An Introduction to Analyzing Political Texts Part I

Arthur Spirling

New York University

November 15, 2019

Boring but important sanity check

Boring but important sanity check

https://github.com/ArthurSpirling/yale_text_course

Overview

Overview


since
race
stand
responsibility
parents
live
law
together
matter
south
need

Overview

- Descriptive inference:

Overview

- Descriptive inference: how to characterize text,

Overview

- Descriptive inference: how to characterize text, vector space model,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity, style,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity, style, bursts.

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity, style, bursts. Uncertainty for texts.

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity, style, bursts. Uncertainty for texts.
- Basic supervised techniques: sentiment,

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity, style, bursts. Uncertainty for texts.

- Basic supervised techniques: sentiment, scaling.

Overview

- Descriptive inference: how to characterize text, vector space model, bag-of-words, (dis)similarity measures, keywords in context, complexity, style, bursts. Uncertainty for texts.

- Basic supervised techniques: sentiment, scaling.

- Basic unsupervised techniques: topics.

Quantitative vs Qualitative

Quantitative vs Qualitative

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.

Quantitative vs Qualitative

- For most of its history text analysis was qualitative.
- Partly still is:

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,

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

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'

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.

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)

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

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*.

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*.

$\rightarrow$ something which we cannot observe directly but which we can make inferences about from things we can observe.

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.

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*.

$\rightarrow$ 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,

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,

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.

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.

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...

And. . .

And. . .

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

And. . .

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

author ‘what does this Senator prioritize?’,

And. . .

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

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

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?',

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?’

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?'

We need to think carefully about. . .

We need to think carefully about...

- the appropriate population and sample

We need to think carefully about. . .

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

We need to think carefully about...

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

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

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

We need to think carefully about...

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

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

$\rightarrow$ 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.

We need to think carefully about...

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

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

$\rightarrow$ 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.

$\rightarrow$ statistical choices

We need to think carefully about...

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

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

$\rightarrow$ 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.

$\rightarrow$ statistical choices

- what we can infer about the latent variables.

We need to think carefully about. . .

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

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

$\rightarrow$ 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.

$\rightarrow$ statistical choices

- what we can infer about the latent variables.

$\rightarrow$ comparing, testing, validating.

In general, we will. . .

In general, we will. . .

Get Texts

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…

In general, we will. . .

Get Texts → Document Term Matrix

```
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…
```

In general, we will...

Get Texts → Document Term Matrix

```
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…
```

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

In general, we will. . .

Get Texts → Document Term Matrix → Operate

```
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…
```

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

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

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

→ Operate

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

...

In general, we will...

Get Texts → Document Term Matrix → Operate → Inference

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

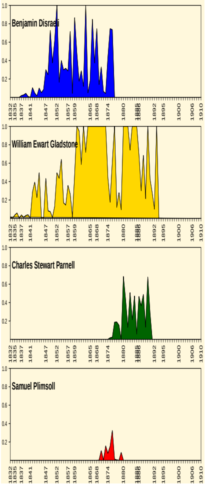
> Order. The Minister must be allowed to reply without interruption.

> 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…

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

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



I. Defining the Corpus

I. Defining the Corpus

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

I. Defining the Corpus

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

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

I. Defining the Corpus

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

$\longrightarrow$ 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.

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.

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.

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)

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)

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,

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,

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,

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

Sampling

The corpus is made up of the documents within it,

Sampling

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

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.

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,

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%.

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,

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,

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,

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.

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.

$\longrightarrow$ depending on your philosophical position,

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.

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.

$\longrightarrow$ 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.

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.

$\longrightarrow$ 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:

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.

II. Reducing Complexity

II. Reducing Complexity

- language is extraordinarily complex,

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

$\rightarrow$ makes the modeling problem much more tractable.

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

$\rightarrow$ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences,

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

$\rightarrow$ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes,

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

→ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes, and the fit of our models.

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

$\rightarrow$ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes, and the fit of our models.

NB inevitably, the degree to which one simplifies is dependent on the particular task at hand.

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

$\rightarrow$ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes, and the fit of our models.

NB inevitably, the degree to which one simplifies is dependent on the particular task at hand.

$\rightarrow$ there is no 'one best way' to go from texts to numeric data.

II. Reducing Complexity

- language is extraordinarily complex, and involves great subtlety and nuanced interpretation.

but remarkably, we can do very well by simplifying, and representing documents as straightforward mathematical objects.

→ makes the modeling problem much more tractable.

by 'do very well', we mean that much more complicated representations add (almost) nothing to the quality of our inferences, our ability to predict outcomes, and the fit of our models.

NB inevitably, the degree to which one simplifies is dependent on the particular task at hand.

→ there is no 'one best way' to go from texts to numeric data. Good idea to check sensitivity.

From Texts to Numeric Data

From Texts to Numeric Data

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

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.

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.

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.

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.

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.

‘superfluous’ material: control characters and punctuation

‘superfluous’ material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`,

‘superfluous’ material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`, do not connote much that is of substantive importance.

‘superfluous’ material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`, do not connote much that is of substantive importance.

$\rightarrow$ remove them.

‘superfluous’ material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`, do not connote much that is of substantive importance.

→ remove them. Same for underlining or **emboldening**.

‘superfluous’ material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`, do not connote much that is of substantive importance.

→ remove them. Same for underlining or **emboldening**.

- punctuation may also be unhelpful

'superfluous' material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`, do not connote much that is of substantive importance.

→ remove them. Same for underlining or **emboldening**.

- punctuation may also be unhelpful

  are `wash`, `wash.`, `wash,`, `wash)` really different words?

‘superfluous’ material: control characters and punctuation

- generally think control characters—non-printing, but cause the document to look different—like `\n`, do not connote much that is of substantive importance.

→ remove them. Same for <u>underlining</u> or **emboldening**.

- punctuation may also be unhelpful

  are `wash`, `wash.`, `wash,`, `wash)` really different words?

→ convert everything to whitespace (?)

Quick Note on Terminology

Quick Note on Terminology

a type is a unique sequence of characters that are grouped together in some meaningful way.

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),

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,

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.

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`’

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`”,

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,

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).

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.

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

Tokens and tokenization

The text is now ‘clean’,

Tokens and tokenization

The text is now 'clean', and we want to pull out the meaningful subunits

Tokens and tokenization

The text is now ‘clean’, and we want to pull out the meaningful subunits—the tokens.

Tokens and tokenization

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

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,

Tokens and tokenization

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

$\rightarrow$ usually the tokens are words, but might include numbers or punctuation too.

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.

Tokens and tokenization

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

$\rightarrow$ 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

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`'

Removing Stop Words

Removing Stop Words

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

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’.

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.

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,

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.

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.

Removing Stop Words

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

$\rightarrow$ 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,

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

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

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

Tagging

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

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).

