$\boxed{TTR = \frac{\text{total types}}{\text{total tokens}}}$$

e.g. authors with limited vocabularies will have a low lexical diversity.

Tabloid vs Broadsheet

NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated

Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images

MORE ON: ISIS

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

$$TTR = \frac{250}{491} = 0.51$$

The New York Times

Obama's 'Boots on the Ground': U.S. Special Forces Are Sent to Tackle Global Threats

Japan and South Korea Settle Dispute Over Wartime 'Comfort Women'

ITINERARIES

T.S.A. Moves Closer to Rejecting Some State Driver's Licenses for...

MIDDLE EAST

Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC. 28, 2015

Iraqi soldiers at the Anbar police headquarters in Ramadi, Iraq, on Monday, after seizing a government complex from the Islamic State. Ahmad Al-Rubaye/Agence France-Presse — Getty Images

Email

Share

BAGHDAD — Iraqi forces said on Monday they had seized a strategic government complex in the western city of Ramadi from the Islamic State after a fierce weeklong battle, putting them on the verge of a crucial victory following a brutal seven-month occupation of the city by the extremist group

$$TTR = \frac{428}{978} = 0.43$$

Hmm...

Unexpected, and mostly product of different text lengths: shorter texts tend to have fewer repetitions (of e.g. common words).

but also case that longer documents cover more topics which presumably adds to richness (?)

so make denominator non-linear:

1954 Guiraud index of lexical richness : $\boxed{R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}}$

so NY Post: $\frac{250}{\sqrt{491}} = 11.28$; NYT: $\frac{428}{\sqrt{978}} = 13.68$.

$\rightarrow$ has been augmented—Advanced Guiraud—to exclude very common words.

Partner Exercise

Restoration of national income, which shows continuing gains for the third successive year, supports the normal and logical policies under which agriculture and industry are returning to full activity. Under these policies we approach a balance of the national budget. National income increases; tax receipts, based on that income, increase without the levying of new taxes.

Some say my tax plan is too big. Others say its too small. I respectfully disagree.

Compare these two speech segments. Which is more difficult to understand? Why: which features are important?

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability': general issue was assigning school texts to pupils of different ages and abilities.
- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

$$= 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

based on $\hat{\beta}$s from linear model where y = average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

- Kincaid et al later translate to US School grade level that would be (on average) required to comprehend text.

Readability Guidelines

in practice, estimated FRE can be outside $[0, 100]$.

However. . .

Score	Education	Description	Clve % US popn
0–30	college graduates	very difficult	28
31–50		difficult	72
51–60		fairly difficult	85
61–70	9th grade	standard	–
71–80		fairly easy	–
81–90		easy	–
91–100	4th grade	very easy	–

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion
37	Spirling
45	life insurance requirement (FL)
48	*New York times*
65	*Reader's Digest*
67	*Al Qaeda* press release
77	Dickens' works
80	children's books: e.g. *Wind in the Willows*
90	death row inmate last statements (TX)
100	this entry right here.

Notes

Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.

e.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"

Widely used because it 'works', not because it is justified from first principles

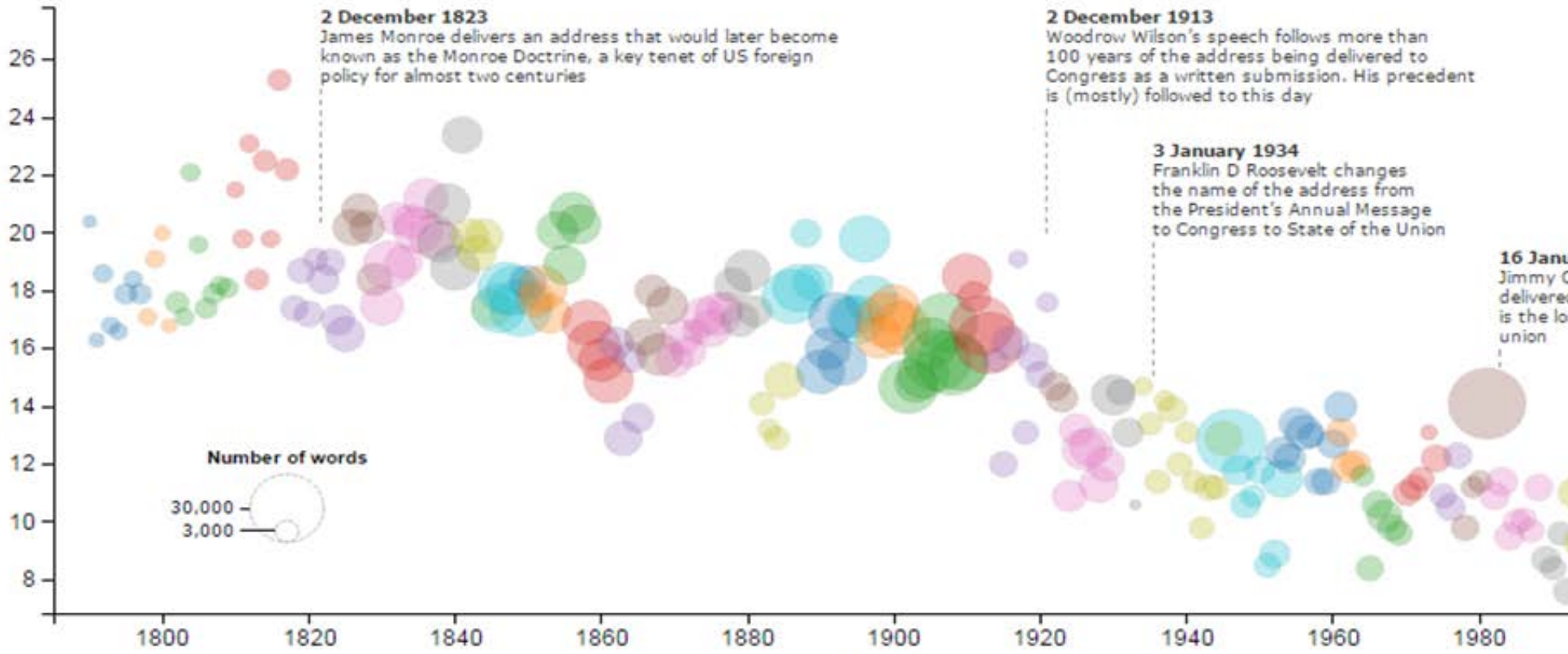
One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

Surprisingly little effort to describe statistical behavior of estimator: sampling distribution etc.

The state of our union is ... dumber:

How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union



Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

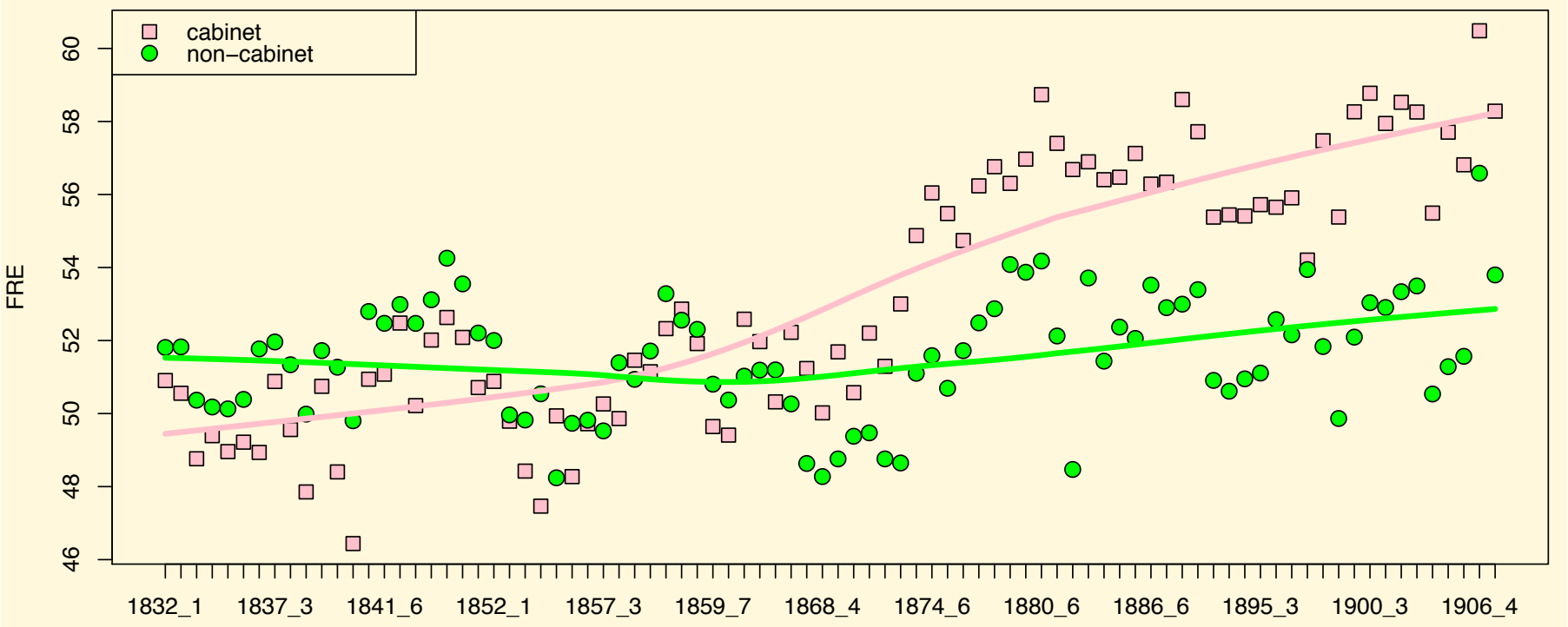
↓ local, clientalistic appeals via bribery...

↑ 'party orientated electorate', with national policies and national leaders

Q how did these leaders respond to new voters?

A by changing nature of their speech: simpler, less complex expressions in parliament

Flesch overtime plot



Dale-Chall, 1948

yields grade level of text sample.

DC

$$0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000) 'familiar' words.

e.g. `about`, `back`, `call`, etc.

Partner Exercise

The FRE of SOTU speeches is declining. Why might it be difficult to make readability comparisons over time? (hint: when were the reading ease measures invented? are topics of speeches constant? were addresses always delivered the same way?)

Does the nature of the decline suggest that speeches are becoming simpler for demand (i.e. voter) or supply (i.e. leader) incentive reasons? (hint: consider the smoothness/jaggedness of the decrease)

Readability scores

TABLE 2: Readability scores

Score	Formula
Flesch Reading Ease	$206.835 - 1.015 \times AWS - 84.6 \times ASW$
Flesch-Kincaid	$-15.59 + 0.390 \times AWS + 11.8 \times ASW$
Gunning Fog	$0.4 \times (AWS + 100 \times PWW)$
SMOG	$3.1291 + 5.7127 \times \sqrt{APS}$
Dale-Chall	$3.6365 + 0.0496 \times AWS + 15.79 \times DWW$

Notes. AWS: average number of words per sentence; ASW: average number of syllables per word; PWW: ratio of polysyllabic words (3+ syllables) to word count; APS: average number of polysyllabic words per sentence; DWW: ratio of difficult words (not on Dale-Chall list) to word count.

Descriptive Statistics: Stylometrics

Mystery of *The Federalist Papers*

85 essays published anonymously in 1787 and 1788

Generally agreed that Alexander Hamilton wrote 51 essays, John Jay wrote 5 essays, James Madison wrote 14 essays, and 3 essays were written jointly by Hamilton and Madison.

That leaves 12 that are disputed.

Mosteller and Wallace, 1963/4

In essence, they...

Count word frequencies of function words (`by`, `from`, `to`, etc.) in the 73 essays with undisputed authorship

then collapse on author to get word frequencies specific to the authors

now model these author-specific rates with Poisson and negative binomial distributions

use Bayes' theorem to determine the posterior probability that Hamilton (Madison) wrote a particular disputed essay for all such essays

i.e. they ask "if rates of function word usage are constant within authors for these documents, which author was most likely to have written essay x given the observed function word usage of these authors on the other documents?"

More Details

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

may think that sentence length distinguishes authors, but Hamilton and Madison "practically twins" on this.

use function words—conjunctions, prepositions, pronouns—for two (related) reasons:

1. authors use them unconsciously
2. therefore, don't vary much by topic.

NB typically assume one instance of a function word is independent of the next, and use is fixed over a lifetime (and constant within a given text).

→ wrong, but models relying on these assns discriminate well (see Peng & Hengartner on e.g. Austin v Shakespeare)

The Vector Space Model of Text

1) Task:
   - Numerous tasks will suppose that we can measure document similarity or dissimiliarity
2) Objective Function
   - For a variety of tasks, will impose some measure or definition of similarity, dissimilarity, or distance.

$$\begin{aligned} d(\boldsymbol{X}_i, \boldsymbol{X}_j) &= \text{Dissimilarity(Distance)} \rightsquigarrow \text{Bigger implies further apart} \\ s(\boldsymbol{X}_i, \boldsymbol{X}_j) &= \text{Similarity} \rightsquigarrow \text{Bigger implies closer together} \end{aligned}$$

   - Objective functions$\rightsquigarrow$ determine which points we compare and aggregate similarity, dissimilarity, and distance
3) Optimization
   - Depends on the particular task, likely arranging/grouping objects to find similarity
4) Validation
   - Are the mathematical definitions of similarity actually similar for our particular purpose?

