Unsupervised Learning and Its Vagaries

Theory, Feature Selection, Discovery

Arthur Spirling

New York University

July 9, 2018

1. The Basics

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*.

$\rightarrow$ 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.

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,

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*.

$\rightarrow$ 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, 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?'

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

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…

$\rightarrow$ Document Term Matrix

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…

$\rightarrow$ Document Term Matrix

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

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…
```

$\rightarrow$ Document Term Matrix

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

$\rightarrow$ 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.

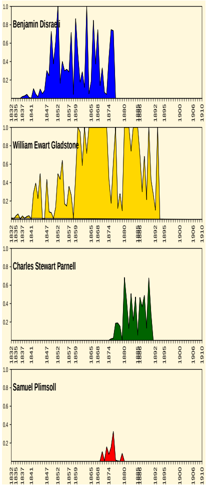
I am grateful to my hon. Friend for her question. I pay tribute to her work with the International Myeloma Foundation…

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

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

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

...



In general, we will. . .

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

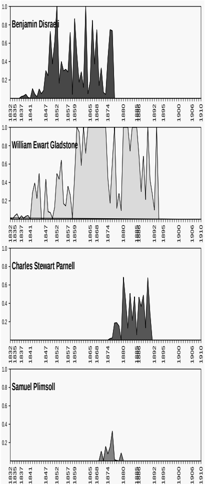
$$
\begin{array}{l c}
 & \begin{array}{cccc} a & an & \dots & ze \end{array} \\
\begin{array}{l} MP_{001} \\ MP_{002} \\ \\ MP_i \\ \\ MP_{654} \\ MP_{655} \end{array} &
\begin{pmatrix}
2 & 0 & \dots & 1 \\
0 & 3 & \dots & 0 \\
\vdots & \vdots & \dots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\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

→ Operate

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

. . .

→ Inference



In general, we will...

Theoretical Model(s)? $\longrightarrow$ Empirical Implications

Get Texts $\rightarrow$ Document Term Matrix $\rightarrow$ Operate $\rightarrow$ Inference

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

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

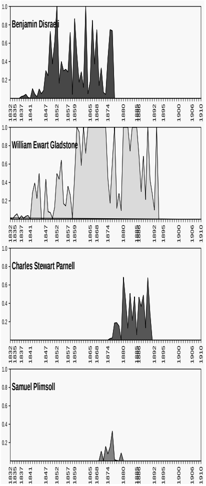
```
I am grateful to my
hon. Friend for her
question. I pay tribute
to her work with the
International Myeloma
Foundation…
```

```
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

...



We have decisions to make. . .

We have decisions to make. . .

- the appropriate population and sample

We have decisions to make...

- the appropriate population and sample

$\rightarrow$ document selection, stochastic view of text

We have decisions to make...

- 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 have decisions to make...

- 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

We have decisions to make. . .

- 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

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

We have decisions to make...

- 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

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

$\rightarrow$ statistical choices

We have decisions to make...

- 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

- 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 have decisions to make. . .

- 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

- 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.

The received wisdom. . .

The received wisdom. . .

- language is extraordinarily complex,

The received wisdom...

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

The received wisdom. . .

- 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.

The received wisdom...

- 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.

The received wisdom...

- 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,

The received wisdom...

- 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,

The received wisdom. . .

- 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.

The received wisdom...

- 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.

The received wisdom. . .

- 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.

The received wisdom...

- 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.

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.

→ 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.

$\rightarrow$ 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.

$\rightarrow$ 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 underlining or **emboldening**.

- punctuation may also be unhelpful

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

→ convert everything to whitespace (?)

Well. . .

Well...

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

Well. . .

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

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

Well. . .

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

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

e.g. social media:

Well. . .

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

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

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

Well. . .

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

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

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

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

Well. . .

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

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

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

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

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

Well. . .

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

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

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

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

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

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

Well. . .

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

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

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

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

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

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

but may not be as important as you think.

‘superfluous’ material: capitalization

‘superfluous’ material: capitalization

Federalist 1

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

‘superfluous’ material: capitalization

Federalist 1

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

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

‘superfluous’ material: capitalization

Federalist 1

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

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

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

'superfluous' material: capitalization

Federalist 1

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

- is the one use of '`The`' the same word as the seven uses of '`the`'?
- is '`UNION`' the same word as '`union`' and '`Union`' as used elsewhere in this essay?

yes $\rightarrow$ lowercase (uppercase) everything

‘superfluous’ material: capitalization

Federalist 1

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

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

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

yes $\rightarrow$ lowercase (uppercase) everything

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

‘superfluous’ material: capitalization

Federalist 1

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

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

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

yes $\rightarrow$ lowercase (uppercase) everything

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

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

‘superfluous’ material: capitalization

Federalist 1

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

- is the one use of ‘`The`’ the same word as the seven uses of ‘`the`’?
- is ‘`UNION`’ the same word as ‘`union`’ and ‘`Union`’ as used elsewhere in this essay?
- yes $\rightarrow$ lowercase (uppercase) everything
- or keep lists (dictionary) of proper nouns, lowercase everything else
- or lowercase words at the beginning of a sentence (how do we know where a sentence begins?) leave everything else as is

Quick Note on Terminology

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.

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.

$\rightarrow$ 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.

→ 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.

→ 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.

Penn POS Tagger

Penn POS Tagger

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

Stemming and Lemmatization

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.

$\rightarrow$ 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.

→ 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`'.

Partner Exercise

Partner Exercise

Consider these elements of a document. Suppose we change all punctuation to whitespace, de-capitalize, remove stop words, and stem what remains.

Partner Exercise

Consider these elements of a document. Suppose we change all punctuation to whitespace, de-capitalize, remove stop words, and stem what remains. What do we get?

Partner Exercise

Consider these elements of a document. Suppose we change all punctuation to whitespace, de-capitalize, remove stop words, and stem what remains. What do we get? Is the original meaning intact?

Partner Exercise

Consider these elements of a document. Suppose we change all punctuation to whitespace, de-capitalize, remove stop words, and stem what remains. What do we get? Is the original meaning intact?