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

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.

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,

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.

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.

Stemming and Lemmatization

Stemming and Lemmatization

Documents may use different forms of words

Stemming and Lemmatization

Documents may use different forms of words (‘`jumped`’, ‘`jumping`’, ‘`jump`’),

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.

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 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.

$\rightarrow$ 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.

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.

$\rightarrow$ 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').

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 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`'.

Snowball stemmer examples

Snowball stemmer examples

Original Word		Stemmed Word
abolish	$\mapsto$	`abolish`
abolished	$\mapsto$	`abolish`
abolishing	$\mapsto$	`abolish`
abolition	$\mapsto$	`abolit`

Snowball stemmer 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`

Snowball stemmer 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.

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

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

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.

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.

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.

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 Don't Care about Word Order

We have now pre-processed our texts.

We Don't Care about Word Order

We have now pre-processed our texts.

Generally,

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.

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.

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.

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

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 $\boxed{\text{bag-of-words}}$ (BOW).

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 $\boxed{\text{bag-of-words}}$ (BOW).

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

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 $\boxed{\text{bag-of-words}}$ (BOW).

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

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 $\boxed{\text{bag-of-words}}$ (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.`"

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 $\boxed{\text{bag-of-words}}$ (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"`

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 $\boxed{\text{bag-of-words}}$ (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`"

III. Vector Space Model

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line,

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions:

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions: one corresponds to how '`Bob`'-ish it is, one corresponds to how '`goes`'-ish it is, one corresponds to how '`home`'-ish it is.

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions: one corresponds to how '`Bob`'-ish it is, one corresponds to how '`goes`'-ish it is, one corresponds to how '`home`'-ish it is.

Features will typically be the *n*-gram (mostly unigram) frequencies of the tokens in the document,

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions: one corresponds to how '`Bob`'-ish it is, one corresponds to how '`goes`'-ish it is, one corresponds to how '`home`'-ish it is.

Features will typically be the *n*-gram (mostly unigram) frequencies of the tokens in the document, or some function of those frequencies.

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions: one corresponds to how '`Bob`'-ish it is, one corresponds to how '`goes`'-ish it is, one corresponds to how '`home`'-ish it is.

Features will typically be the *n*-gram (mostly unigram) frequencies of the tokens in the document, or some function of those frequencies.

e.g. '`the cat sat on the mat`' becomes `(2,1,1,1,1)`

III. Vector Space Model

We can think about a document as being a collection of W features (tokens, words etc)

if each feature can be placed on the real line, then the document can be thought of as a point $\mathbb{R}^W$.

e.g. "`Bob goes home`" can be thought of a vector in 3 dimensions: one corresponds to how '`Bob`'-ish it is, one corresponds to how '`goes`'-ish it is, one corresponds to how '`home`'-ish it is.

Features will typically be the *n*-gram (mostly unigram) frequencies of the tokens in the document, or some function of those frequencies.

e.g. '`the cat sat on the mat`' becomes `(2,1,1,1,1)` if we define the dimensions as `(the, cat, sat, on, mat)` and use simple counts.

Notation and Terminology

Notation and Terminology

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

Notation and Terminology

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

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

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

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,

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...

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).

Notation and Terminology

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

$w = 1, \dots, W$ 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).

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...

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

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

partial DTM from Roosevelt's Inaugural Addresses

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
```

→ could re-weight in various ways, but will leave as raw counts for now.

partial TDM from Roosevelt's Inaugural Addresses

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
```

$\rightarrow$ could re-weight in various ways, but will leave as raw counts for now.

Distance Metrics and Measures

Comparing Texts: Distance

Comparing Texts: Distance

Recall that the vector space model represents a document as a point in the feature space.

Comparing Texts: Distance

Recall that the vector space model represents a document as a point in the feature space.

i.e. $\mathbf{y}_d \in \mathbb{R}^W$ is a representation of document d.

Comparing Texts: Distance

Recall that the vector space model represents a document as a point in the feature space.

i.e. $\mathbf{y}_d \in \mathbb{R}^W$ is a representation of document d.

q how 'far' is that document from some other document (in the same space)?

Comparing Texts: Distance

Recall that the vector space model represents a document as a point in the feature space.

i.e. $\mathbf{y}_d \in \mathbb{R}^W$ is a representation of document d.

q how 'far' is that document from some other document (in the same space)?

$\rightarrow$ tells us about similarity of documents

Comparing Texts: Distance

Recall that the vector space model represents a document as a point in the feature space.

i.e. $\mathbf{y}_d \in \mathbb{R}^W$ is a representation of document d.

q how 'far' is that document from some other document (in the same space)?

$\rightarrow$ tells us about similarity of documents

and is typically required for application of multivariate techniques, anyway

Comparing Texts: Distance

Recall that the vector space model represents a document as a point in the feature space.

i.e. $\mathbf{y}_d \in \mathbb{R}^W$ is a representation of document d.

q how 'far' is that document from some other document (in the same space)?

$\rightarrow$ tells us about similarity of documents

and is typically required for application of multivariate techniques, anyway

e.g. principal components analysis operates on distance matrix.

Euclidean Distance

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

$$\boxed{\|\mathbf{y}_i - \mathbf{y}_j\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)}} = \sqrt{\sum (\mathbf{y}_i - \mathbf{y}_j)^2}$$

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

$$\boxed{\|\mathbf{y}_i - \mathbf{y}_j\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)}} = \sqrt{\sum (\mathbf{y}_i - \mathbf{y}_j)^2}$$

e.g. $\mathbf{y}_i = [0.00, 0.00, 1.38, 1.52, 0.00]$ and $\mathbf{y}_j = [0.00, 2.13, 3.24, 0.01, 0.06]$

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

$$\boxed{\|\mathbf{y}_i - \mathbf{y}_j\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)}} = \sqrt{\sum (\mathbf{y}_i - \mathbf{y}_j)^2}$$

e.g. $\mathbf{y}_i = [0.00, 0.00, 1.38, 1.52, 0.00]$ and $\mathbf{y}_j = [0.00, 2.13, 3.24, 0.01, 0.06]$

well $(\mathbf{y}_i - \mathbf{y}_j) = [0.00, -2.13, -1.86, 1.51, -0.06]$

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

$$\boxed{\|\mathbf{y}_i - \mathbf{y}_j\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)}} = \sqrt{\sum (\mathbf{y}_i - \mathbf{y}_j)^2}$$

e.g. $\mathbf{y}_i = [0.00, 0.00, 1.38, 1.52, 0.00]$ and $\mathbf{y}_j = [0.00, 2.13, 3.24, 0.01, 0.06]$

well $(\mathbf{y}_i - \mathbf{y}_j) = [0.00, -2.13, -1.86, 1.51, -0.06]$

and $(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j) =$

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

$$\boxed{\|\mathbf{y}_i - \mathbf{y}_j\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)}} = \sqrt{\sum (\mathbf{y}_i - \mathbf{y}_j)^2}$$

e.g. $\mathbf{y}_i = [0.00, 0.00, 1.38, 1.52, 0.00]$ and $\mathbf{y}_j = [0.00, 2.13, 3.24, 0.01, 0.06]$

well $(\mathbf{y}_i - \mathbf{y}_j) = [0.00, -2.13, -1.86, 1.51, -0.06]$

and $(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j) = (0 \times 0) + (-2.13 \times -2.13) + (-1.86 \times -1.86) + (1.51 \times 1.51) + (-0.06 \times -0.06) = 10.2802$

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space.

Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

$$\boxed{\|\mathbf{y}_i - \mathbf{y}_j\| = \sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)}} = \sqrt{\sum (\mathbf{y}_i - \mathbf{y}_j)^2}$$

e.g. $\mathbf{y}_i = [0.00, 0.00, 1.38, 1.52, 0.00]$ and $\mathbf{y}_j = [0.00, 2.13, 3.24, 0.01, 0.06]$

well $(\mathbf{y}_i - \mathbf{y}_j) = [0.00, -2.13, -1.86, 1.51, -0.06]$

and $(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j) = (0 \times 0) + (-2.13 \times -2.13) + (-1.86 \times -1.86) + (1.51 \times 1.51) + (-0.06 \times -0.06) = 10.2802$

and $\sqrt{(\mathbf{y}_i - \mathbf{y}_j) \cdot (\mathbf{y}_i - \mathbf{y}_j)} = 3.206275$

larger distances imply lower similarity.

Partner exercise

Partner exercise

1. consider three documents in term frequency space:

   $[5, 4, 3]$
   $[50, 40, 30]$
   $[3, 3, 4]$

   Which documents will Euclidean distance place closest together?

Partner exercise

1. consider three documents in term frequency space:

   $[5, 4, 3]$
   $[50, 40, 30]$
   $[3, 3, 4]$

   Which documents will Euclidean distance place closest together? Why?

Partner exercise

1. consider three documents in term frequency space:

   $[5, 4, 3]$
   $[50, 40, 30]$
   $[3, 3, 4]$

   Which documents will Euclidean distance place closest together? Why?

2. now suppose the second document is simply the first document copied 10 times.

Partner exercise

1. consider three documents in term frequency space:

   $[5, 4, 3]$
   $[50, 40, 30]$
   $[3, 3, 4]$

   Which documents will Euclidean distance place closest together? Why?

2. now suppose the second document is simply the first document copied 10 times. Does the Euclidean distance seem intuitively suitable given how similar you know the content to be?

Better Approach

Better Approach

Euclidean distance rewards magnitude, rather than direction.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. doesn't reward being close in relative use of terms.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. doesn't reward being close in relative use of terms. Instead, rewards documents that are similarly 'far' from the origin.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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,

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

$\rightarrow$ divide out each of the components (the documents) by their length:

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

$\rightarrow$ divide out each of the components (the documents) by their length: the L^2 norm, $||\mathbf{y}_i|| = \sqrt{\sum w^2}$,

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

$\rightarrow$ divide out each of the components (the documents) by their length: the L^2 norm, $||\mathbf{y}_i|| = \sqrt{\sum w^2}$, where w refers to the (weighted) frequency of a feature in the document vector.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

$\rightarrow$ divide out each of the components (the documents) by their length: the L^2 norm, $||\mathbf{y}_i|| = \sqrt{\sum w^2}$, where w refers to the (weighted) frequency of a feature in the document vector.

so when the document has generally high term frequencies (because it is longer),

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

$\rightarrow$ divide out each of the components (the documents) by their length: the L^2 norm, $||\mathbf{y}_i|| = \sqrt{\sum w^2}$, where w refers to the (weighted) frequency of a feature in the document vector.

so when the document has generally high term frequencies (because it is longer), w^2 will be larger,

Better Approach

Euclidean distance rewards magnitude, rather than direction.

i.e. 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 relatively similar uses of terms.

→ divide out each of the components (the documents) by their length: the L^2 norm, $||\mathbf{y}_i|| = \sqrt{\sum w^2}$, where w refers to the (weighted) frequency of a feature in the document vector.

so when the document has generally high term frequencies (because it is longer), w^2 will be larger, which makes $||\mathbf{y}_i||$ larger.

Cosine Similarity

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \; \|\mathbf{y}_j\|}}$$

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \, \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \, \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

If $\mathbf{y}_i$ and $\mathbf{y}_j$ are vectors, c_{ij} is the cosine of the angle between them.

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \ \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

If $\mathbf{y}_i$ and $\mathbf{y}_j$ are vectors, c_{ij} is the cosine of the angle between them.

and document length is controlled for.

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \; \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

If $\mathbf{y}_i$ and $\mathbf{y}_j$ are vectors, c_{ij} is the cosine of the angle between them.

and document length is controlled for.

so intuitively,

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \ \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

If $\mathbf{y}_i$ and $\mathbf{y}_j$ are vectors, c_{ij} is the cosine of the angle between them.

and document length is controlled for.

so intuitively, cosine similarity captures some notion of relative 'direction' (e.g. style or topics in the document)

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \, \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

If $\mathbf{y}_i$ and $\mathbf{y}_j$ are vectors, c_{ij} is the cosine of the angle between them.

and document length is controlled for.

so intuitively, cosine similarity captures some notion of relative 'direction' (e.g. style or topics in the document) rather than 'magnitude' (distance from origin).

Cosine Similarity

$$c_{ij} = \boxed{\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{\|\mathbf{y}_i\| \ \|\mathbf{y}_j\|}}$$

$\rightarrow$ we have a measure of similarity, which (since $\mathbf{y}_i$ and $\mathbf{y}_j$ are non-negative) must be between 0 and 1.

If $\mathbf{y}_i$ and $\mathbf{y}_j$ are vectors, c_{ij} is the cosine of the angle between them.

and document length is controlled for.

so intuitively, cosine similarity captures some notion of relative 'direction' (e.g. style or topics in the document) rather than 'magnitude' (distance from origin). Is the Pearson correlation between two vectors that have been demeaned.