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry$\rightsquigarrow$ modify with word weighting

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry$\rightsquigarrow$ modify with word weighting
- Natural notions of distance

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry$\rightsquigarrow$ modify with word weighting
- Natural notions of distance
- Kernel Trick: richer comparisons of large feature spaces

Texts and Geometry

Consider a document-term matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 2 & 0 & \dots & 0 \\ 0 & 0 & 3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 3 \end{pmatrix}$$

Suppose documents live in a space $\rightsquigarrow$ rich set of results from linear algebra

- Provides a geometry$\rightsquigarrow$ modify with word weighting
- Natural notions of distance
- Kernel Trick: richer comparisons of large feature spaces
- Building block for clustering, supervised learning, and scaling

Texts in Space

Texts in Space

$$\text{Doc1} \quad = \quad (1, 1, 3, \ldots, 5)$$

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \end{aligned}$$

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \textbf{Doc1}, \textbf{Doc2} &\in \Re^J \end{aligned}$$

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \mathbf{Doc1}, \mathbf{Doc2} &\in \Re^J \end{aligned}$$

Inner Product between documents:

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \mathbf{Doc1}, \mathbf{Doc2} &\in \Re^J \end{aligned}$$

Inner Product between documents:

$$\mathbf{Doc1} \cdot \mathbf{Doc2} = (1, 1, 3, \ldots, 5)'(2, 0, 0, \ldots, 1)$$

Texts in Space

$$\begin{aligned} \text{Doc1} &= (1, 1, 3, \ldots, 5) \\ \text{Doc2} &= (2, 0, 0, \ldots, 1) \\ \mathbf{Doc1}, \mathbf{Doc2} &\in \Re^J \end{aligned}$$

Inner Product between documents:

$$\begin{aligned} \mathbf{Doc1} \cdot \mathbf{Doc2} &= (1, 1, 3, \ldots, 5)'(2, 0, 0, \ldots, 1) \\ &= 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1 \end{aligned}$$

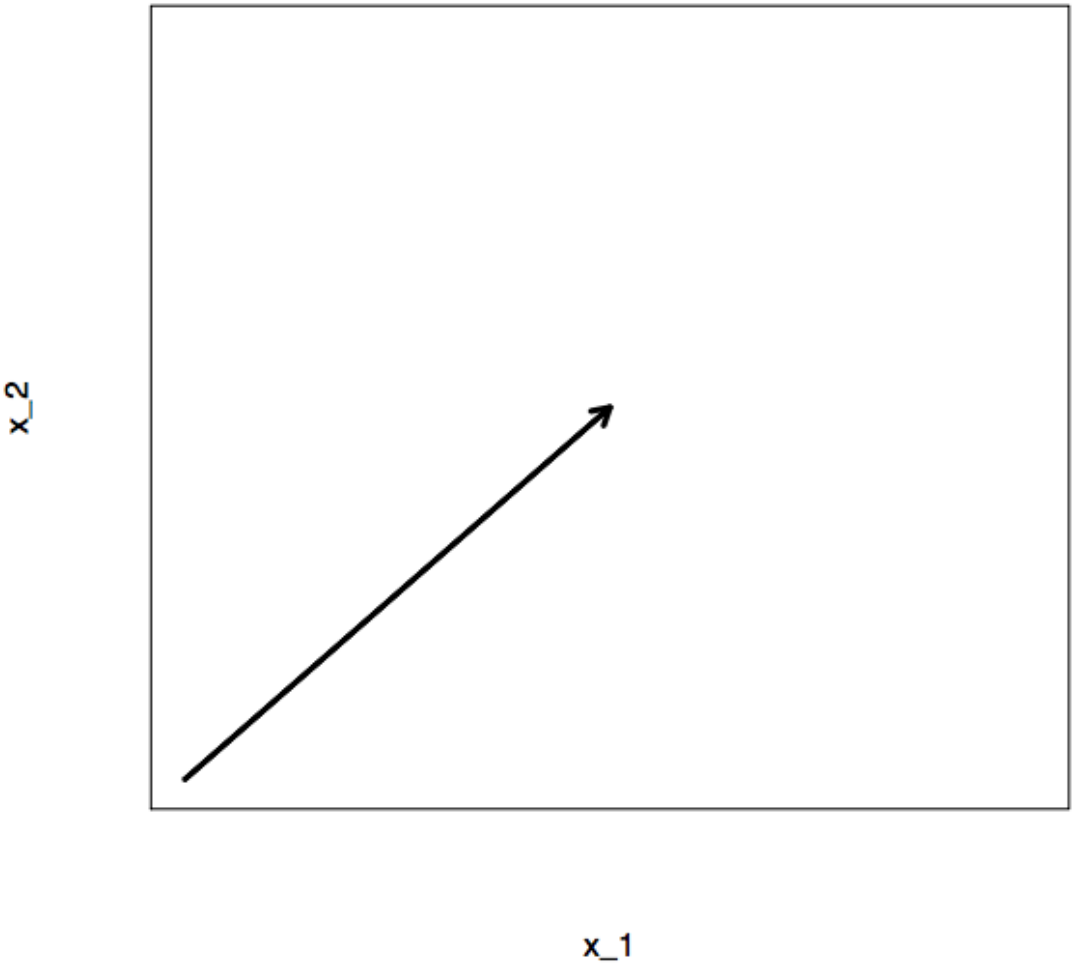
Texts in Space

$$
\begin{aligned}
\text{Doc1} &= (1, 1, 3, \ldots, 5) \\
\text{Doc2} &= (2, 0, 0, \ldots, 1) \\
\mathbf{Doc1}, \mathbf{Doc2} &\in \Re^J
\end{aligned}
$$

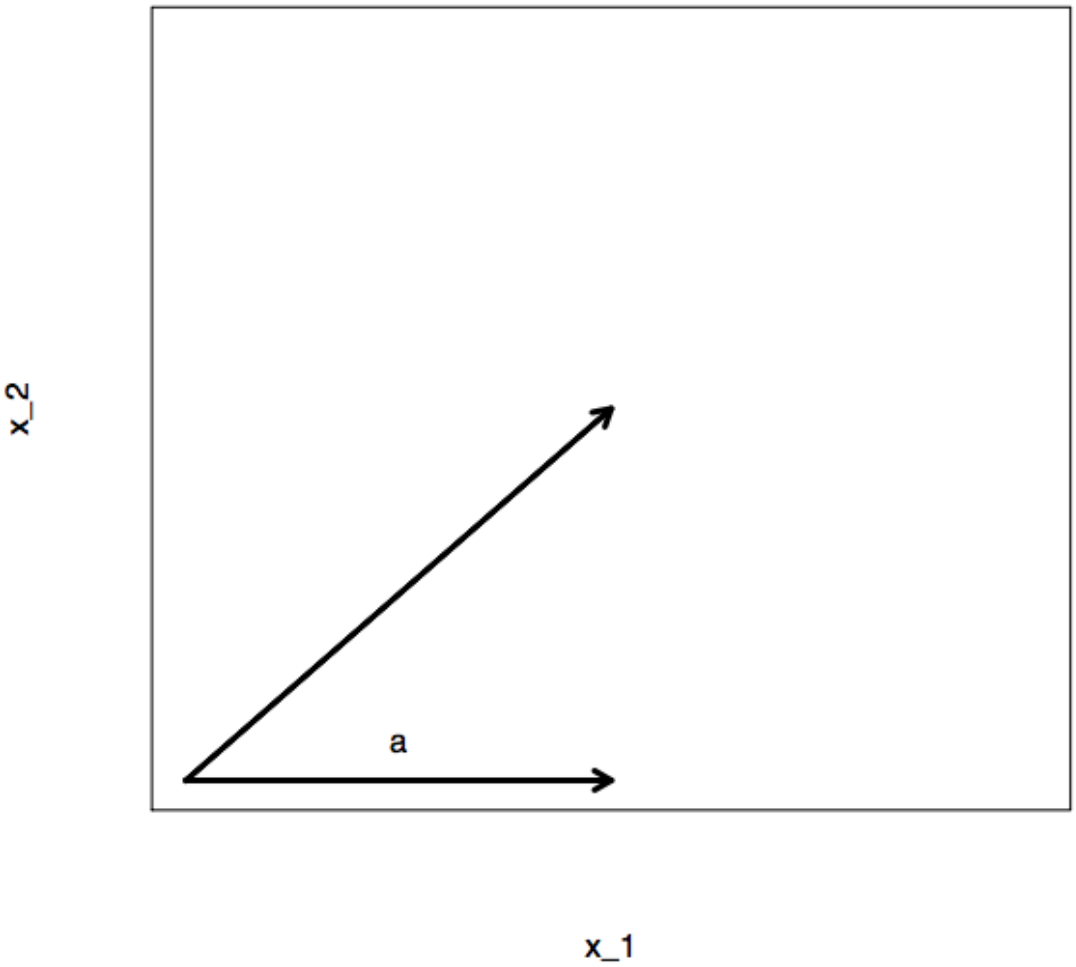
Inner Product between documents:

$$
\begin{aligned}
\mathbf{Doc1} \cdot \mathbf{Doc2} &= (1, 1, 3, \ldots, 5)'(2, 0, 0, \ldots, 1) \\
&= 1 \times 2 + 1 \times 0 + 3 \times 0 + \ldots + 5 \times 1 \\
&= 7
\end{aligned}
$$

Vector Length

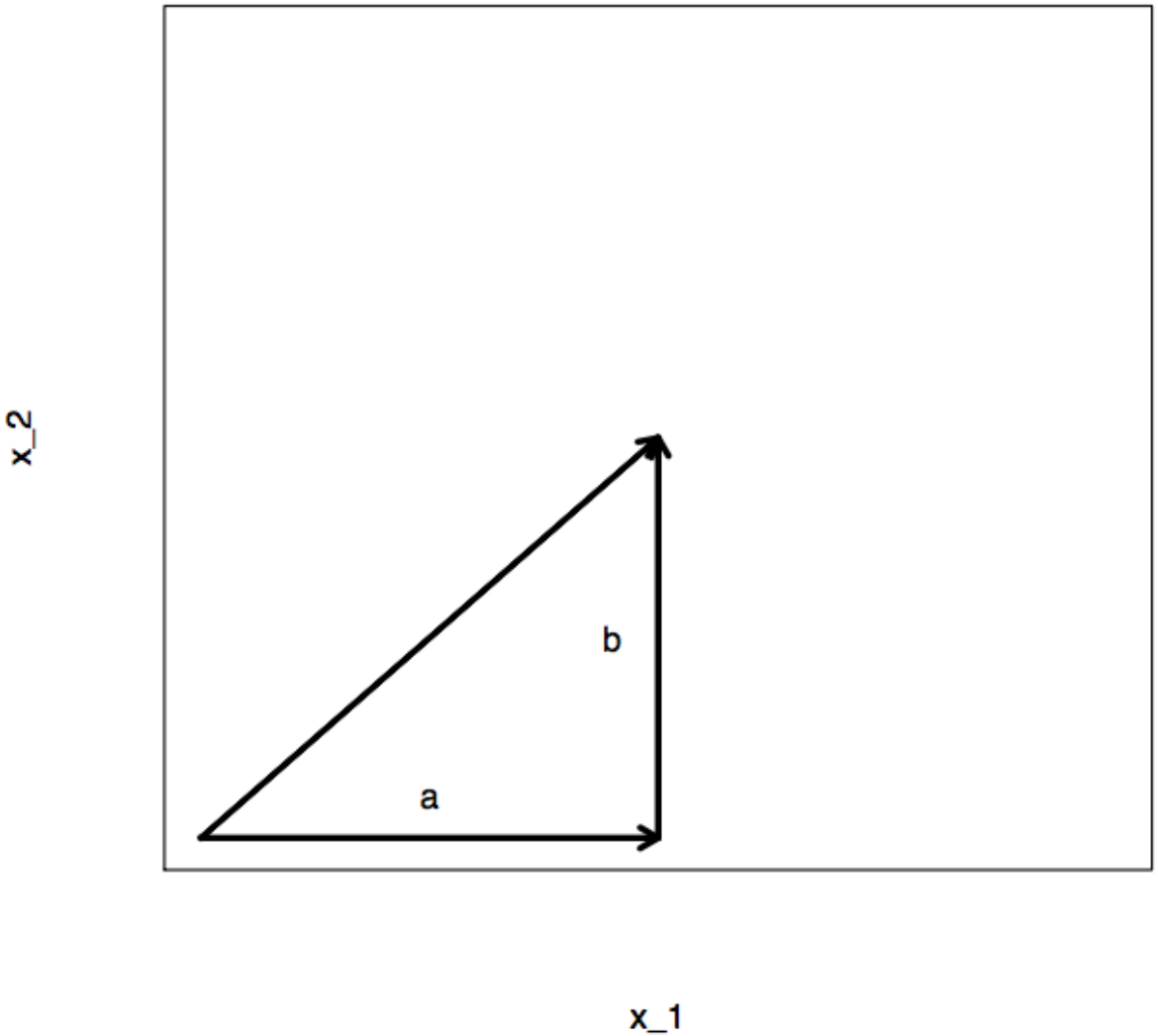


Vector Length



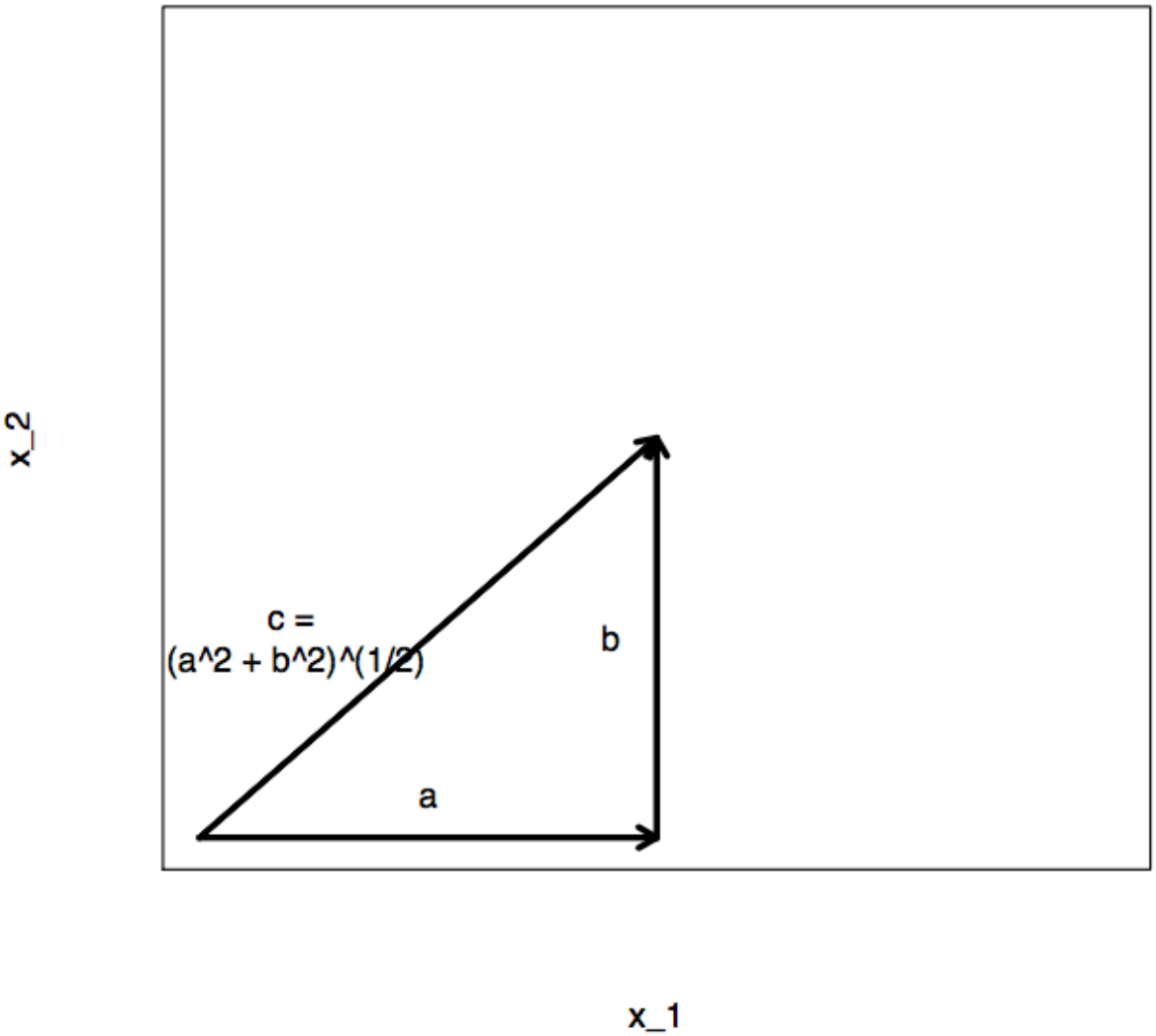
- Pythogorean Theorem: Side with length a