1. `The mountains are beautiful in Ore. and Wash.`
2. `http://www.wsj.com/articles/son-of-saul-not-about-the-survivors-1449590175`
3. `I can't go with him to Beijing.`

We Don’t Care about Word Order

We Don't Care about Word Order

We have now preprocessed our texts.

We Don't Care about Word Order

We have now preprocessed our texts.

Generally,

We Don't Care about Word Order

We have now preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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 preprocessed 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.`"

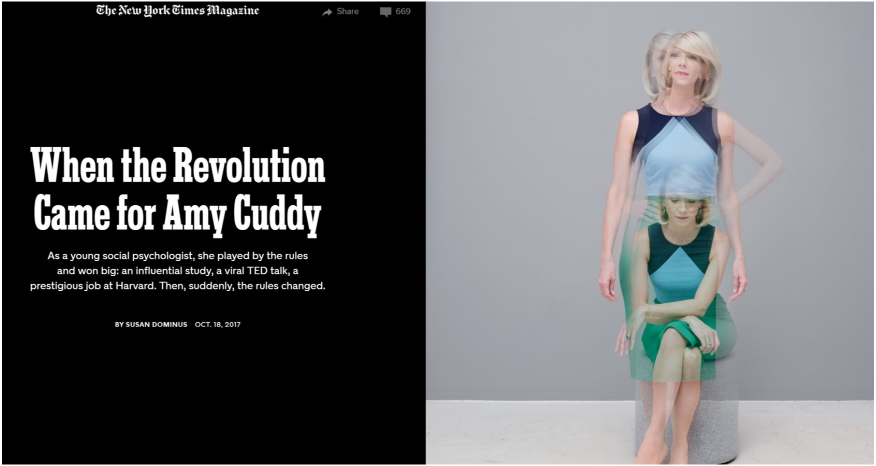
$\rightarrow$ "`lead republican presidenti candid said muslim ban enter us`"

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

2. Record Scratch

Recent Happenings. . .

Recent Happenings. . .



Gelman & Fung in *Slate*

Gelman & Fung in *Slate*

This is not such a surprise. Cuddy's scientific claim was, as is typically the case, based on finding "statistically significant" results in experiments. We know, though, that it is easy for researchers to find statistically significant comparisons even in a single, small, noisy study.

Gelman & Fung in *Slate*

This is not such a surprise. Cuddy's scientific claim was, as is typically the case, based on finding "statistically significant" results in experiments. We know, though, that it is easy for researchers to find statistically significant comparisons even in a single, small, noisy study.Through the mechanism called p-hacking or the garden of forking paths, any specific reported claim typically represents only one of many analyses that could have been performed on a dataset.

→ huh. Seems we're making *a lot* of decisions when we preprocess.

Hang on. . .

Hang on. . .

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature,

Hang on. . .

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Hang on. . .

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

Hang on...

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Hang on. . .

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Q So just check sensitivity by running multiple models?

Hang on...

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Q So just check sensitivity by running multiple models?

A Well, 7 binary steps $\Rightarrow 2^7 = 128$ possible combinations.

Hang on. . .

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Q So just check sensitivity by running multiple models?

A Well, 7 binary steps $\Rightarrow 2^7 = 128$ possible combinations. Good luck.

Hang on...

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Q So just check sensitivity by running multiple models?

A Well, 7 binary steps $\Rightarrow 2^7 = 128$ possible combinations. Good luck.

Q What *can* we do?

Hang on...

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Q So just check sensitivity by running multiple models?

A Well, 7 binary steps $\Rightarrow 2^7 = 128$ possible combinations. Good luck.

Q What *can* we do?

A Check how pairwise distances move between texts as we make choices,

Hang on...

Almost all the advice about preprocessing comes from the supervised ('machine-learning') literature, and is then applied without thought to unsupervised approaches.

Q Does this matter?

A Yes, probably—results are not robust to different preprocessing steps.

Q So just check sensitivity by running multiple models?

A Well, 7 binary steps $\Rightarrow 2^7 = 128$ possible combinations. Good luck.

Q What *can* we do?

A Check how pairwise distances move between texts as we make choices, esp important when 'theory' is weak. See `preText`.

Let’s investigate. . .

Let's investigate. . .

P – **Punctuation Removal**

Let's investigate...

P – **Punctuation Removal**

N – **Number Removal**

Let's investigate. . .

- P – **Punctuation Removal**
- N – **Number Removal**
- L – **Lowercasing**

Let's investigate...

- P – **Punctuation Removal**
- N – **Number Removal**
- L – **Lowercasing**
- S – **Stemming**

Let's investigate...

- P – **Punctuation Removal**
- N – **Number Removal**
- L – **Lowercasing**
- S – **Stemming**
- W – **Stopword Removal**

Let's investigate...

- P – **Punctuation Removal**
- N – **Number Removal**
- L – **Lowercasing**
- S – **Stemming**
- W – **Stopword Removal**
- I – **Infrequent Term Removal**

Let's investigate...

- P – **Punctuation Removal**
- N – **Number Removal**
- L – **Lowercasing**
- S – **Stemming**
- W – **Stopword Removal**
- I – **Infrequent Term Removal**
- '3' – **n-gram Inclusion**

Let's investigate...

P – **Punctuation Removal**

N – **Number Removal**

L – **Lowercasing**

S – **Stemming**

W – **Stopword Removal**

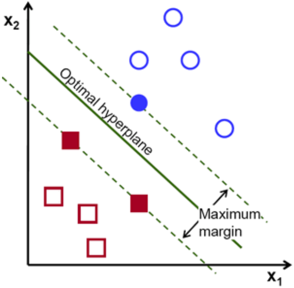
I – **Infrequent Term Removal**

'3' – **n-gram Inclusion**

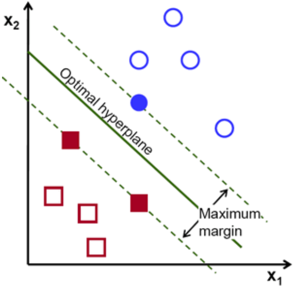
7 binary choices $\longrightarrow 2^7 = 128$ specifications.