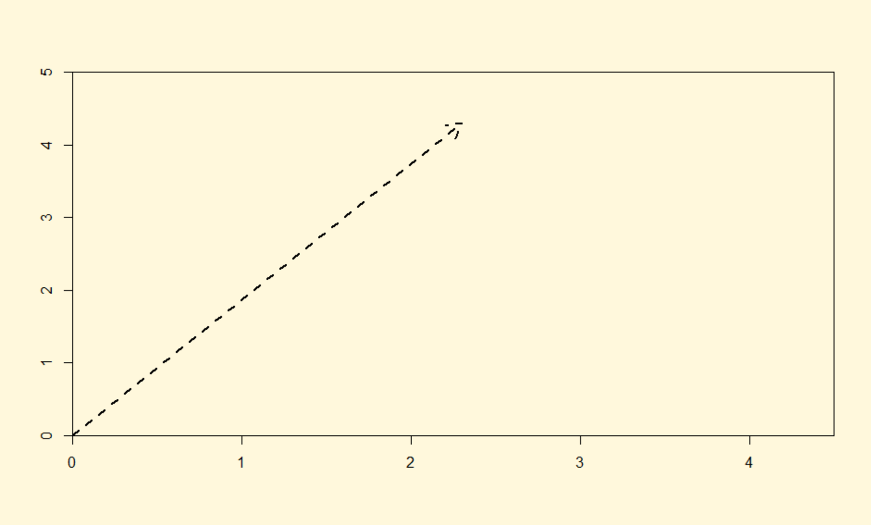
Graphically

Graphically

$$\mathbf{y_i} = [2.3, 4.3];\ \mathbf{y_j} = [3.9, 2.1]$$

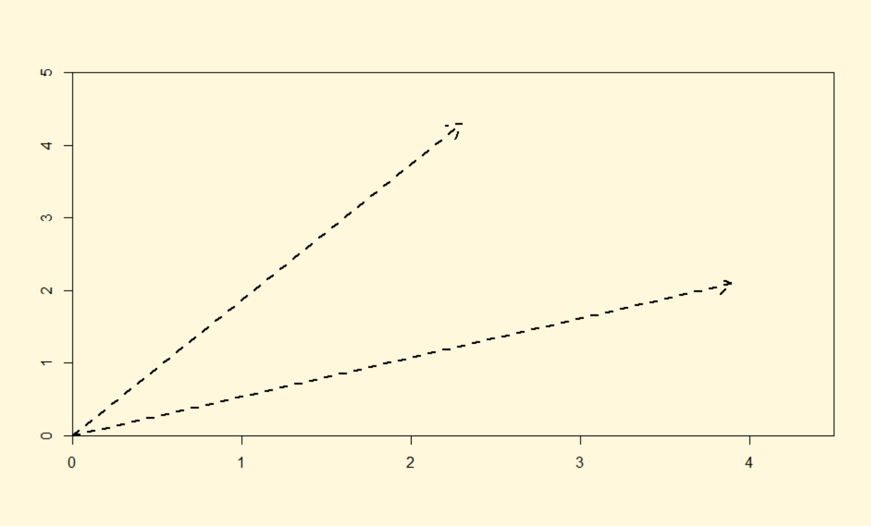
Graphically

$$\mathbf{y_i} = [2.3, 4.3]; \ \mathbf{y_j} = [3.9, 2.1]$$



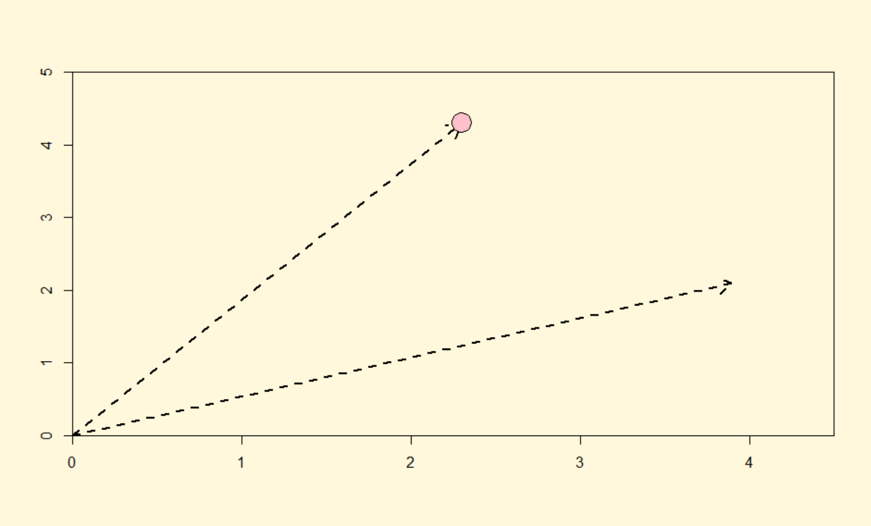
Graphically

$\mathbf{y_i} = [2.3, 4.3]$; $\mathbf{y_j} = [3.9, 2.1]$



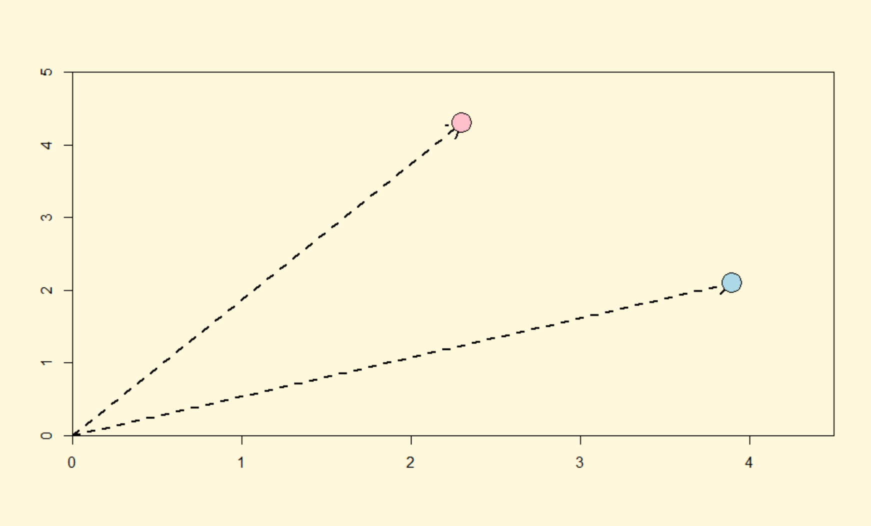
Graphically

$\mathbf{y_i} = [2.3, 4.3]$; $\mathbf{y_j} = [3.9, 2.1]$



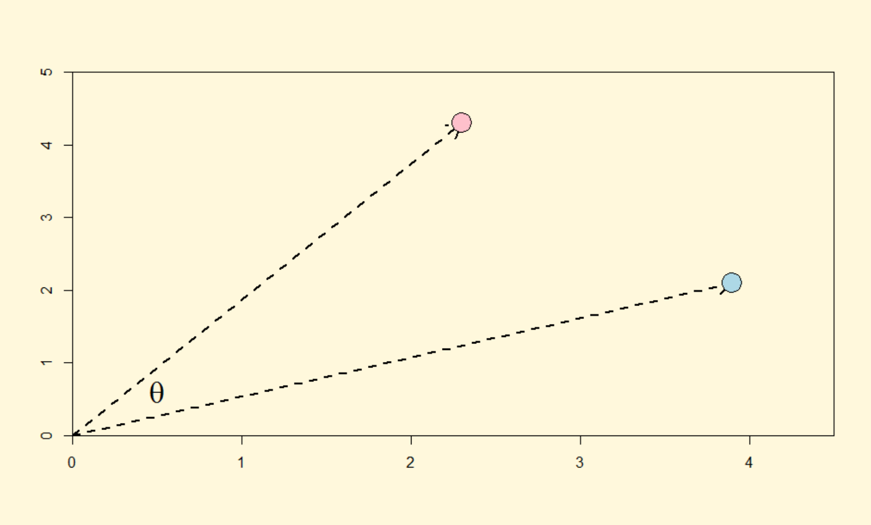
Graphically

$\mathbf{y_i} = [2.3, 4.3]$; $\mathbf{y_j} = [3.9, 2.1]$



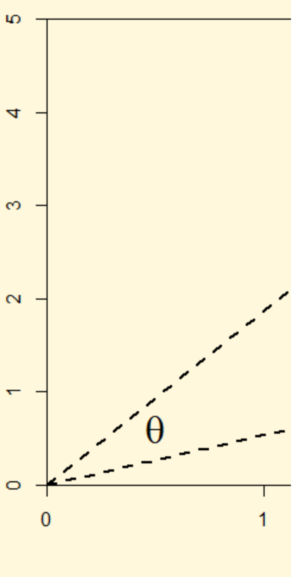
Graphically

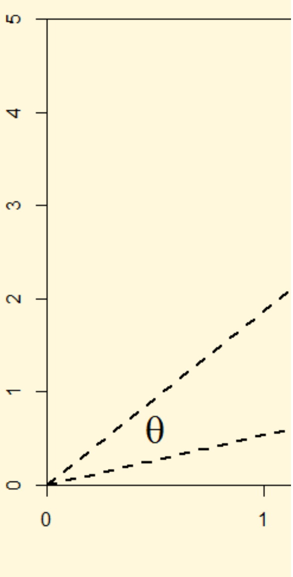
$$\mathbf{y_i} = [2.3, 4.3];\ \mathbf{y_j} = [3.9, 2.1]$$



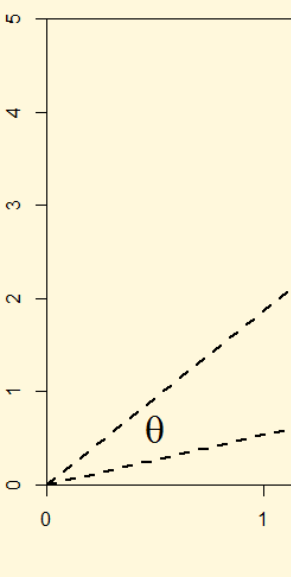
Algebra

Algebra



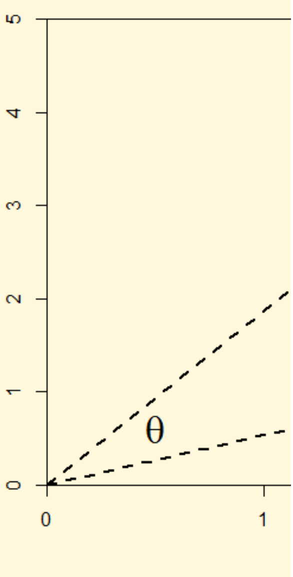


know dot product of vectors:



know dot product of vectors:
$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$

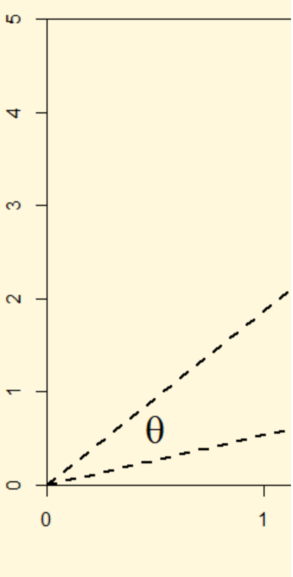
Algebra



know dot product of vectors:
$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i|| ||\mathbf{y}_j|| \cos\theta$

then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i|| ||\mathbf{y}_j||}$

Algebra