Vector Length



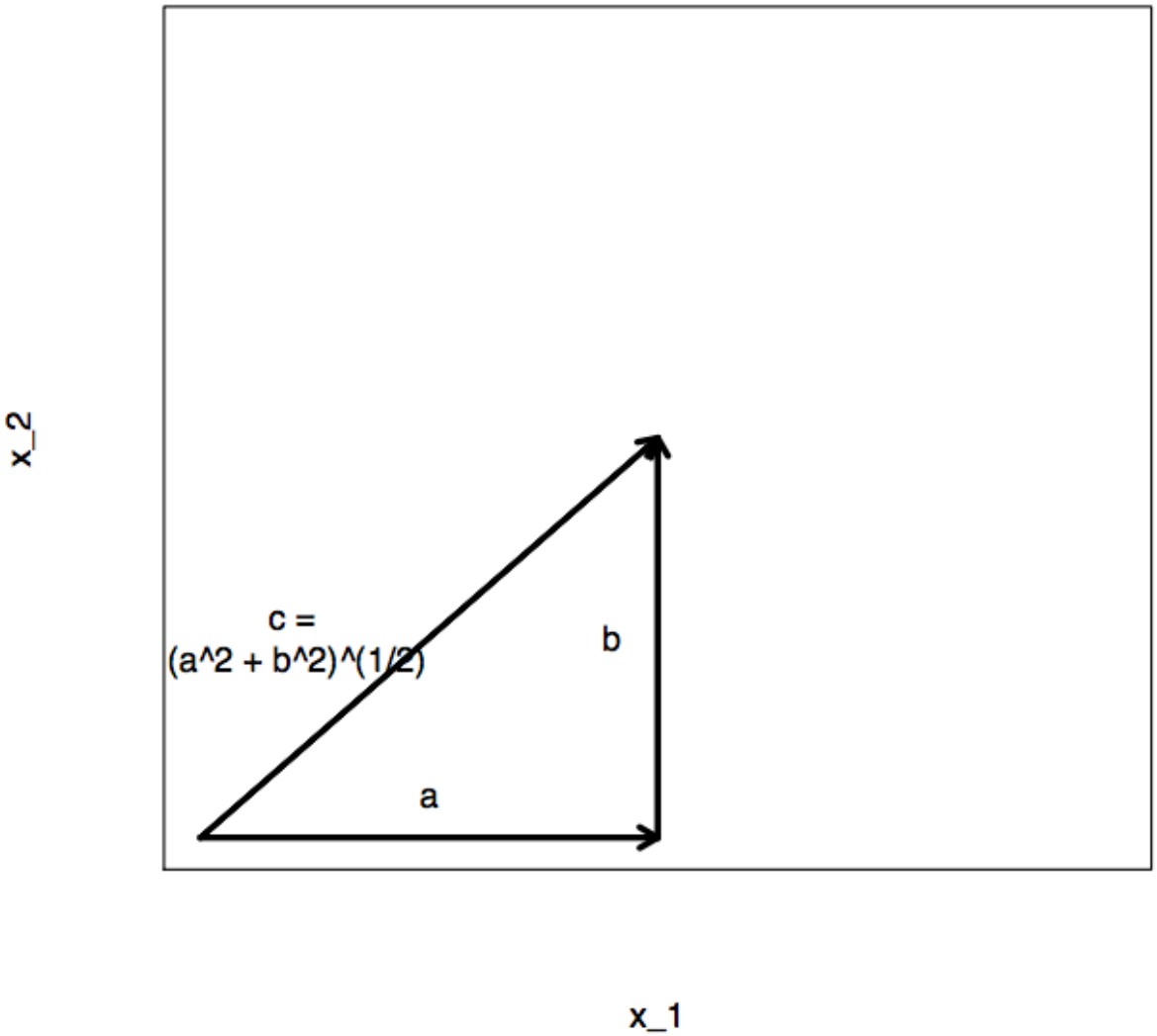
- Pythogorean Theorem: Side with length a
- Side with length b and right triangle

Vector Length



- Pythogorean Theorem: Side with length a
- Side with length b and right triangle
- $c = \sqrt{a^2 + b^2}$

Vector Length



- Pythogorean Theorem: Side with length a
- Side with length b and right triangle
- $c = \sqrt{a^2 + b^2}$
- This is generally true

Vector (Euclidean) Length

Definition

Suppose $\boldsymbol{v} \in \Re^J$. *Then, we will define its length as*

$$\begin{aligned} ||\boldsymbol{v}|| &= (\boldsymbol{v} \cdot \boldsymbol{v})^{1/2} \\ &= (v_1^2 + v_2^2 + v_3^2 + \ldots + v_J^2)^{1/2} \end{aligned}$$

Normalized length of a document is equal to each of the document's coordinates squared, added together, and taken the squared root; this will be useful for many distance measures. I allows us to think about measuring distance in some principled way.

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$
4) $d(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k)$

$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$
4) $d(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k)$

Explore distance functions to compare documents$\rightsquigarrow$

Measures of Dissimilarity

Initial guess$\rightsquigarrow$ Distance metrics

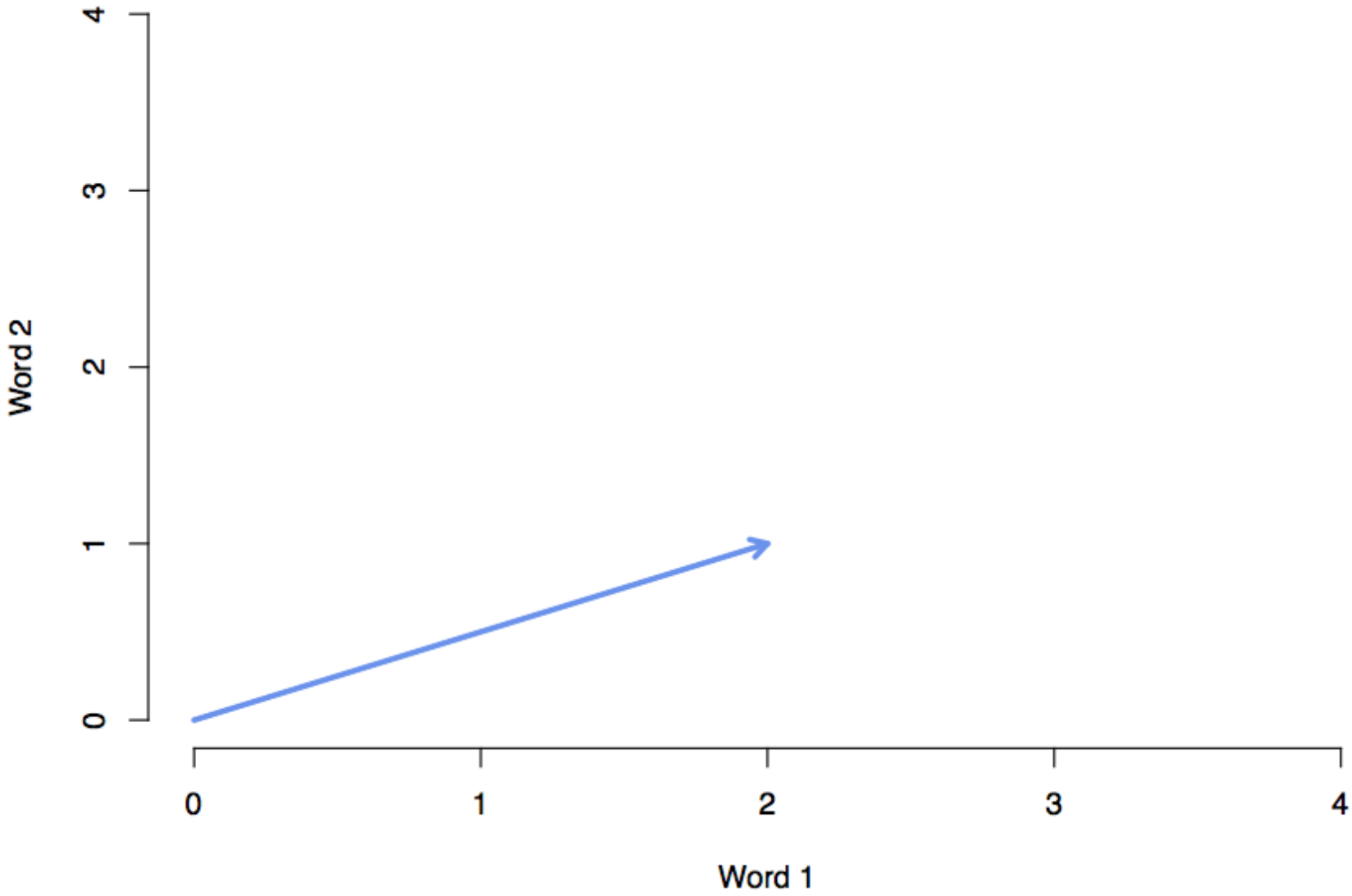
Properties of a metric: (distance function) $d(\cdot, \cdot)$. Consider arbitrary documents $\boldsymbol{X}_i$, $\boldsymbol{X}_j$, $\boldsymbol{X}_k$

1) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$
2) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0$ if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$
3) $d(\boldsymbol{X}_i, \boldsymbol{X}_j) = d(\boldsymbol{X}_j, \boldsymbol{X}_i)$
4) $d(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq d(\boldsymbol{X}_i, \boldsymbol{X}_j) + d(\boldsymbol{X}_j, \boldsymbol{X}_k)$

Explore distance functions to compare documents$\rightsquigarrow$Do we want additional assumptions/properties?

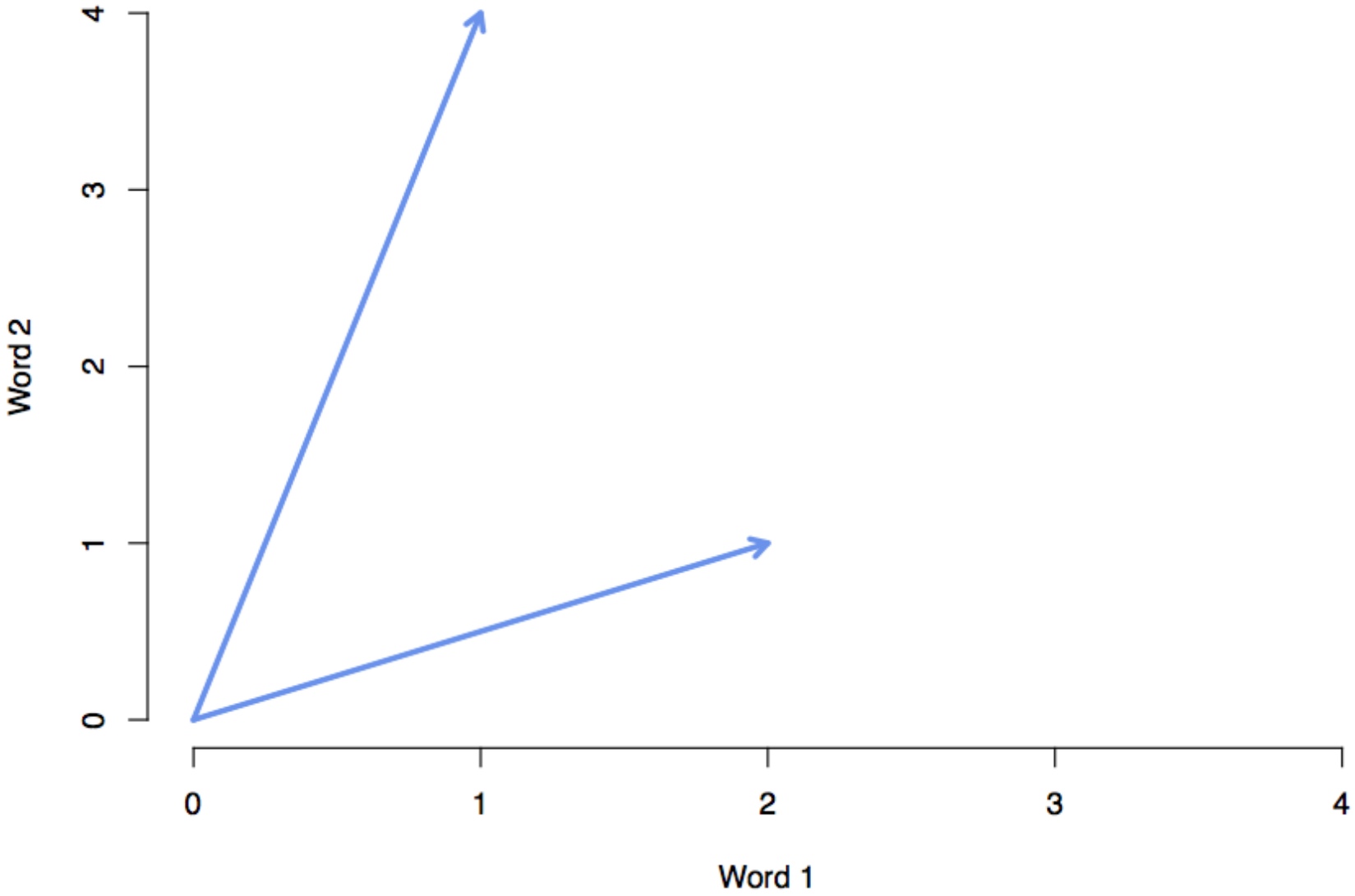
Measuring the Distance Between Documents