Does it Matter? Supervised Learning Edition

Does it Matter? Supervised Learning Edition

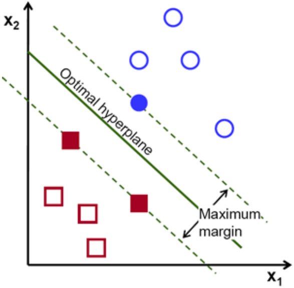


Does it Matter? Supervised Learning Edition



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Does it Matter? Supervised Learning Edition

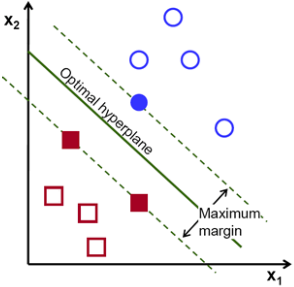


Predicted class

		P	N
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Well-defined: either step improves ability to predict target,

Does it Matter? Supervised Learning Edition

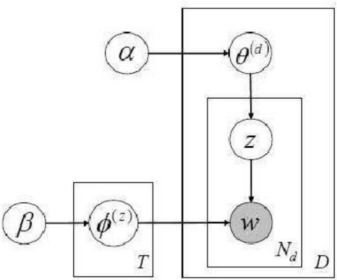


		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

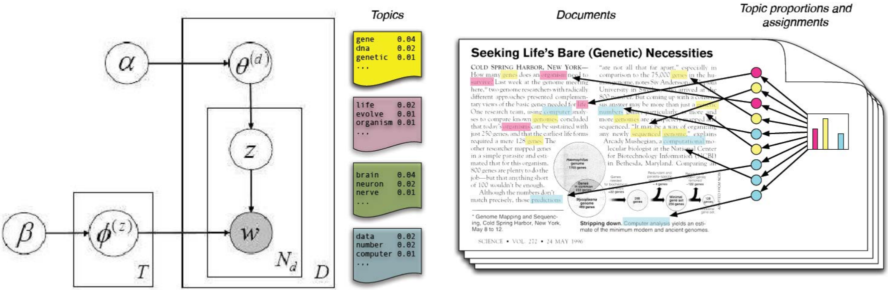
Well-defined: either step improves ability to predict target, or it doesn't.

Does it Matter? Unsupervised Learning Edition

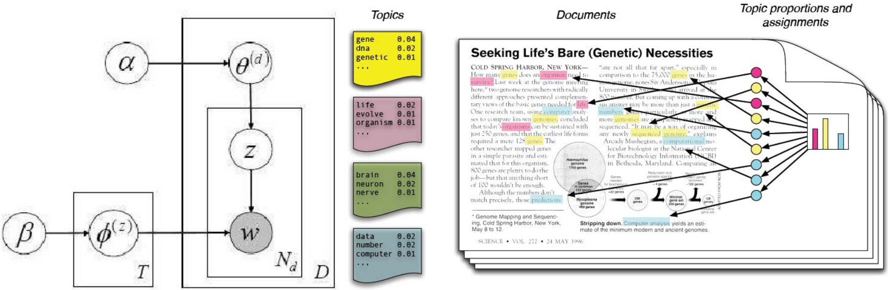
Does it Matter? Unsupervised Learning Edition



Does it Matter? Unsupervised Learning Edition

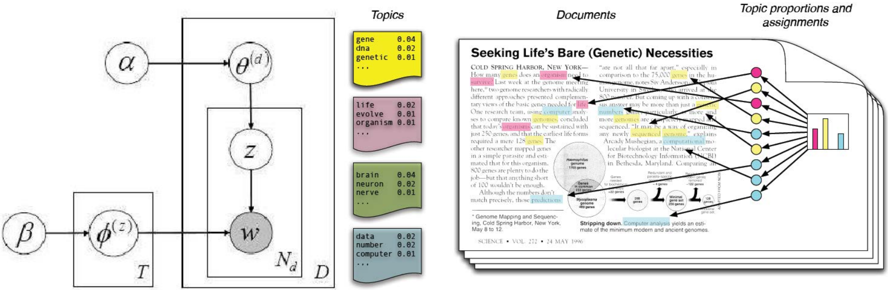


Does it Matter? Unsupervised Learning Edition



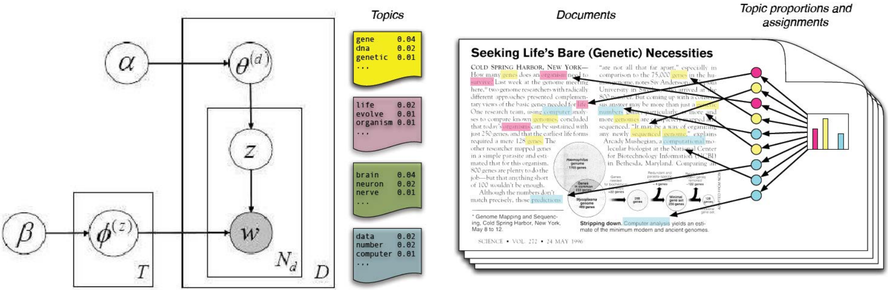
No well-defined/general performance measure:

Does it Matter? Unsupervised Learning Edition



No well-defined/general performance measure: what matters is ‘discovery’ and ‘description’.

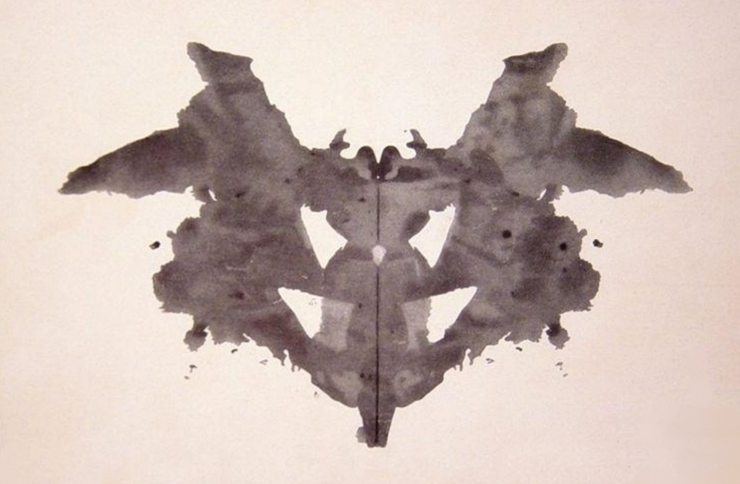
Does it Matter? Unsupervised Learning Edition