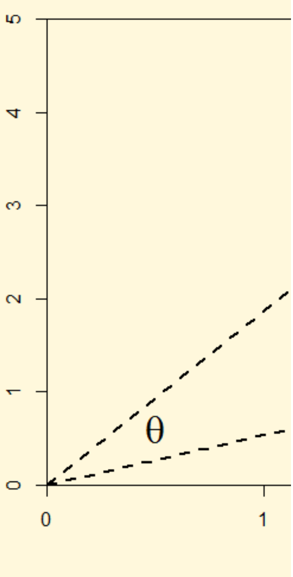


know dot product of vectors: $\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i|| ||\mathbf{y}_j|| \cos\theta$

then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i|| ||\mathbf{y}_j||}$

and $\theta = \arccos\left(\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i|| ||\mathbf{y}_j||}\right)$.

Algebra



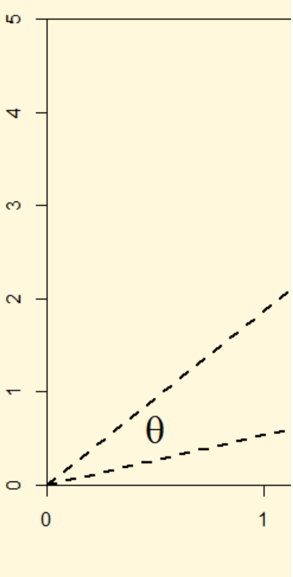
know dot product of vectors:
$$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i|| \, ||\mathbf{y}_j|| \cos\theta$$

then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i|| \, ||\mathbf{y}_j||}$

and $\theta = \arccos\left(\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i|| \, ||\mathbf{y}_j||}\right).$

so $\theta = \arccos\left(\frac{18}{21.62}\right)$

Algebra



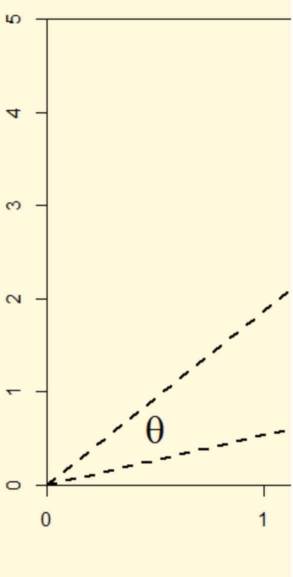
know dot product of vectors:
$$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$$

then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}$

and $\theta = \arccos\left(\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}\right).$

so $\theta = \arccos\left(\frac{18}{21.62}\right) = 0.58$

Algebra



know dot product of vectors:
$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$