Euclidean Distance [e.g. "yo" and "bro"] – length of red vector



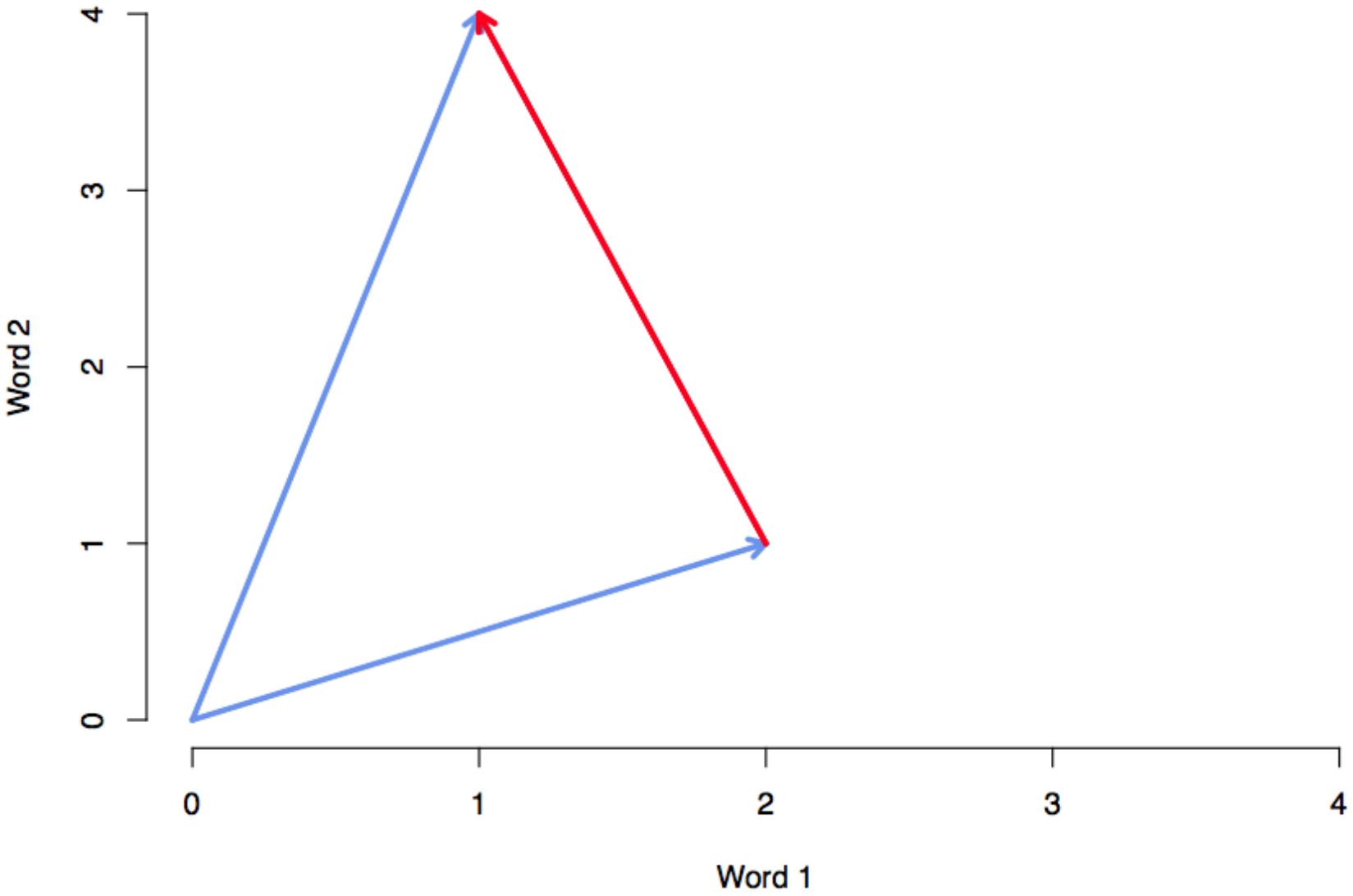
Measuring the Distance Between Documents

Euclidean Distance [e.g. "yo" and "bro"] – length of red vector



Measuring the Distance Between Documents

Euclidean Distance [e.g. "yo" and "bro"] – length of red vector



Measuring the Distance Between Documents

Definition

The Euclidean distance between documents $\boldsymbol{X}_i$ *and* $\boldsymbol{X}_j$ *as*

$$||\boldsymbol{X}_i - \boldsymbol{X}_j|| = \sqrt{\sum_{m=1}^{J} (x_{im} - x_{jm})^2}$$

Measuring the Distance Between Documents

Definition

The Euclidean distance between documents $\boldsymbol{X}_i$ *and* $\boldsymbol{X}_j$ *as*

$$||\boldsymbol{X}_i - \boldsymbol{X}_j|| = \sqrt{\sum_{m=1}^{J} (x_{im} - x_{jm})^2}$$

Suppose $\boldsymbol{X}_i = (1, 4)$ and $\boldsymbol{X}_j = (2, 1)$. The distance between the documents is:

$$\begin{aligned} ||(1,4) - (2,1)|| &= \sqrt{(1-2)^2 + (4-1)^2} \\ &= \sqrt{10} \end{aligned}$$

Measuring the Distance Between Documents

Euclidean distance rewards magnitude, rather than direction. i.e. it doesn't reward being close in relative use of terms. Instead, rewards documents that are similarly 'far' from the origin.

We can do better by normalizing document length, and rewarding similar uses of terms. To do this, we divide each of the components (the documents) by their length.

Measuring the Distance Between Documents

Euclidean distance rewards magnitude, rather than direction. i.e. it doesn't reward being close in relative use of terms. Instead, rewards documents that are similarly 'far' from the origin.
We can do better by normalizing document length, and rewarding similar uses of terms. To do this, we divide each of the components (the documents) by their length.

Measuring the Distance Between Documents

Many distance metrics

Measuring the Distance Between Documents

Many distance metrics Consider the Minkowski family

Measuring the Distance Between Documents

Many distance metrics Consider the Minkowski family

Definition

The Minkowski Distance between documents $\mathbf{X}_i$ and $\mathbf{X}_j$ for value p is

$$d_p(\mathbf{X}_i, \mathbf{X}_j) = \left(\sum_{m=1}^{J} |x_{im} - x_{jm}|^p \right)^{1/p}$$

Members of the Minkowski Family

Members of the Minkowski Family

Manhattan metric

Members of the Minkowski Family

Manhattan metric

$$d_1(\mathbf{X}_i, \mathbf{X}_j) = \sum_{m=1}^{J} |x_{im} - x_{jm}|$$

Members of the Minkowski Family

Manhattan metric

$$
\begin{aligned}
d_1(\mathbf{X}_i, \mathbf{X}_j) &= \sum_{m=1}^{J} |x_{im} - x_{jm}| \\
d_1((1,4),(2,1)) &= |1| + |3| = 4
\end{aligned}
$$

Members of the Minkowski Family

Manhattan metric

$$
\begin{aligned}
d_1(\mathbf{X}_i, \mathbf{X}_j) &= \sum_{m=1}^{J} |x_{im} - x_{jm}| \\
d_1((1,4),(2,1)) &= |1| + |3| = 4
\end{aligned}
$$

Minkowski (p) metric

Members of the Minkowski Family

Manhattan metric

$$\begin{aligned} d_1(\mathbf{X}_i, \mathbf{X}_j) &= \sum_{m=1}^{J} |x_{im} - x_{jm}| \\ d_1((1,4),(2,1)) &= |1| + |3| = 4 \end{aligned}$$

Minkowski (p) metric

$$\begin{aligned} d_p(\mathbf{X}_i, \mathbf{X}_j) &= \left(\sum_{m=1}^{J} |x_{im} - x_{jm}|^p \right)^{1/p} \\ d_p((1,4),(2,1)) &= (|1-2|^p + |4-1|^p)^{1/p} \end{aligned}$$

What Does p Do?

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain Chebyshev's Metric: all that matters is the max difference on a particular coordinate, because every other difference will be overwhelmed.

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain Chebyshev's Metric: all that matters is the max difference on a particular coordinate, because every other difference will be overwhelmed.

$$\lim_{p \to \infty} d_p(\mathbf{X}_i, \mathbf{X}_j) \quad = \quad \max_{m=1}^{J} |x_{im} - x_{jm}|$$

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain Chebyshev's Metric: all that matters is the max difference on a particular coordinate, because every other difference will be overwhelmed.

$$\lim_{p \to \infty} d_p(\mathbf{X}_i, \mathbf{X}_j) \quad = \quad \max_{m=1}^{J} |x_{im} - x_{jm}|$$

In words: distance between documents only the biggest difference

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain Chebyshev's Metric: all that matters is the max difference on a particular coordinate, because every other difference will be overwhelmed.

$$\lim_{p \to \infty} d_p(\mathbf{X}_i, \mathbf{X}_j) \quad = \quad \max_{m=1}^{J} |x_{im} - x_{jm}|$$

In words: distance between documents only the biggest difference
All other differences do not contribute to distance measure

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain Chebyshev's Metric: all that matters is the max difference on a particular coordinate, because every other difference will be overwhelmed.

$$\lim_{p \to \infty} d_p(\mathbf{X}_i, \mathbf{X}_j) \quad = \quad \max_{m=1}^{J} |x_{im} - x_{jm}|$$

In words: distance between documents only the biggest difference
All other differences do not contribute to distance measure
Decreasing $p \rightsquigarrow$ greater importance of coordinates with smallest differences

What Does p Do?

Increasing $p \rightsquigarrow$ greater importance of coordinates with largest differences
If we let $p \to \infty$ Obtain Chebyshev's Metric: all that matters is the max difference on a particular coordinate, because every other difference will be overwhelmed.

$$\lim_{p \to \infty} d_p(\mathbf{X}_i, \mathbf{X}_j) = \max_{m=1}^{J} |x_{im} - x_{jm}|$$

In words: distance between documents only the biggest difference
All other differences do not contribute to distance measure
Decreasing $p \rightsquigarrow$ greater importance of coordinates with smallest differences

$$\lim_{p \to -\infty} d_p(\mathbf{X}_i, \mathbf{X}_j) = \min_{m=1}^{J} |x_{im} - x_{jm}|$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) = 10$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$\begin{aligned} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\ d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \end{aligned}$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$\begin{aligned} d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\ d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \\ d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \end{aligned}$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$
\begin{aligned}
d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^2 + 4^2 + 3^2} = \sqrt{125} = 11.18
\end{aligned}
$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$
\begin{aligned}
d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^2 + 4^2 + 3^2} = \sqrt{125} = 11.18 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10
\end{aligned}
$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$
\begin{aligned}
d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^2 + 4^2 + 3^2} = \sqrt{125} = 11.18 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^4 + 4^4 + 3^4} = (10337)^{1/4} = 10.08
\end{aligned}
$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$
\begin{aligned}
d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^2 + 4^2 + 3^2} = \sqrt{125} = 11.18 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^4 + 4^4 + 3^4} = (10337)^{1/4} = 10.08 \\
d_\infty(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10
\end{aligned}
$$

Comparing the Metrics

Suppose $\boldsymbol{X}_i = (10, 4, 3)$, $\boldsymbol{X}_j = (0, 4, 3)$, and $\boldsymbol{X}_k = (0, 0, 0)$
Then:

$$
\begin{aligned}
d_1(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_1(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10 + 4 + 3 = 17 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_2(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^2 + 4^2 + 3^2} = \sqrt{125} = 11.18 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_4(\boldsymbol{X}_i, \boldsymbol{X}_k) &= \sqrt{10^4 + 4^4 + 3^4} = (10337)^{1/4} = 10.08 \\
d_\infty(\boldsymbol{X}_i, \boldsymbol{X}_j) &= 10 \\
d_\infty(\boldsymbol{X}_i, \boldsymbol{X}_k) &= 10
\end{aligned}
$$

Are all differences equal?

Previous metrics treat all dimensions as equal

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting
Mahalanobis Distance

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting
Mahalanobis Distance

Definition

Suppose that we have a covariance matrix $\mathbf{\Sigma}$

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting
Mahalanobis Distance

Definition

Suppose that we have a covariance matrix $\mathbf{\Sigma}$. *Then we can define the Mahalanobis Distance between documents* $\mathbf{X}_i$ *and* $\mathbf{X}_j$ *as* ,

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting
Mahalanobis Distance

Definition

Suppose that we have a covariance matrix $\mathbf{\Sigma}$. *Then we can define the Mahalanobis Distance between documents* $\mathbf{X}_i$ *and* $\mathbf{X}_j$ *as ,*

$$d_{Mah}(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \mathbf{\Sigma}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting
Mahalanobis Distance

Definition

Suppose that we have a covariance matrix $\mathbf{\Sigma}$. *Then we can define the Mahalanobis Distance between documents* $\mathbf{X}_i$ *and* $\mathbf{X}_j$ *as ,*

$$d_{Mah}(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \mathbf{\Sigma}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

More generally: $\mathbf{\Sigma}$ could be symmetric and positive-definite

Are all differences equal?