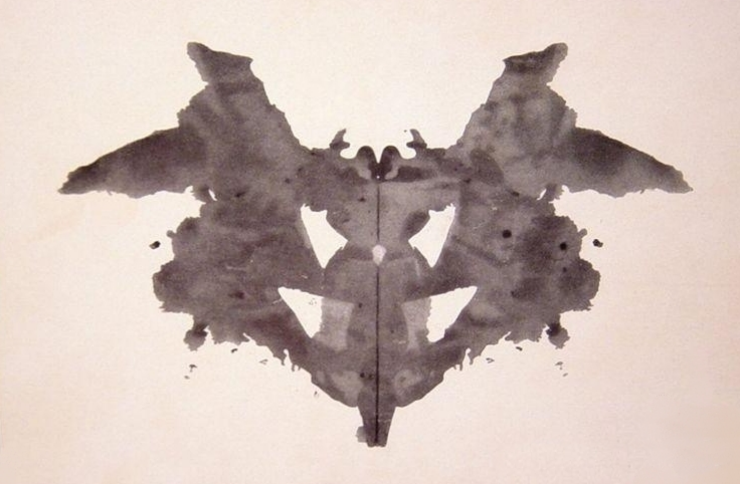
No well-defined/general performance measure: what matters is 'discovery' and 'description'. So, it might.

Aside: The ‘discovery’ problem

Aside: The 'discovery' problem



Aside: The ‘discovery’ problem

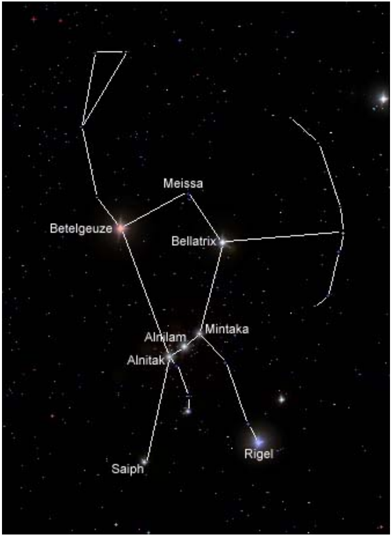


→ what do you see?

The ‘discovery’ problem: what do you see?

The 'discovery' problem: what do you see?

The 'discovery' problem: what do you see?



The ‘discovery’ problem

The ‘discovery’ problem

Humans are very good at abstraction.

The ‘discovery’ problem

Humans are very good at abstraction.

Researchers are very bad at recognizing they are human.

The 'discovery' problem

Humans are very good at abstraction.

Researchers are very bad at recognizing they are human.

$\rightarrow$ very easy to 'make sense' of pretty much anything,

The ‘discovery’ problem

Humans are very good at abstraction.

Researchers are very bad at recognizing they are human.

$\rightarrow$ very easy to ‘make sense’ of pretty much anything, or ‘file-drawer’ it.

The 'discovery' problem

Humans are very good at abstraction.

Researchers are very bad at recognizing they are human.

→ very easy to 'make sense' of pretty much anything, or 'file-drawer' it.

Very unclear how to make 'discovery' unfalsifiable as a criteria of research.

The 'discovery' problem

Humans are very good at abstraction.

Researchers are very bad at recognizing they are human.

$\rightarrow$ very easy to 'make sense' of pretty much anything, or 'file-drawer' it.

Very unclear how to make 'discovery' unfalsifiable as a criteria of research. Could we preregister what would count as a discovery?

Advice from the field...

Advice from the field...

Citation	Steps	Cites
Slapin & Proksch, 2008	P-S-L-N-W	427
Grimmer, 2010	L-P-S-I-W	258
Quinn et al, 2012	P-L-S-I	275
Grimmer & King, 2011	L-P-S-I	109
Roberts et al, 2014	P-L-S-W	117

Related advice from a related field (?)

Related advice from a related field (?)

3. What Could Possibly Go Wrong?

Motivating Example

Motivating Example

Motivating Example

UK Manifesto Corpus
(1918–2001)

Motivating Example

UK Manifesto Corpus (1918–2001): Labour, Liberal, Conservatives.

Motivating Example

UK Manifesto Corpus (1918–2001): Labour, Liberal, Conservatives.

Use Wordfish, unsupervised (Poisson based) scaling algorithm fit by EM.

Motivating Example

UK Manifesto Corpus (1918–2001): Labour, Liberal, Conservatives.

Use Wordfish, unsupervised (Poisson based) scaling algorithm fit by EM.

$\rightarrow$ place documents on one (ideological) dimension.

Motivating Example

UK Manifesto Corpus (1918–2001): Labour, Liberal, Conservatives.

Use Wordfish, unsupervised (Poisson based) scaling algorithm fit by EM.

→ place documents on one (ideological) dimension.

Preprocess DTM 128 ways, and hopefully resulting rank order is robust.

Motivating Example

UK Manifesto Corpus (1918–2001): Labour, Liberal, Conservatives.

Use Wordfish, unsupervised (Poisson based) scaling algorithm fit by EM.

$\rightarrow$ place documents on one (ideological) dimension.

Preprocess DTM 128 ways, and hopefully resulting rank order is robust. Hopefully.

1983 Labour Manifesto

1983 Labour Manifesto

What we do propose to do is to get rid of the nuclear boomerangs which offer no genuine protection to our people but, first and foremost, to help stop the nuclear arms race which is the most dangerous threat to us all.

1983 Labour Manifesto

What we do propose to do is to get rid of the nuclear boomerangs which offer no genuine protection to our people but, first and foremost, to help stop the nuclear arms race which is the most dangerous threat to us all.

Exercise, through the Bank of England, much closer direct control over bank lending. Agreed development plans will be concluded with the banks and other financial institutions. Create a public bank operating through post offices, by merging the National Girobank, National Savings Bank and the Paymaster General's Office.