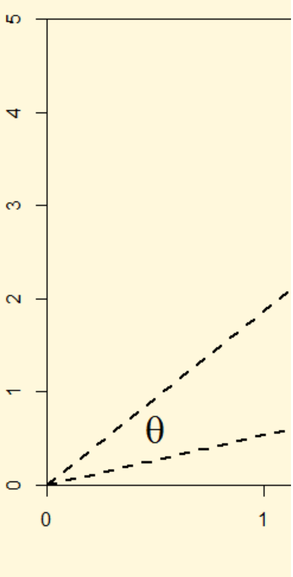
then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}$

and $\theta = \arccos\left(\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}\right).$

so $\theta = \arccos\left(\frac{18}{21.62}\right) = 0.58$

$\rightarrow 0.58 \times \frac{180}{\pi} = 33.63^\circ.$

Algebra



know dot product of vectors:

$$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$$

then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}$

and $\theta = \arccos\left(\frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}\right)$.

so $\theta = \arccos\left(\frac{18}{21.62}\right) = 0.58$

$$\rightarrow 0.58 \times \frac{180}{\pi} = 33.63^\circ.$$

Looks about right.

1983 General Election Manifestos

1983 General Election Manifestos

1983 General Election Manifestos

- Labour manifesto as 'longest suicide note in history'

1983 General Election Manifestos

- Labour manifesto as 'longest suicide note in history'
- unilateral nuclear disarmament, withdrawal from the EEC, abolition of the Lords, re-nationalisation

1983 General Election Manifestos

- Labour manifesto as 'longest suicide note in history'
- unilateral nuclear disarmament, withdrawal from the EEC, abolition of the Lords, re-nationalisation

1983 General Election Manifestos

- Labour manifesto as 'longest suicide note in history'
- unilateral nuclear disarmament, withdrawal from the EEC, abolition of the Lords, re-nationalisation

- Conservative manifesto promised trade union curbs, deflation etc.

1983 General Election Manifestos

- Labour manifesto as 'longest suicide note in history'
- unilateral nuclear disarmament, withdrawal from the EEC, abolition of the Lords, re-nationalisation

- Conservative manifesto promised trade union curbs, deflation etc.

$$\boxed{c_{ij} \approx 0.70}$$

1997 General Election Manifestos

1997 General Election Manifestos

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

- 'New Labour' and 'Third Way'

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

- 'New Labour' and 'Third Way'
- committed to Conservative spending plans (for next two years),

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

- ‘New Labour’ and ‘Third Way’
- committed to Conservative spending plans (for next two years), no income tax rises.

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

- 'New Labour' and 'Third Way'
- committed to Conservative spending plans (for next two years), no income tax rises.

$\boxed{c_{ij} \approx 0.90}$

Descriptive Statistics: Key Words

Key Words in Context

Key Words in Context

In Information Retrieval it is often extremely helpful to know how and where a particular token of interest appears,

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.

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,

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.

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,

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.

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.

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.

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

Example: ‘democratic’ and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

Example: '`democratic`' and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

1867 House of Commons considers extending suffrage to urban working class men,

Example: 'democratic' and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

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

Example: ‘democratic’ and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

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

→ represents approximate doubling of electorate.

Example: 'democratic' and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

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.

Example: 'democratic' and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

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.

Example: ‘`democratic`’ and the Second Reform Act

DERBY, 1867. DIZZY WINS WITH "REFORM BILL."

A LEAP IN THE DARK.

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.

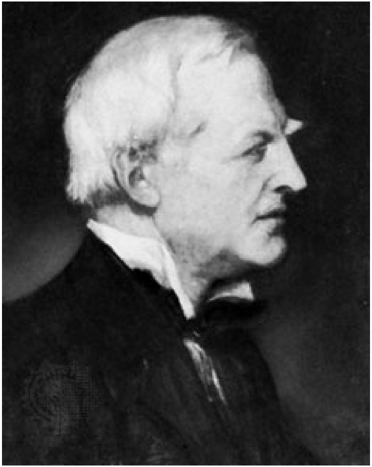
Detail: s267738.txt

Detail: s267738.txt

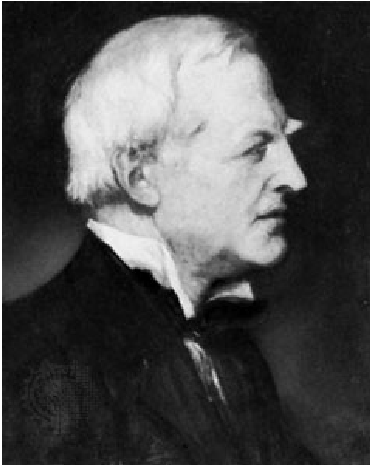
preword	word	postword
swept away in some further	democratic	change. And it is for
throne. When you get a	democratic	basis for your institutions, you
differences between ourselves and other	democratic	legislatures? Where is the democratic
democratic legislatures? Where is the	democratic	legislature which 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

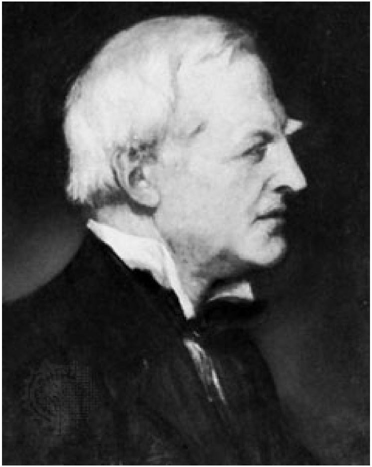
The Original Speaker and Speech



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.



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.

Partner Exercise

Partner Exercise

The context of key words is especially important when comparing usage across time and space.

Partner Exercise

The context of key words is especially important when comparing usage across time and space.

Suppose you were studying the history of entertainment technology. Consider the key word '`wireless`'. How has the frequency of this term changed over time? How has the context changed?

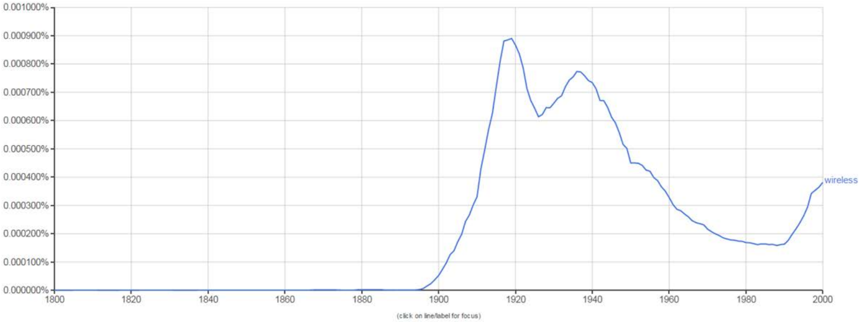
Partner Exercise

The context of key words is especially important when comparing usage across time and space.