Previous metrics treat all dimensions as equal
We may want to engage in some scaling/reweighting
Mahalanobis Distance

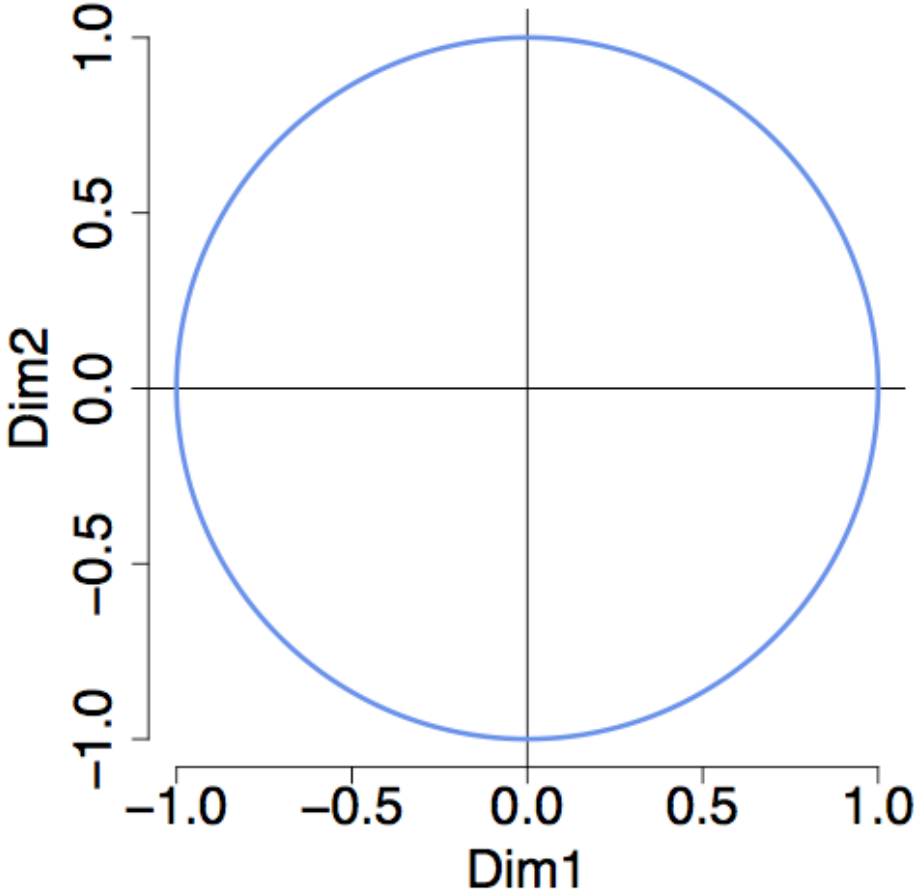
Definition

Suppose that we have a covariance matrix $\mathbf{\Sigma}$. *Then we can define the Mahalanobis Distance between documents* $\mathbf{X}_i$ *and* $\mathbf{X}_j$ *as ,*

$$d_{Mah}(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \mathbf{\Sigma}^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

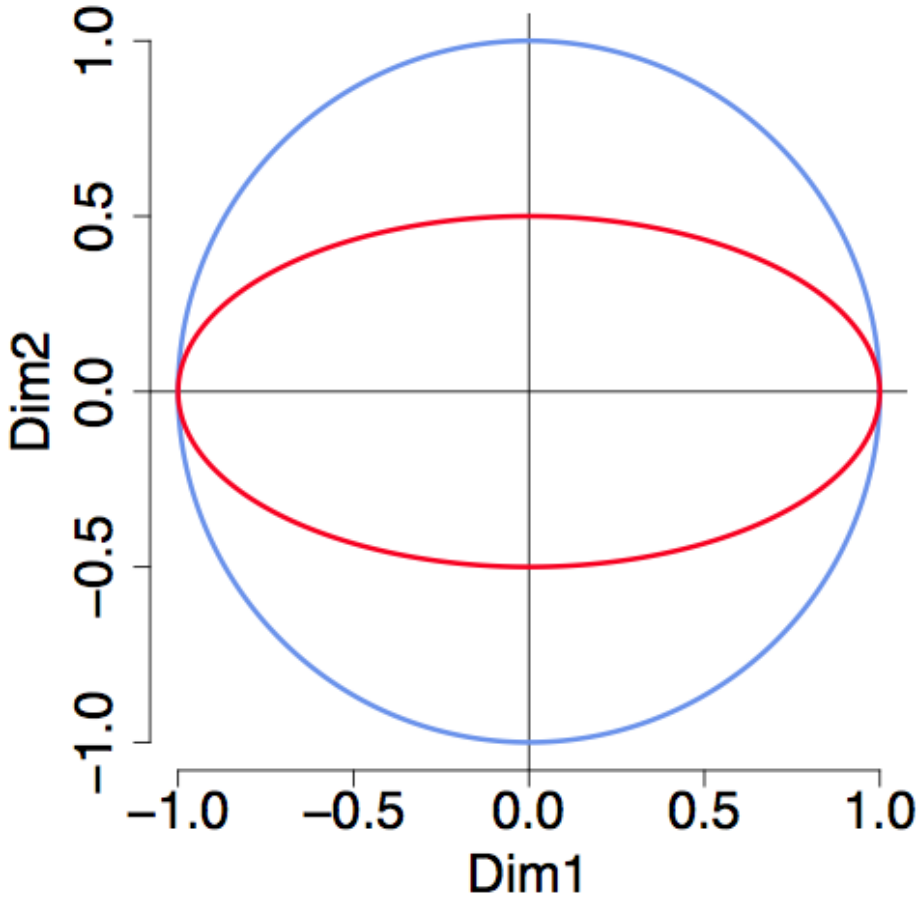
More generally: $\mathbf{\Sigma}$ could be symmetric and positive-definite
What does $\mathbf{\Sigma}$ do?

Some Intuition: The Unit Circle



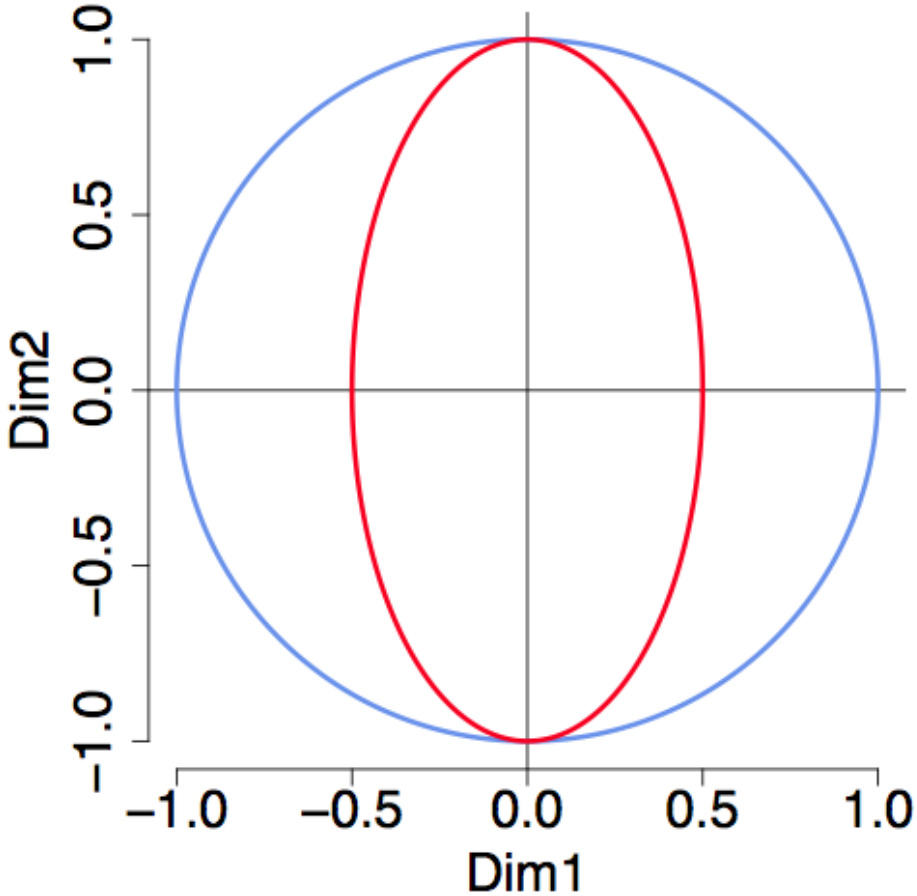
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Some Intuition: The Unit Circle



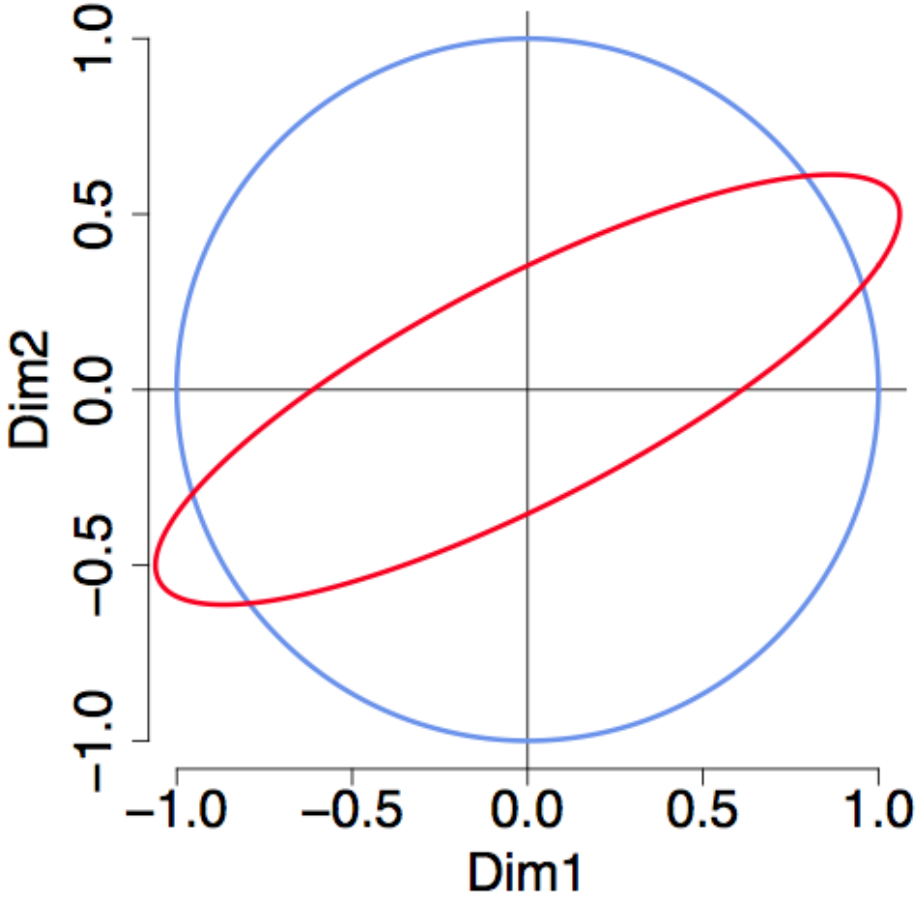
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}$$

Some Intuition: The Unit Circle



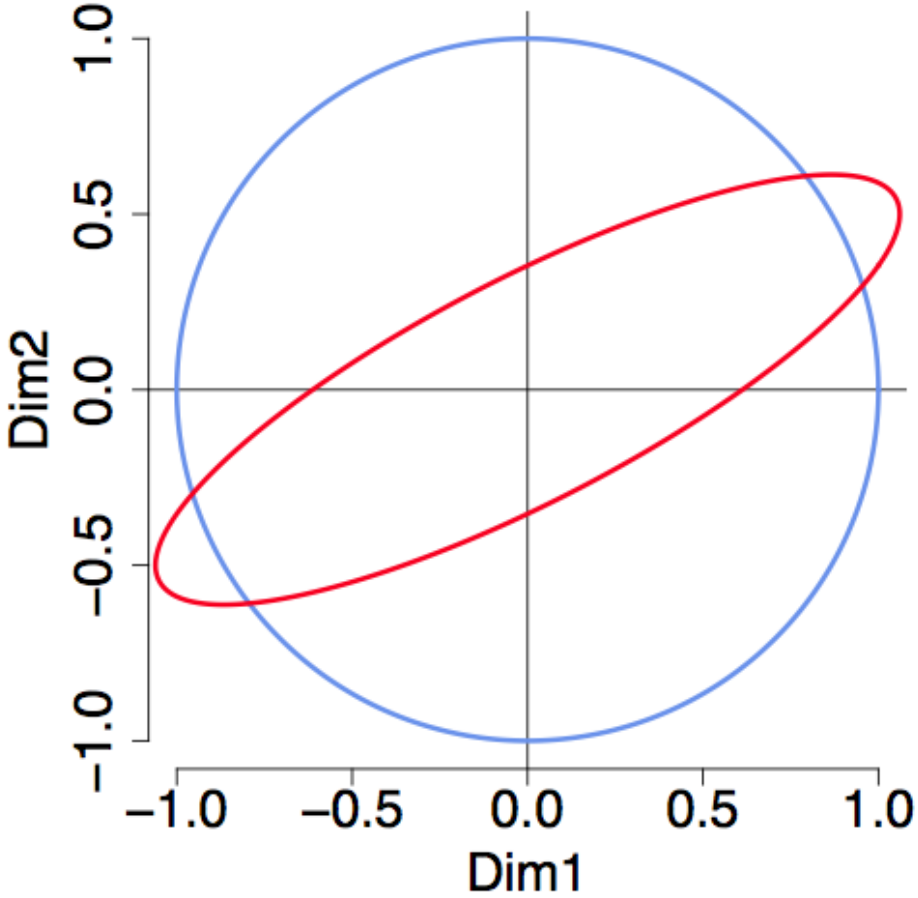
$$\mathbf{\Sigma} = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}$$

Some Intuition: The Unit Circle



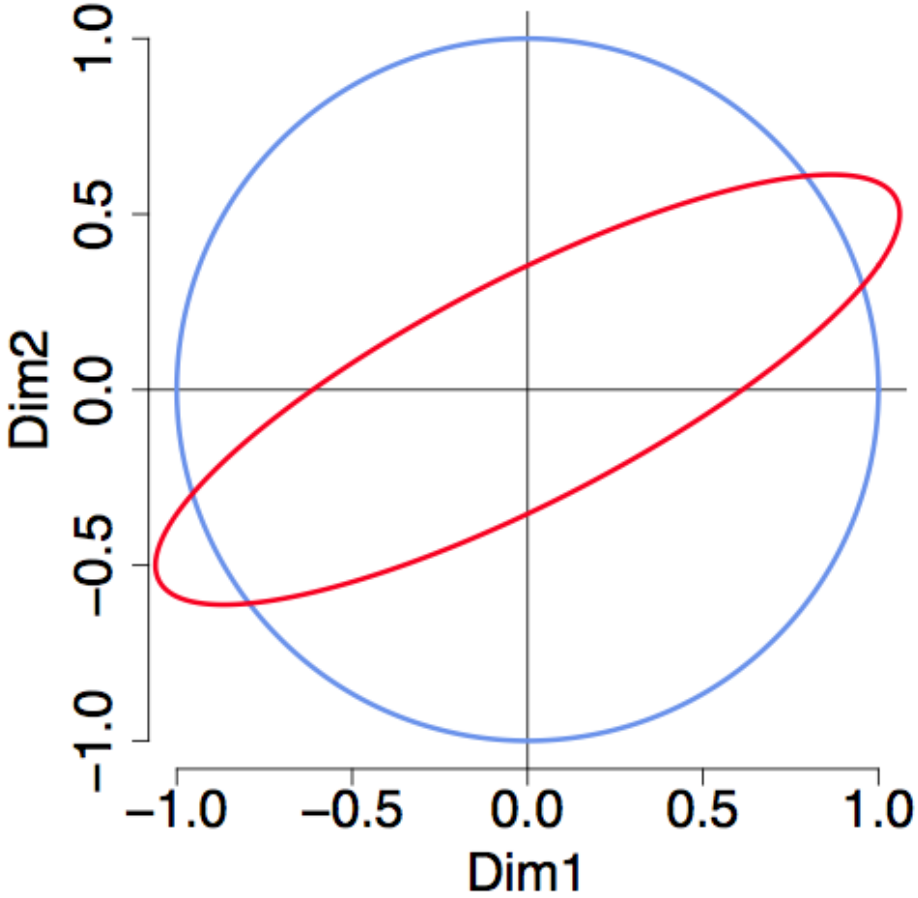
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Some Intuition: The Unit Circle



$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Some Intuition: The Unit Circle



$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

Measuring Similarity

Measuring Similarity

What properties should similarity measure have?