1983 Labour Manifesto

What we do propose to do is to get rid of the nuclear boomerangs which offer no genuine protection to our people but, first and foremost, to help stop the nuclear arms race which is the most dangerous threat to us all.

Exercise, through the Bank of England, much closer direct control over bank lending. Agreed development plans will be concluded with the banks and other financial institutions. Create a public bank operating through post offices, by merging the National Girobank, National Savings Bank and the Paymaster General's Office.

For all these reasons, British withdrawal from the Community is the right policy for Britain

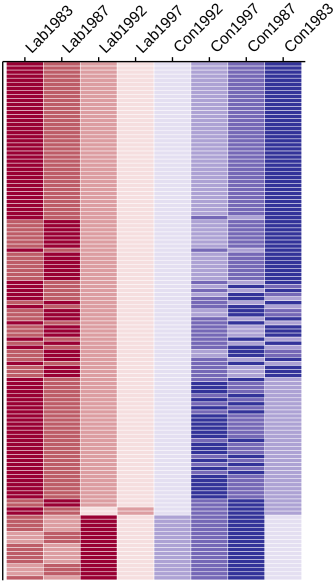
Fixing Ideas: *a priori* rankings

Fixing Ideas: *a priori* rankings

$$\text{Lab}_{1983} < \text{Lab}_{1987} < \text{Lab}_{1992} < \text{Lab}_{1997} <$$
$$\text{Con}_{1992} < \text{Con}_{1997} < \text{Con}_{1987} < \text{Con}_{1983}$$

Wordfish Rankings

Wordfish Rankings



Forking Paths

Forking Paths

12 *unique* document rankings

Forking Paths

12 *unique* document rankings

and substantially different conclusions.

Forking Paths

12 *unique* document rankings

and substantially different conclusions.

Specification	Most Left	Most Right

12 *unique* document rankings

and substantially different conclusions.

Specification	Most Left	Most Right
P-N-S-W-I-3	Lab $_{1983}$	Cons $_{1983}$
N-S-W-3	Lab $_{1987}$	Cons $_{1987}$
N-L-3	Lab $_{1992}$	Cons $_{1987}$
N-L-S	Lab $_{1983}$	Cons $_{1992}$

4. A Solution

A 'Solution': `preText`

A 'Solution': `preText`

1. Assess consequences of preprocessing choices,

A ‘Solution’: `preText`

1. Assess consequences of preprocessing choices, and provide ‘early warning’ of trouble

A 'Solution': preText

1. Assess consequences of preprocessing choices, and provide 'early warning' of trouble

2. Characterize a number of (representative?) corpora

A 'Solution': `preText`

1. Assess consequences of preprocessing choices, and provide 'early warning' of trouble
2. Characterize a number of (representative?) corpora
3. Easy to (ab)use `R` package

A ‘Solution’: preText

1. Assess consequences of preprocessing choices, and provide ‘early warning’ of trouble
2. Characterize a number of (representative?) corpora
3. Easy to (ab)use R package

preText: Diagnostics to Assess the Effects of Text Preprocessing Decisions

Functions to assess the effects of different text preprocessing decisions on the inferences drawn from the resulting document-term matrices they generate.

Version:	0.4.4
Depends:	R ($\geq$ 3.3.0)
Imports:	quanteda, gridExtra, ggplot2, vegan, grid, parallel, topicmodels, cowplot, ecodist, proxy, reshape2
Suggests:	testthat, knitr, rmarkdown
Published:	2016-10-08
Author:	Matthew J. Denny, Arthur Spirling.
Maintainer:	Matthew J. Denny <mdenny at psu.edu>
License:	GPL-3
NeedsCompilation:	no
Materials:	README
CRAN checks:	preText results

Fundamental Idea

Fundamental Idea

Start with (no preprocessing) base case

Fundamental Idea

Start with (no preprocessing) base case

Compare how pairwise document distances change with different preprocessing decisions

Fundamental Idea

Start with (no preprocessing) base case

Compare how pairwise document distances change with different preprocessing decisions

Measure how 'unusual' these changes are:

Fundamental Idea

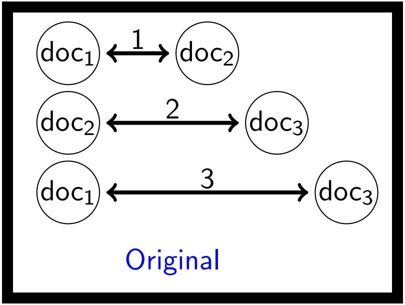
Start with (no preprocessing) base case

Compare how pairwise document distances change with different preprocessing decisions

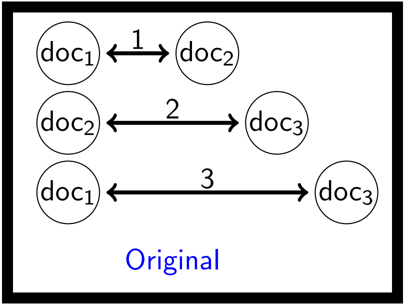
Measure how 'unusual' these changes are: more unusual $\Rightarrow$ be more cautious