Suppose you were studying the history of entertainment technology. Consider the key word '`wireless`'. How has the frequency of this term changed over time? How has the context changed?

Give an example of a political key word that might appear in a different *context* if we study the US vs some other country.

Use of 'Wireless'



Descriptive Statistics: Diversity and Complexity

Lexical Diversity

Lexical Diversity

Recall that the elementary components of a text are called tokens.

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.

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.

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,

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.

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:

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:

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

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.

Partner Exercise

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 it's too small. I respectfully disagree.

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 it's too small. I respectfully disagree.

Compare these two speech segments. Which is more difficult to understand?

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 it's 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

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability':

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.

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

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

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

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.

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).

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

Readability Guidelines

in practice,

Readability Guidelines

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

Readability Guidelines

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

However. . .

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

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
33	mean political science article; judicial opinion

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)

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

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*

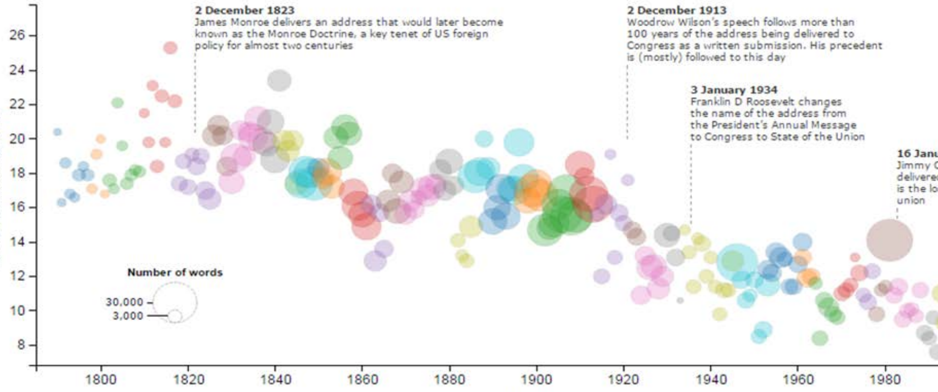
Examples

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.

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

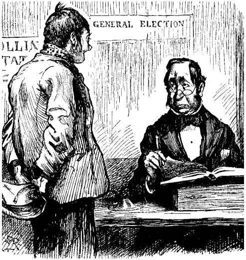
Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate



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...

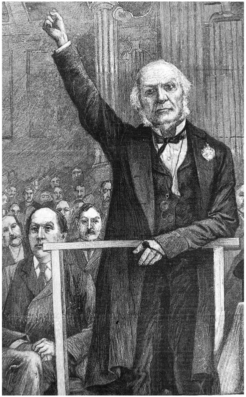
Leaders and their incentives

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



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?

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:

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,

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

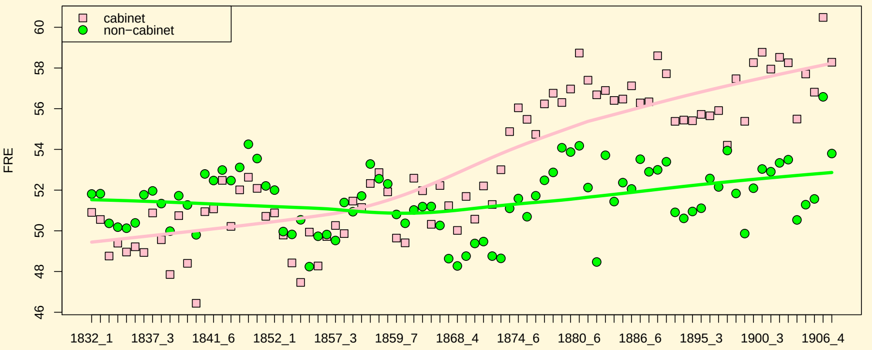
↓ local, clientalistic appeals via bribery...

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

Q how did these leaders respond to new voters?

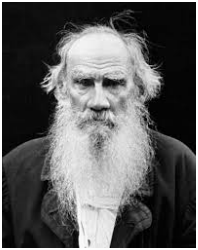
A by changing nature of their speech: simpler, less complex expressions in parliament

Flesch overtime plot

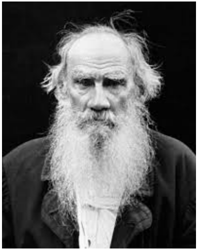


Partner Exercise

Partner Exercise

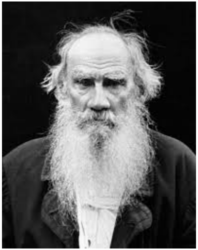


Partner Exercise



Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

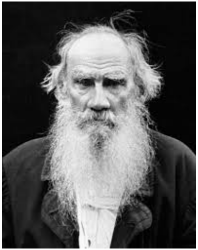
Partner Exercise



Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

1. You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in?

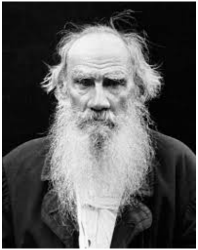
Partner Exercise



Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

1. You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?

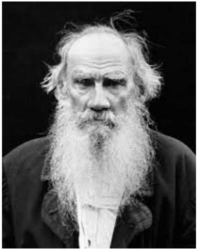
Partner Exercise



Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

1. You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?
2. You estimate the (average) FRE scores for all their novels, respectively. Which estimate (for Tolstoy or Salinger) are you more confident in?

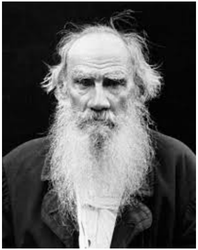
Partner Exercise



Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

1. You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?
2. You estimate the (average) FRE scores for all their novels, respectively. Which estimate (for Tolstoy or Salinger) are you more confident in? Why?

Partner Exercise



Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

1. You estimate the FRE score for *War and Peace* and compare it to the FRE of *The Catcher in the Rye*. Which estimate are you more confident in? Why?

2. You estimate the (average) FRE scores for all their novels, respectively. Which estimate (for Tolstoy or Salinger) are you more confident in? Why?

Sampling and Uncertainty

Sampling and Uncertainty

Sampling and Uncertainty

To now,

Sampling and Uncertainty

To now, we've been concerned with point estimates

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43,

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is unsatisfying:

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is unsatisfying: it says nothing of the variance of such estimates,

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is unsatisfying: it says nothing of the variance of such estimates, yet surely we are more certain of an estimate from a long text (or many texts) than we are from a short text.