Measuring Similarity

What properties should similarity measure have?

- Maximum: document with itself

Measuring Similarity

What properties should similarity measure have?

- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)

Measuring Similarity

What properties should similarity measure have?

- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)
- Increasing when more of same words used

Measuring Similarity

What properties should similarity measure have?

- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)
- Increasing when more of same words used
- ? $s(a, b) = s(b, a)$.

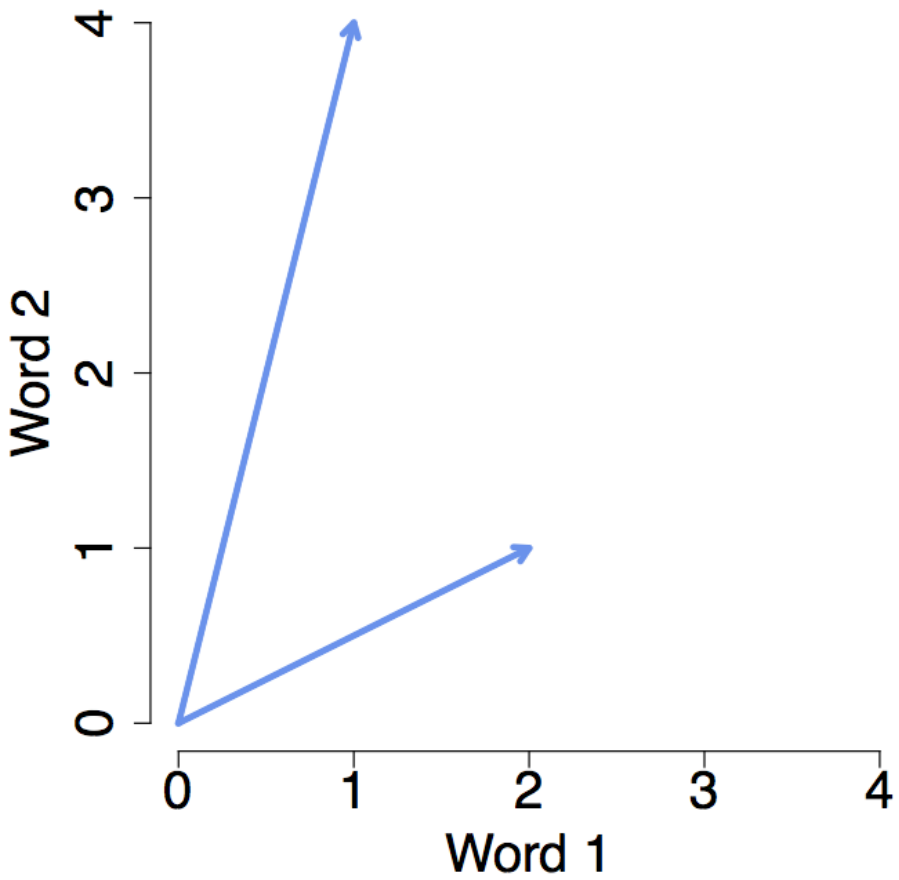
Measuring Similarity

What properties should similarity measure have?

- Maximum: document with itself
- Minimum: documents have no words in common (orthogonal)
- Increasing when more of same words used
- ? $s(a, b) = s(b, a)$.

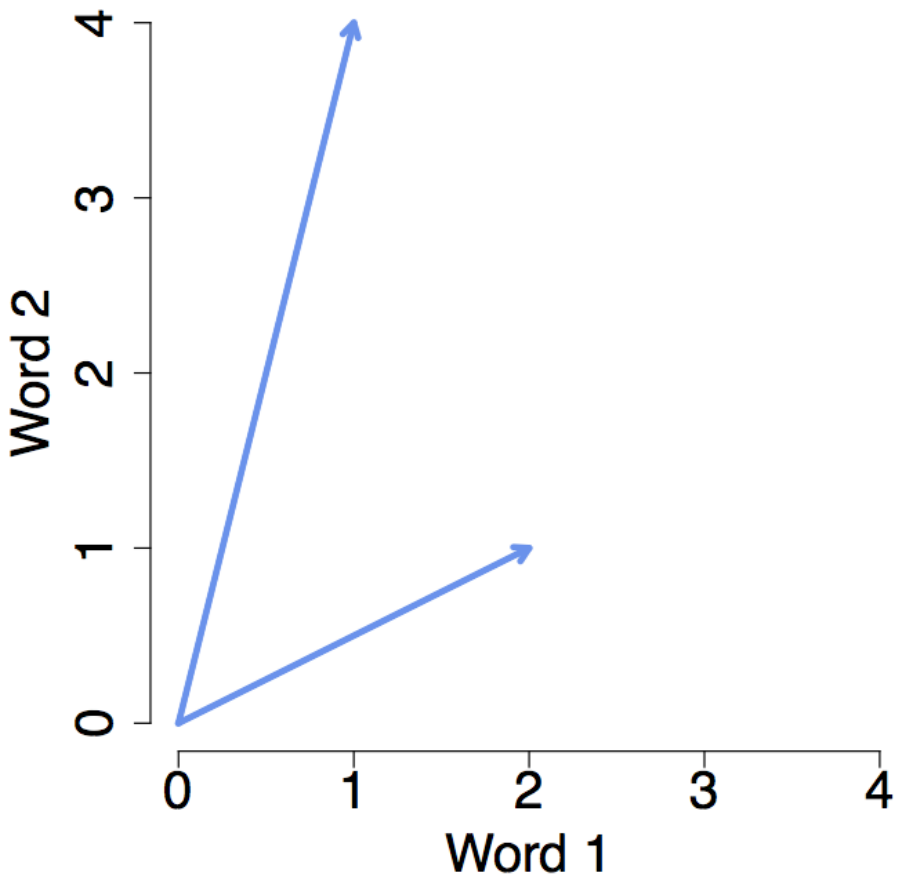
How should additional words be treated?

Measuring Similarity



Measure 1: Inner product

Measuring Similarity



Measure 1: Inner product

$$(2,1)' \cdot (1,4) = 6$$

If we have two vectors (unit or otherwise):

$a = (a_1, a_2, \ldots, a_n)$

$b = (b_1, b_2, \ldots, b_n)$

ten their dot/inner product is defined as:

$a \bullet b = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n + b_n$

That is, it is just the sum of the termwise multiplication between elements;

If we have two vectors (unit or otherwise):

$a = (a_1, a_2, \ldots, a_n)$

$b = (b_1, b_2, \ldots, b_n)$

ten their dot/inner product is defined as:

$a \bullet b = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n + b_n$

That is, it is just the sum of the termwise multiplication between elements;

If we have two vectors (unit or otherwise):

$a = (a_1, a_2, \ldots, a_n)$

$b = (b_1, b_2, \ldots, b_n)$

ten their dot/inner product is defined as:

$a \bullet b = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n + b_n$

That is, it is just the sum of the termwise multiplication between elements;

If we have two vectors (unit or otherwise):

$a = (a_1, a_2, \ldots, a_n)$

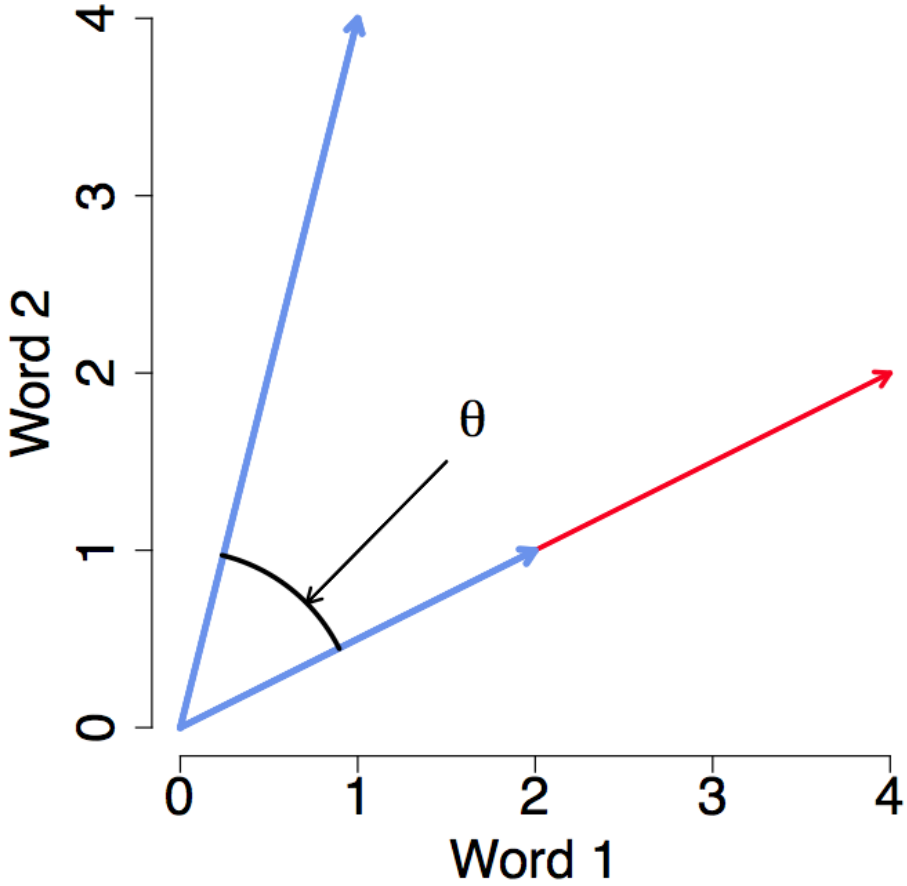
$b = (b_1, b_2, \ldots, b_n)$

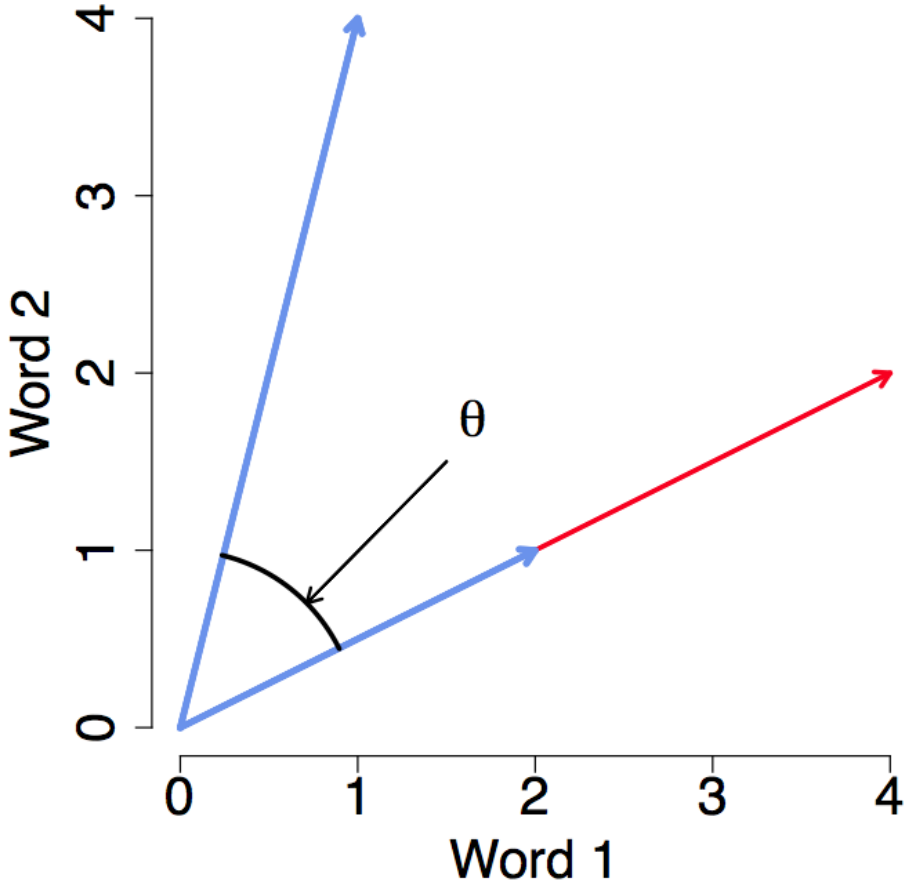
ten their dot/inner product is defined as:

$a \bullet b = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n + b_n$

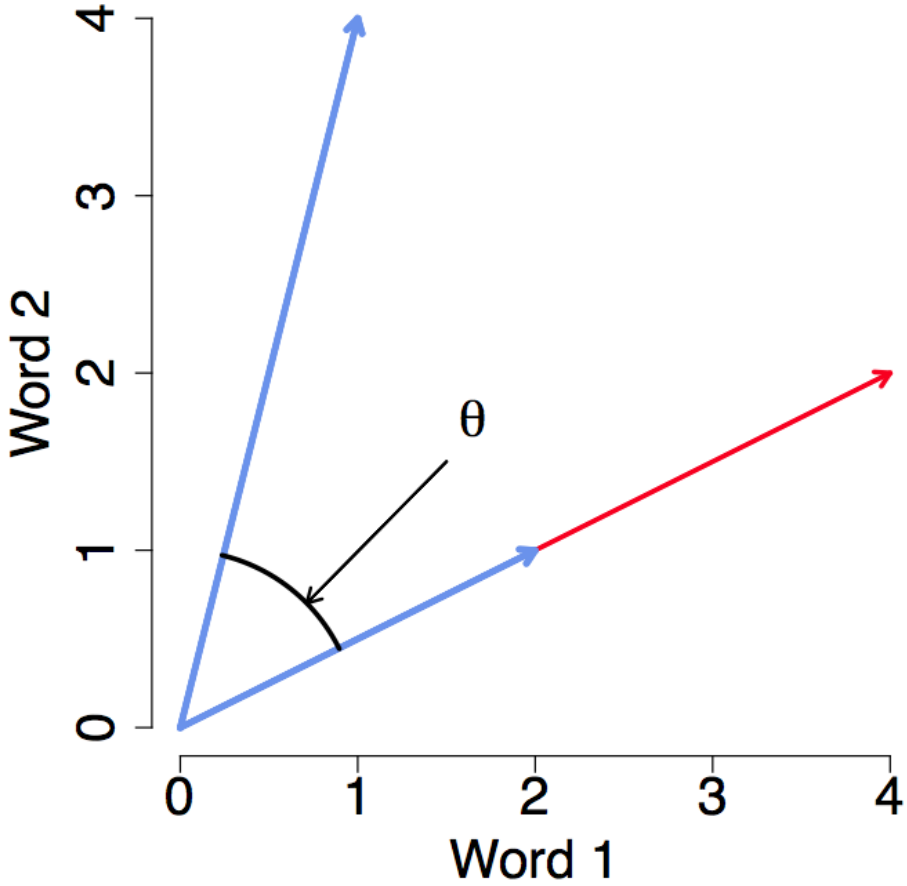
That is, it is just the sum of the termwise multiplication between elements;