Toy Example

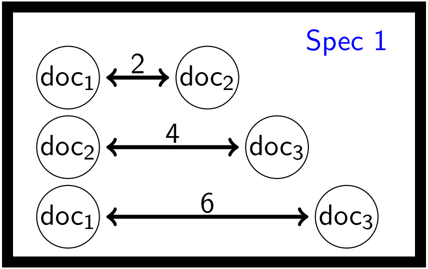
Toy Example



Toy Example

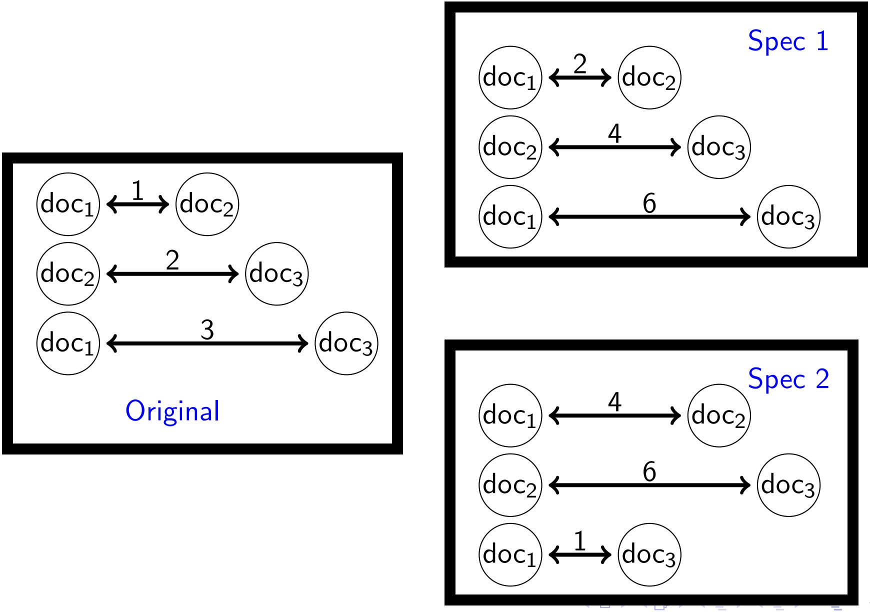


doc1
1
doc2
doc2
2
doc3
doc1
3
doc3
Original



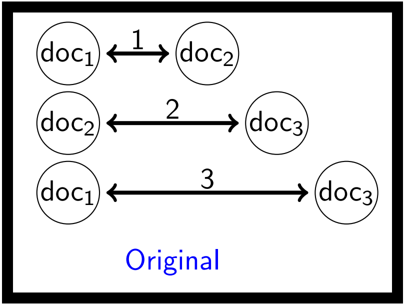
Spec 1
doc1
2
doc2
doc2
4
doc3
doc1
6
doc3

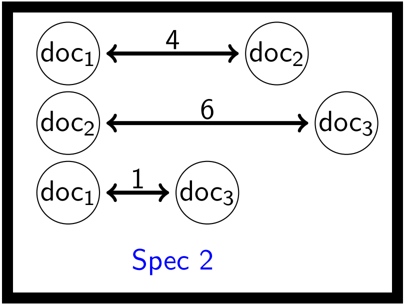
Toy Example


Spec 1
doc1 2 doc2
doc2 4 doc3
doc1 6 doc3
doc1 1 doc2
doc2 2 doc3
doc1 3 doc3
Original
Spec 2
doc1 4 doc2
doc2 6 doc3
doc1 1 doc3

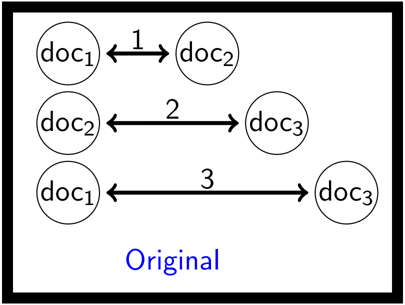
Ranking Distance Changes

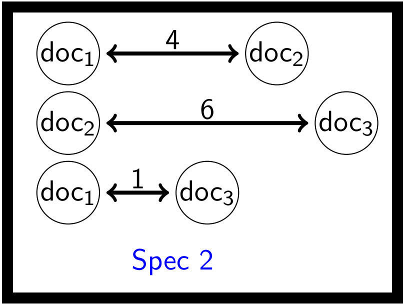
Ranking Distance Changes





Ranking Distance Changes





Original	Specification 2	Abs Rank Difference
$d(1,3) = 3$	$d(2,3) = 6$	$\Delta d(1,3) = 2$
$d(2,3) = 2$	$d(1,2) = 4$	$\Delta d(2,3) = 1$
$d(1,2) = 1$	$d(1,3) = 1$	$\Delta d(1,2) = 1$

Comparing Preprocessing Specifications

Comparing Preprocessing Specifications

Start with first specification, M_1.

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case.

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case. This pair is the largest mover.

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case. This pair is the largest mover.

Ask:

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case. This pair is the largest mover.

Ask: what is the rank of that largest mover pair in terms of the distances changes induced by every other specification (the M_i st $i \neq 1$)?

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case. This pair is the largest mover.

Ask: what is the rank of that largest mover pair in terms of the distances changes induced by every other specification (the M_i st $i \neq 1$)?

e.g.

$$\mathbf{v_{M_1}} = (2_{M_2}, 14_{M_3}, 2_{M_4}, 3_{M_5}, \ldots, 15_{M_{127}})$$

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case. This pair is the largest mover.

Ask: what is the rank of that largest mover pair in terms of the distances changes induced by every other specification (the M_i st $i \neq 1$)?

e.g.

$$\mathbf{v_{M_1}} = (2_{M_2}, 14_{M_3}, 2_{M_4}, 3_{M_5}, \ldots, 15_{M_{127}})$$

Average of absolute difference between $\mathbf{v_{M_1}}$ and '1' yields measure of unusualness.

Comparing Preprocessing Specifications

Start with first specification, M_1. Every specification will have a pair that moves most relative to base case. This pair is the largest mover.

Ask: what is the rank of that largest mover pair in terms of the distances changes induced by every other specification (the M_i st $i \neq 1$)?

e.g.

$$\mathbf{v_{M_1}} = (2_{M_2}, 14_{M_3}, 2_{M_4}, 3_{M_5}, \ldots, 15_{M_{127}})$$

Average of absolute difference between $\mathbf{v_{M_1}}$ and '1' yields measure of unusualness. We can do this for every M_i.

Generalizing: `preText` score (of ith specification)

Generalizing: `preText` score (of ith specification)

Now, consider top k largest moving document pairs

Generalizing: `preText` score (of *i*th specification)

Now, consider top k largest moving document pairs

Average across $\mathbf{v}_{\mathbf{M_i}} \longrightarrow {\mathbf{v}_{\mathbf{M_i}}}^{(k)}$

Generalizing: `preText` score (of *i*th specification)

Now, consider top k largest moving document pairs

Average across $\mathbf{v_{M_i}} \longrightarrow \mathbf{v_{M_i}}^{(k)}$

Normalize by $\frac{n(n-1)}{2}$ (n = number of documents)

Generalizing: `preText` score (of ith specification)

Now, consider top k largest moving document pairs

Average across $\mathbf{v}_{\mathbf{M_i}} \longrightarrow {\mathbf{v}_{\mathbf{M_i}}}^{(k)}$

Normalize by $\frac{n(n-1)}{2}$ (n = number of documents)

$$\boxed{\texttt{preText}\ \text{score}_i = \frac{2{\mathbf{v}_{\mathbf{M_i}}}^{(k)}}{n(n-1)}}$$

Interpreting `preText` scores

Interpreting preText scores

preText scores range between 0 and 1.