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is unsatisfying: it says nothing of the variance of such estimates, yet surely we are more certain of an estimate from a long text (or many texts) than we are from a short text.

e.g. how much would you trust a one word response vs a 1000-word speech from a member of parliament in terms of its complexity or policy content?

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is unsatisfying: it says nothing of the variance of such estimates, yet surely we are more certain of an estimate from a long text (or many texts) than we are from a short text.

e.g. how much would you trust a one word response vs a 1000-word speech from a member of parliament in terms of its complexity or policy content?

→ think a little more systematically about the sampling distribution of a statistic.

Sampling Distributions: Reminder

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean,

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important,

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important, because it tells us the uncertainty around our statistic,

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important, because it tells us the uncertainty around our statistic, and we can use that to produce

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals and make statements about the statistical significance of differences between means of different groups.

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals and make statements about the statistical significance of differences between means of different groups.

Bootstrapping for Sampling Distributions

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population,

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it,

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest.

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method,

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.
Remarkably,

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.
Remarkably, it works well,

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.
Remarkably, it works well, even in quite small samples (e.g. $N < 20$)

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.
Remarkably, it works well, even in quite small samples (e.g. $N < 20$)

NB many forms:

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.
Remarkably, it works well, even in quite small samples (e.g. $N < 20$)

NB many forms: non-parametric is most common,

Bootstrapping for Sampling Distributions

Bootstrapping is a method to obtain the properties of an estimator—the variance, here—via random sampling with replacement from our sample.

It treats the sample as the population, and draws (say, 1000) samples from it, each time calculating some statistic of interest. Putting those statistics together yields a sampling distribution giving us access to the standard error etc.

NB it is a simulation method, in the sense that we are not analytically deriving the formula for the sampling distribution, but approximating it via random sampling.
Remarkably, it works well, even in quite small samples (e.g. $N < 20$)

NB many forms: non-parametric is most common, though parametric is more precise (but requires additional assumptions)

Bootstrap Unit

Bootstrap Unit

When we have a document,

When we have a document, need to think about what it represents a sample of.

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens?

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs?

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap,

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos:

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

tho this requires segmenting the text into sentences (i.e. cannot use DTM),

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

tho this requires segmenting the text into sentences (i.e. cannot use DTM), and performing the relevant operation (e.g. FRE) as many times as there are sentences.

Bootstrap Unit

When we have a document, need to think about what it represents a sample of.

so tokens? paragraphs? sentences?

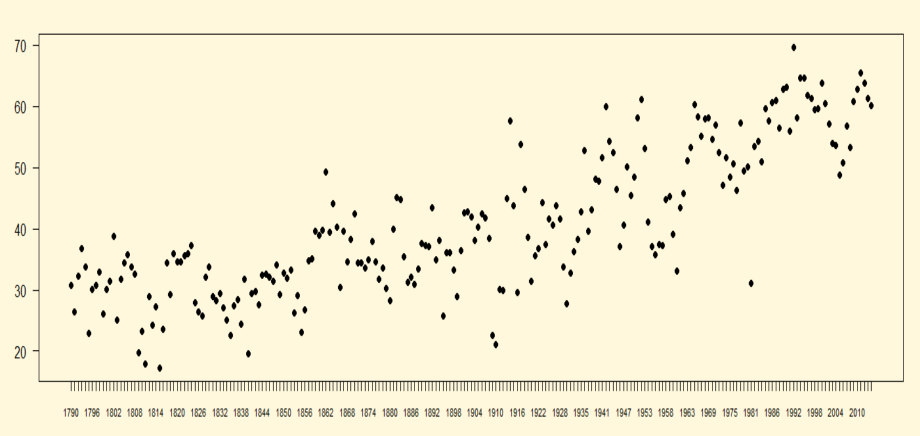
Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

tho this requires segmenting the text into sentences (i.e. cannot use DTM), and performing the relevant operation (e.g. FRE) as many times as there are sentences.

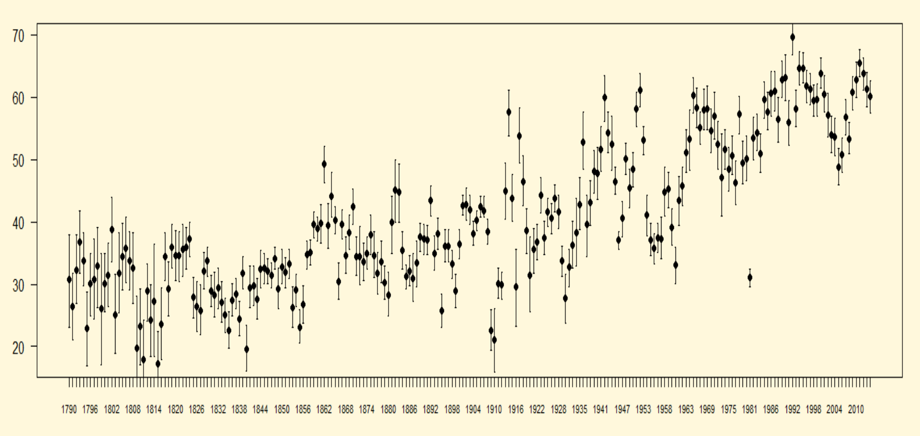
btw long texts give rise to smaller SEs than short ones, which makes sense!

SOU: 1000 bootstrap samples

SOU: 1000 bootstrap samples



SOU: 1000 bootstrap samples



Descriptive Statistics: Stylometrics & Burstiness

Mystery of *The Federalist Papers*

Mystery of *The Federalist Papers*

Mystery of *The Federalist Papers*

85 essays published anonymously in 1787 and 1788

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.

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.

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

Mosteller and Wallace, 1963/4

In essence, they. . .

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

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

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

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

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

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			

More Details

may think that sentence length distinguishes authors

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			

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.

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—

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:

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

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.

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,

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).

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,

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)

Burstiness

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter,

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Surges must be long,

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Surges must be long, and/or intense,

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Surges must be long, and/or intense, depending on specification of model.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

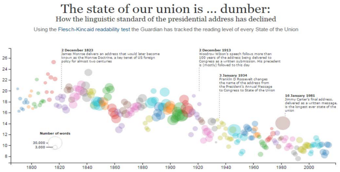
Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

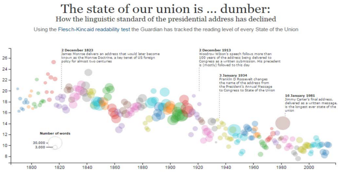
Surges must be long, and/or intense, depending on specification of model.

Applying to SOTU, 1790–2002

Applying to SOTU, 1790–2002



Applying to SOTU, 1790–2002



word	burst
gentlemen	1790–1800
british	1809–1814
slaves	1859–1863
japanese	1942–1945
health	1992–1994
help	1998–