If we have two vectors (unit or otherwise):
$a = (a_1, a_2, \ldots, a_n)$
$b = (b_1, b_2, \ldots, b_n)$
ten their dot/inner product is defined as:
$a \bullet b = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n + b_n$
That is, it is just the sum of the termwise multiplication between elements;


Word 2
0
1
2
3
4
θ
0
1
2
3
4
Word 1

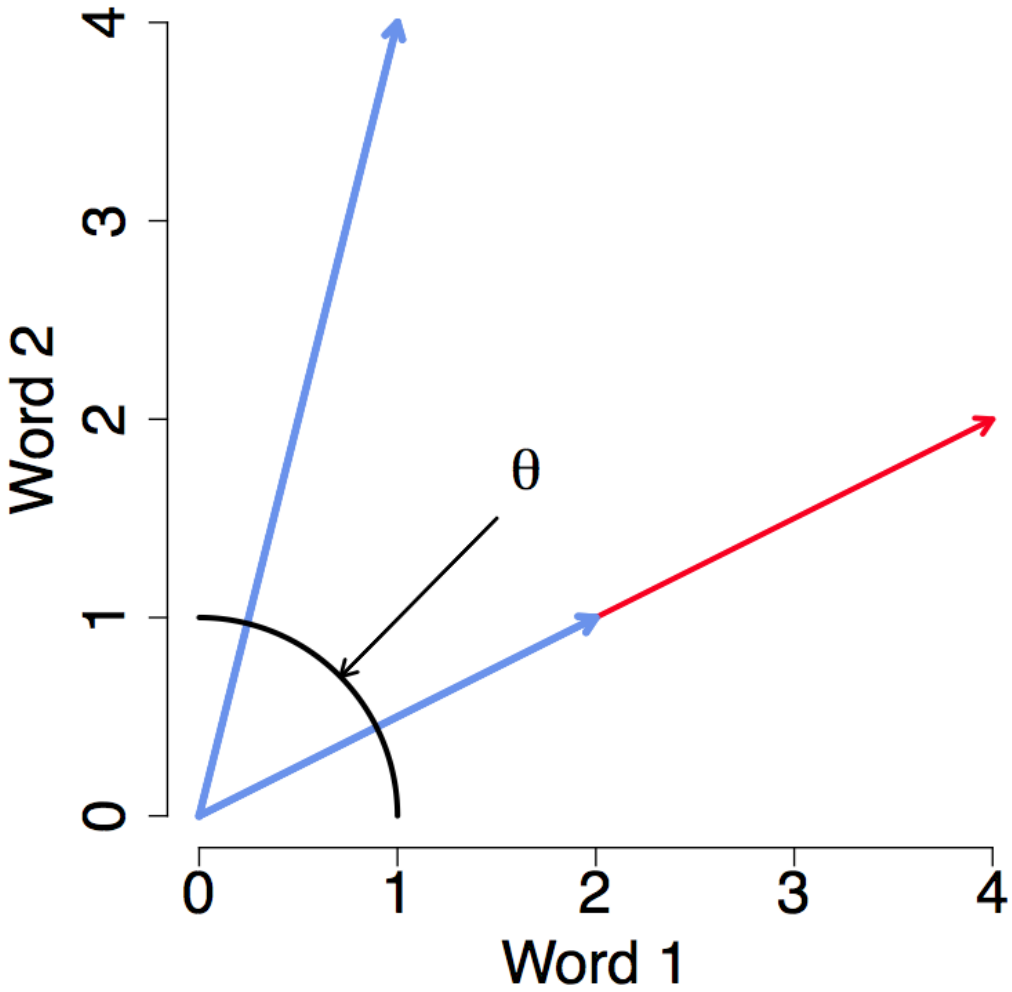


Problem(?): length dependent



Problem(?): length dependent

$$(4,2)'(1,4) = 12$$



Problem(?): length dependent

$$(4,2)'(1,4) = 12$$

$$a \cdot b = ||a|| \times ||b|| \times \cos\theta$$

Cosine Similarity

Cosine Similarity

$$\cos\theta = \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right)$$

Cosine Similarity

$$\cos\theta = \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right)$$

$$\frac{(4,2)}{||(4,2)||} = (0.89, 0.45)$$

Cosine Similarity

$$\cos\theta = \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right)$$

$$\frac{(4,2)}{||(4,2)||} = (0.89, 0.45)$$

$$\frac{(2,1)}{||(2,1)||} = (0.89, 0.45)$$

Cosine Similarity

$$\cos\theta = \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right)$$

$$\frac{(4,2)}{||(4,2)||} = (0.89, 0.45)$$

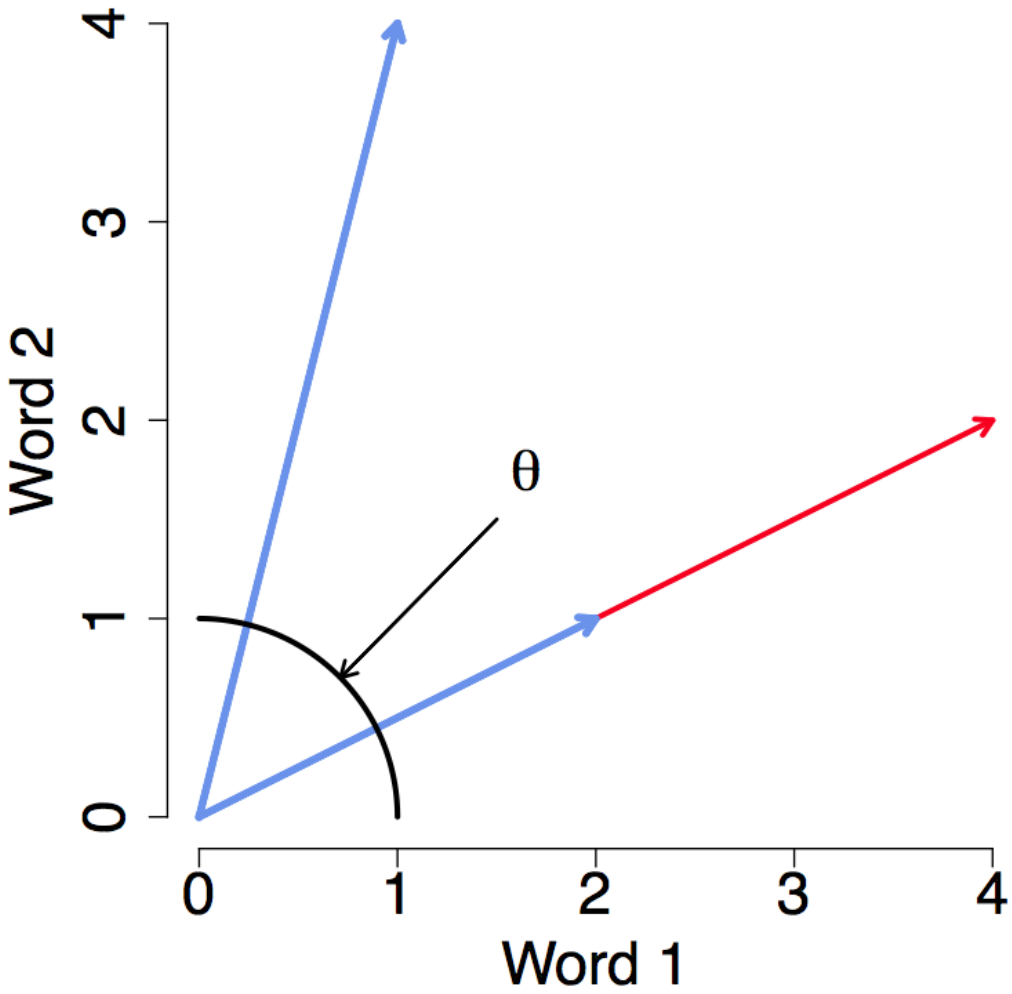
$$\frac{(2,1)}{||(2,1)||} = (0.89, 0.45)$$

$$\frac{(1,4)}{||(1,4)||} = (0.24, 0.97)$$

Cosine Similarity

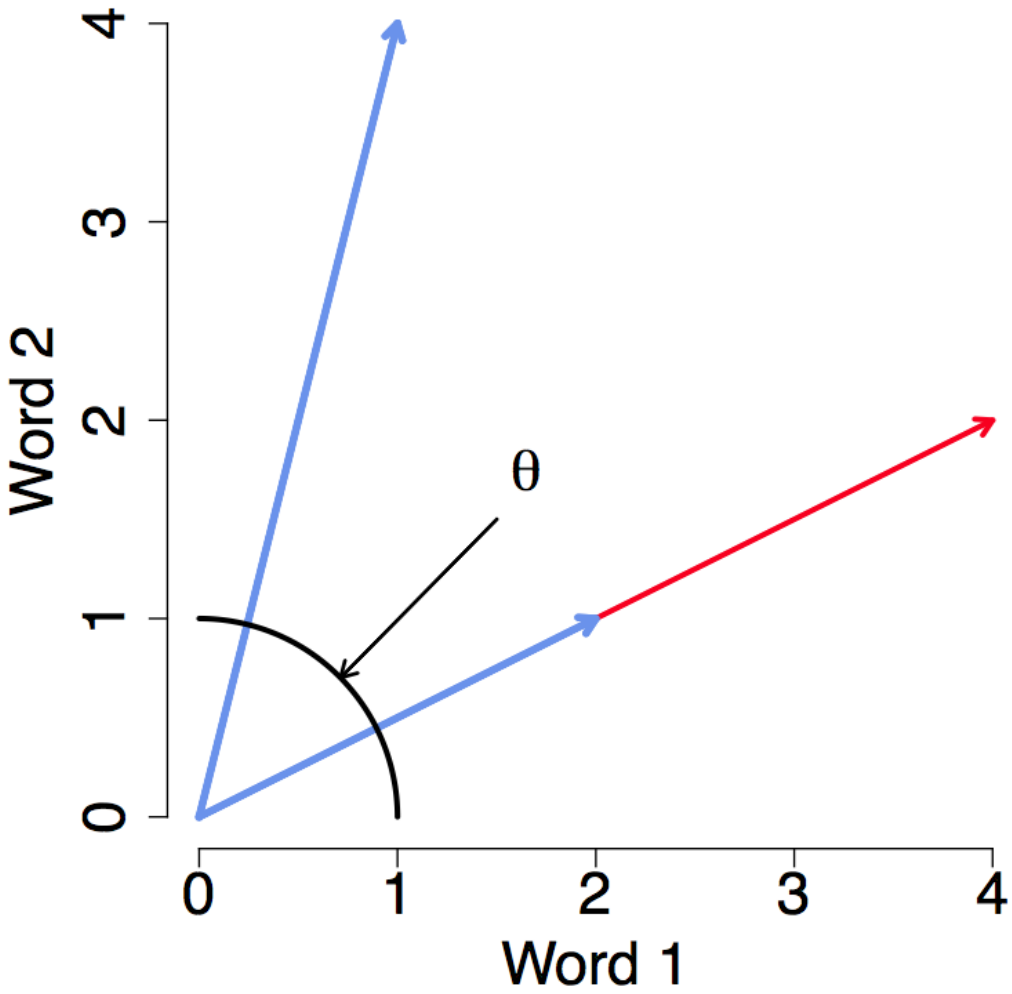
$$
\begin{aligned}
\cos\theta &= \left(\frac{a}{||a||}\right) \cdot \left(\frac{b}{||b||}\right) \\
\frac{(4,2)}{||(4,2)||} &= (0.89, 0.45) \\
\frac{(2,1)}{||(2,1)||} &= (0.89, 0.45) \\
\frac{(1,4)}{||(1,4)||} &= (0.24, 0.97) \\
(0.89, 0.45)'(0.24, 0.97) &= 0.65
\end{aligned}
$$