Interpreting `preText` scores

`preText` scores range between 0 and 1.

Lower score $\longrightarrow$ **"typical"** changes in document distances.

Interpreting `preText` scores

`preText` scores range between 0 and 1.

Lower score $\longrightarrow$ **"typical"** changes in document distances. That is, pair that was ranked as k top mover in given M_i was also ranked (near) top k mover elsewhere.

Interpreting `preTextx` scores

`preText` scores range between 0 and 1.

Lower score $\longrightarrow$ **"typical"** changes in document distances. That is, pair that was ranked as k top mover in given M_i was also ranked (near) top k mover elsewhere.

Higher score $\longrightarrow$ **"atypical"** changes in document distances.

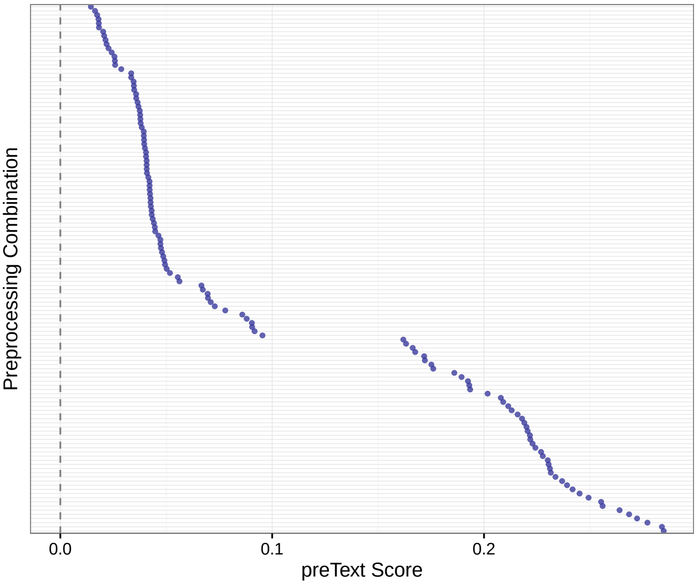
Interpreting `preText` scores

`preText` scores range between 0 and 1.

Lower score $\longrightarrow$ **“typical”** changes in document distances. That is, pair that was ranked as k top mover in given M_i was also ranked (near) top k mover elsewhere.

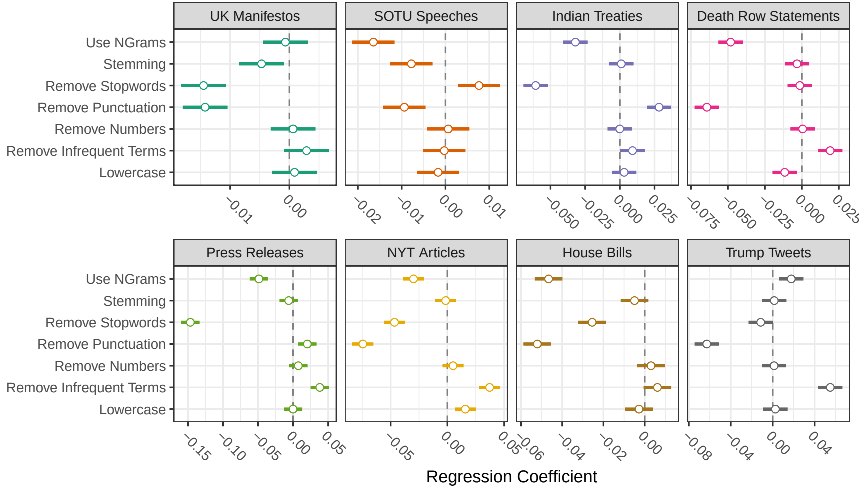
Higher score $\longrightarrow$ **“atypical”** changes in document distances. That is, pair that was ranked as k top mover in given M_i was not ranked (near) top k top mover elsewhere.

preText Scores for Press Releases



Regression Analysis Results

Regression Analysis Results



What To Do About It

What To Do About It

1. Significant parameter estimates serve as an “early warning”.

What To Do About It

1. Significant parameter estimates serve as an "early warning".
2. Conservative approach: average results over all specifications.

What To Do About It

1. Significant parameter estimates serve as an "early warning".

2. Conservative approach: average results over all specifications.

3. Depends on how good your "theory" is.

What To Do About It

1. Significant parameter estimates serve as an "early warning".
2. Conservative approach: average results over all specifications.
3. Depends on how good your "theory" is.
4. *A priori* reasons for selecting a particular specification.

Three Cases

Three Cases

1. All parameter estimates are not significantly different from zero.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic. Reconsider 'theory' of preprocessing,

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic. Reconsider 'theory' of preprocessing, check robustness.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic. Reconsider ‘theory’ of preprocessing, check robustness.

3. Weak theory, some parameter estimates are significantly different from zero.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic. Reconsider 'theory' of preprocessing, check robustness.

3. Weak theory, some parameter estimates are significantly different from zero.

$\rightarrow$ curl up in ball, cry.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic. Reconsider ‘theory’ of preprocessing, check robustness.

3. Weak theory, some parameter estimates are significantly different from zero.

$\rightarrow$ curl up in ball, cry. Reconsider life choices.

Three Cases

1. All parameter estimates are not significantly different from zero.

$\rightarrow$ everything will be fine.

2. Strong theory, some parameter estimates are significantly different from zero.

$\rightarrow$ moderate panic. Reconsider 'theory' of preprocessing, check robustness.

3. Weak theory, some parameter estimates are significantly different from zero.

$\rightarrow$ curl up in ball, cry. Reconsider life choices. Replicate across all combinations: aggregate over results.

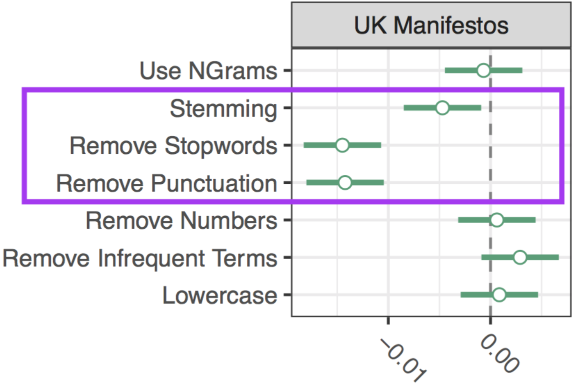
Returning To The UK Wordfish Example

Returning To The UK Wordfish Example

- Weak "theory" $\longrightarrow$ P-N-L-S-W-I

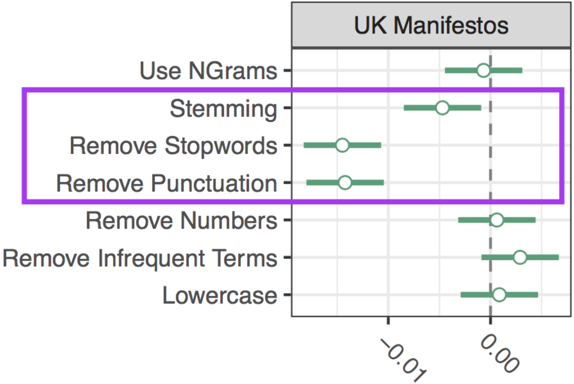
Returning To The UK Wordfish Example

- Weak "theory" $\longrightarrow$ P-N-L-S-W-I



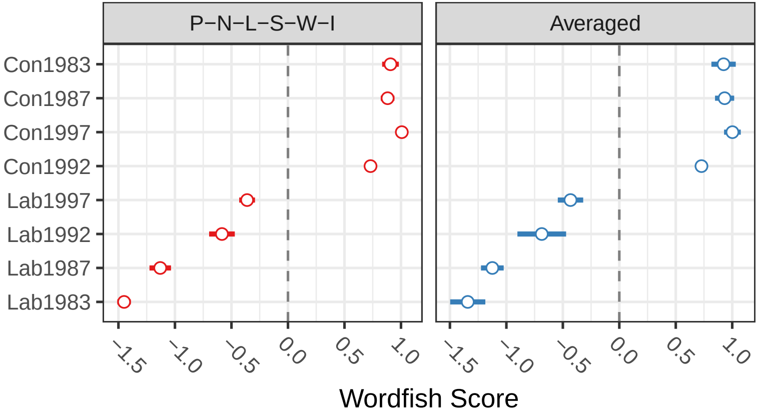
Returning To The UK Wordfish Example

- Weak "theory" $\longrightarrow$ P-N-L-S-W-I

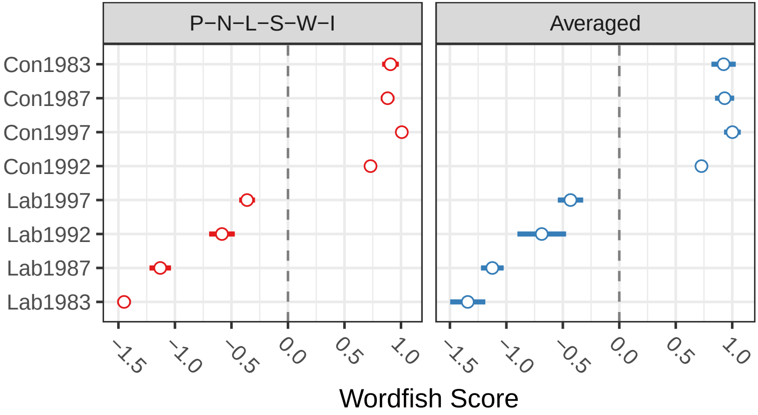


$2^3 = 8$ combinations of choices to average over.

Model Averaging



Model Averaging



Theoretical Specification: "Wrong"

Averaged: Less "Wrong"

So. . . Does it Work?

So. . . Does it Work?

In theory:

So. . . Does it Work?

In theory: probably.

So. . . Does it Work?

In theory: probably.

In practice:

So. . . Does it Work?

In theory: probably.

In practice: definitely.

So. . . Does it Work?

In theory: probably.

In practice: definitely.

$\rightarrow$ every (scaling) example we've looked at,

So. . . Does it Work?

In theory: probably.

In practice: definitely.

$\rightarrow$ every (scaling) example we've looked at, when we say a step doesn't matter, it doesn't.

So. . . Does it Work?

In theory: probably.

In practice: definitely.

→ every (scaling) example we've looked at, when we say a step doesn't matter, it doesn't. When we say a step is consequential,

So...Does it Work?

In theory: probably.

In practice: definitely.

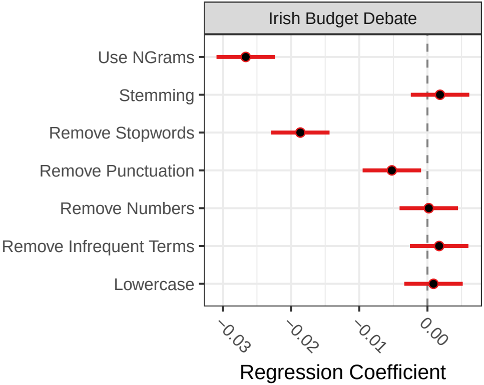
$\rightarrow$ every (scaling) example we've looked at, when we say a step doesn't matter, it doesn't. When we say a step is consequential, it is.

So. . . Does it Work?

In theory: probably.

In practice: definitely.

→ every (scaling) example we've looked at, when we say a step doesn't matter, it doesn't. When we say a step is consequential, it is.



Software and Paper

```
install.packages("preText")
```

Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." Political Analysis 26.2 (2018): 168-189.

github.com/matthewjdenny/preText