Cosine Similarity



$\cos\theta$: removes document length from similarity measure

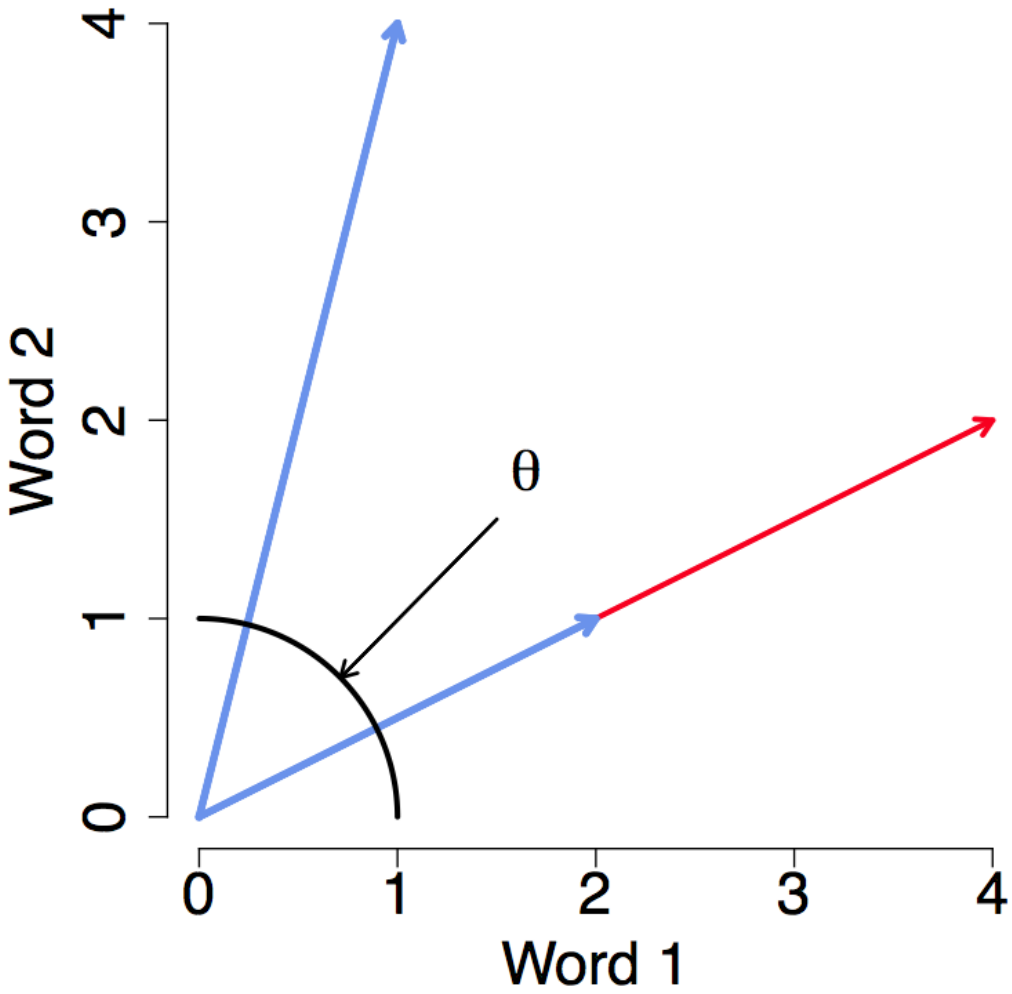
Cosine Similarity



$\cos\theta$: removes document length from similarity measure

Projects texts to unit length representation$\rightsquigarrow$ onto sphere

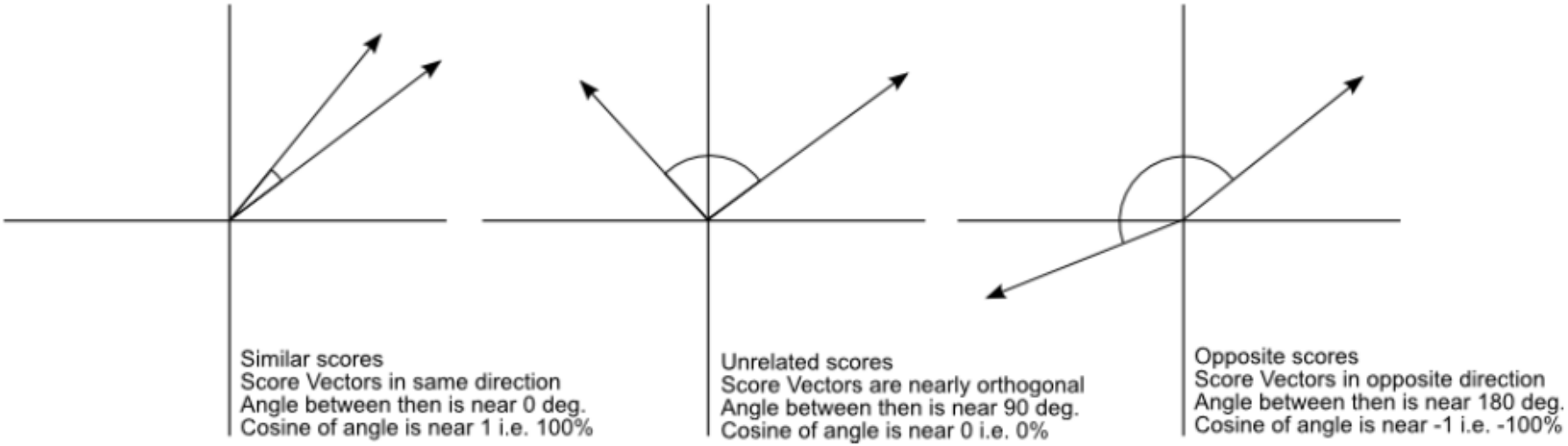
Cosine Similarity



$\cos\theta$: removes document length from similarity measure

Projects texts to unit length representation$\rightsquigarrow$ onto sphere

Cosine similarity illustrated



Example text

Hurricane Gilbert swept toward the Dominican Republic Sunday , and the Civil Defense alerted its heavily populated south coast to prepare for high **winds**, heavy **rains** and high seas.

The **storm** was approaching from the southeast with sustained **winds** of 75 mph gusting to 92 mph .

"There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly before midnight Saturday .

Cabral said residents of the province of Barahona should closely follow **Gilbert** 's movement .

An estimated 100,000 people live in the province, including 70,000 in the city of Barahona , about 125 miles west of Santo Domingo .

Tropical **Storm Gilbert** formed in the eastern Caribbean and strengthened into a **hurricane** Saturday night

The National **Hurricane** Center in Miami reported its position at 2a.m. Sunday at latitude 16.1 north , longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.

The National Weather Service in San Juan , Puerto Rico , said **Gilbert** was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the **storm**.

The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6p.m. Sunday.

Strong **winds** associated with the **Gilbert** brought coastal flooding , strong southeast **winds** and up to 12 feet to Puerto Rico 's south coast.

Example text: selected terms

- Document 1
  Gilbert: 3, hurricane: 2, rains: 1, storm: 2, winds: 2
- Document 2
  Gilbert: 2, hurricane: 1, rains: 0, storm: 1, winds: 2

Example text: cosine similarity in R

```
> toyDfm <- matrix(c(3,2,1,2,2, 2,1,0,1,2), nrow=2, byrow=TRUE)
> colnames(toyDfm) <- c("Gilbert", "hurricane", "rain", "storm", "winds")
> rownames(toyDfm) <- c("doc1", "doc2")
> toyDfm
     Gilbert hurricane rain storm winds
doc1       3         2    1     2     2
doc2       2         1    0     1     2
> simil(toyDfm, "cosine")
          doc1
doc2 0.9438798
```

Document length bias ilustrated

Example TDM

doc1	Two for tea and tea for two
doc2	Tea for me and tea for you
doc3	You for me and me for you

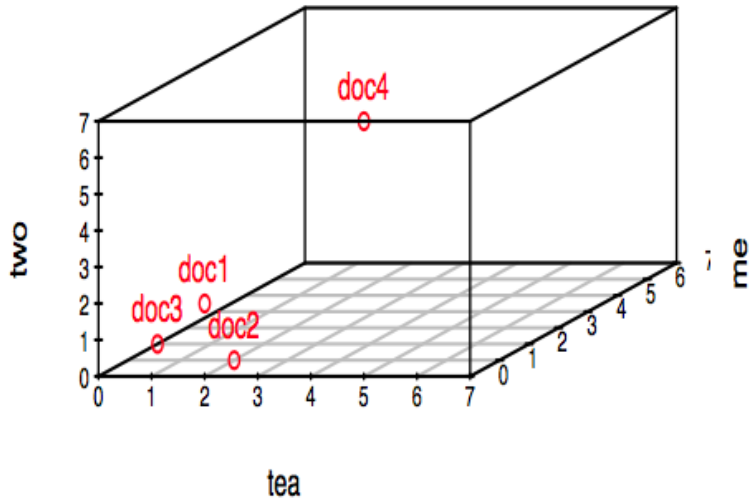
	two	tea	me	you
doc1	2	2	0	0
doc2	0	2	1	1
doc3	0	0	2	2

Document length bias ilustrated

- Measuring literal distance between documents in term space has problem:
- Documents with lots of terms will be further from origin...
- Documents with few terms closer to it...
- So we?ll find all short documents relatively similar...
- Even if they?re unrelated

Document length bias ilustrated

Point	tea	me	two
doc1	2	0	2
doc2	2	1	0
doc3	0	2	0
doc4	5	0	7

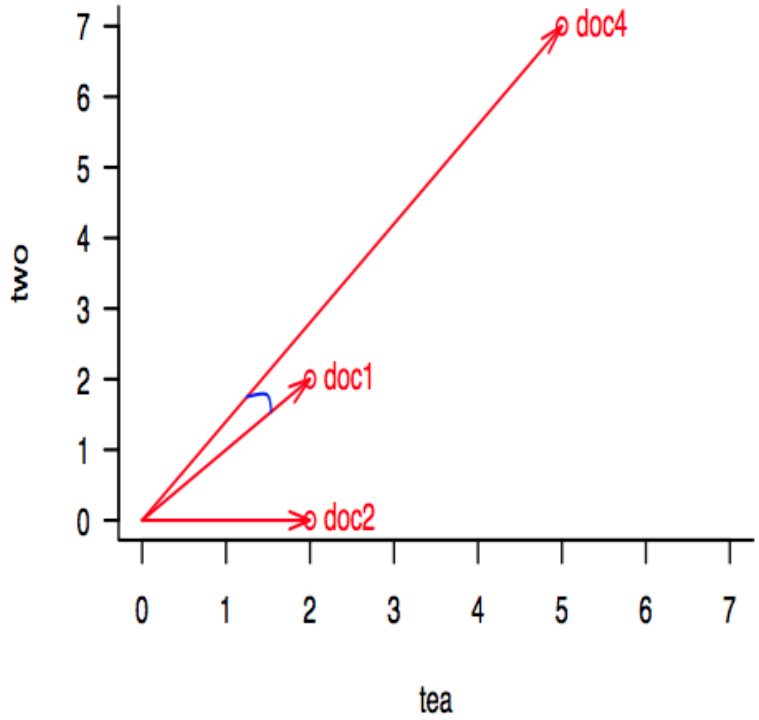


- Doc4, like Doc2, is all about "tea" and "two".
- But because it is longer, it is in a space by itself.

angular distance

to avoid length issue, we treat documents as vectors (lines from the origin in spce) and measure similarity by angle between vectors. Here, we see that Doc1 and Doc4 are indeed similar.

Point	tea	two
doc1	2	2
doc2	2	0
doc4	5	7



A different type of distance: Edit distances

- Edit distance refers to the number of operations required to transform one string into another
- Common edit distance: the Levenshtein distance
- Example: the Levenshtein distance between "kitten" and "sitting" is 3
  - kitten *rightarrow* sitten (substitution of "s" for "k")
  - sitten *rightarrow* sittin (substitution of "i" for "e")
  - sittin *rightarrow* sitting (insertion of "g" at the end).
- Not common, as at a textual level this is hard to implement possibly meaningless; great for string matching exercises.

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

$$n_j = \text{No. documents in which word } j \text{ occurs}$$

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

$$\begin{aligned} n_j &= \text{No. documents in which word } j \text{ occurs} \\ \text{idf}_j &= \log \frac{N}{n_j} \end{aligned}$$

Weighting Words: TF-IDF Weighting

What properties do words need to separate concepts?

- Used frequently
- But not too frequently

Ex. If all statements about OBL contain `Bin Laden` than this contributes nothing to similarity/dissimilarity measures

Inverse document frequency:

$$
\begin{aligned}
n_j &= \text{No. documents in which word } j \text{ occurs} \\
\text{idf}_j &= \log \frac{N}{n_j} \\
\mathbf{idf} &= (\text{idf}_1, \text{idf}_2, \ldots, \text{idf}_J)
\end{aligned}
$$

Weighting Words: TF-IDF Weighting

Why log ?

Weighting Words: TF-IDF Weighting

Why log ?

- Maximum at $n_j = 1$

Weighting Words: TF-IDF Weighting

Why log ?

- Maximum at $n_j = 1$
- Decreases at rate $\frac{1}{n_j} \Rightarrow$ diminishing "penalty" for more common use

Weighting Words: TF-IDF Weighting

Why log ?

- Maximum at $n_j = 1$
- Decreases at rate $\frac{1}{n_j} \Rightarrow$ diminishing "penalty" for more common use
- Other functional forms are fine, embed assumptions about penalization of common use

Weighting Words: TF-IDF

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} \quad = \quad (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

Weighting Words: TF-IDF

$$\begin{aligned}
\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} &= (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J) \\
\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} &= (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)
\end{aligned}$$

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product

$$\mathbf{X}_{i,\text{idf}} \cdot \mathbf{X}_{j,\text{idf}} = (\mathbf{X}_i \times \mathbf{idf})'(\mathbf{X}_j \times \mathbf{idf})$$

Weighting Words: TF-IDF

$$\mathbf{X}_{i,\text{idf}} \equiv \underbrace{\mathbf{X}_i}_{\text{tf}} \times \mathbf{idf} = (X_{i1} \times \text{idf}_1, X_{i2} \times \text{idf}_2, \ldots, X_{iJ} \times \text{idf}_J)$$

$$\mathbf{X}_{j,\text{idf}} \equiv \mathbf{X}_j \times \mathbf{idf} = (X_{j1} \times \text{idf}_1, X_{j2} \times \text{idf}_2, \ldots, X_{jJ} \times \text{idf}_J)$$

How Does This Matter For Measuring Similarity/Dissimilarity?

Inner Product

$$\begin{aligned}\mathbf{X}_{i,\text{idf}} \cdot \mathbf{X}_{j,\text{idf}} &= (\mathbf{X}_i \times \mathbf{idf})'(\mathbf{X}_j \times \mathbf{idf}) \\ &= (\text{idf}_1^2 \times X_{i1} \times X_{j1}) + (\text{idf}_2^2 \times X_{i2} \times X_{j2}) + \\ &\quad \ldots + (\text{idf}_J^2 \times X_{iJ} \times X_{jJ})\end{aligned}$$

Final Product

Applying some measure of distance, similarity (if symmetric) yields:

$$\mathbf{D} = \begin{pmatrix} 0 & d(1,2) & d(1,3) & \dots & d(1,N) \\ d(2,1) & 0 & d(2,3) & \dots & d(2,N) \\ d(3,1) & d(3,2) & 0 & \dots & d(3,N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(N,1) & d(N,2) & d(N,3) & \dots & 0 \end{pmatrix}$$

Lower Triangle contains unique information $N(N-1)/2$

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Paper does a lot. We're going to focus on

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Paper does a lot. We're going to focus on

- Today: Text representation and similarity calculation

Spirling and Indian Treaties

Spirling (2013): model Treaties between US and Native Americans
Why?

- American political development
- IR Theories of Treaties and Treaty Violations
- Comparative studies of indigenous/colonialist interaction
- Political Science question: how did Native Americans lose land so quickly?

Paper does a lot. We're going to focus on

- Today: Text representation and similarity calculation
- Tuesday: Projecting to low dimensional space

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Spirling and Indian Treaties

How do we preserve word order and semantic language?
After stemming, stopping, bag of wording:

- `Peace Between Us`
- `No Peace Between Us`

are identical.
Spirling uses complicated representation of texts to preserve word order$\rightsquigarrow$ broad application
`Peace Between Us`
`Analyzes K-substrings`

Kernel Trick

Kernel Trick

- Kernel Methods: Represent texts, measure similarity

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
    - Arthur gives all his money to Justin

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Uses Kernel methods to measure similarity

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Uses Kernel methods to measure similarity

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Uses Kernel methods to measure similarity

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Uses Kernel methods to measure similarity

Kernel Trick

- Kernel Methods: Represent texts, measure similarity simultaneously
- Compare only substrings in both documents (without explicitly quantifying entire documents)
- Problem solved:
  - Arthur gives all his money to Justin
  - Justin gives all his money to Arthur
  - Discard word order: same sentence Kernel : different sentences.

Uses Kernel methods to measure similarity