3. Descriptive Inference II

DS-GA 3001, Text as Data
Arthur Spirling

February 13, 2018

Housekeeping

Housekeeping

1. Looking to push a homework out ∼Feb 20

Housekeeping

1. Looking to push a homework out ∼Feb 20
2. No speaker this week

Housekeeping

1. Looking to push a homework out $\sim$Feb 20
2. No speaker this week
3. Next week, AS will lecture Tues and Thurs.

From Last week: Identifying characteristic words

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.

Can use similar idea to find most characteristic words by author or speaker in a given corpus.

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says,

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says, compare the frequency with which he uses it (observed),

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says, compare the frequency with which he uses it (observed), with the frequency that it is used in everyone else's speeches (expected).

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says, compare the frequency with which he uses it (observed), with the frequency that it is used in everyone else's speeches (expected).

If $O > E$ and statistically significantly (via X^2 or G^2),

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says, compare the frequency with which he uses it (observed), with the frequency that it is used in everyone else's speeches (expected).

If $O > E$ and statistically significantly (via X^2 or G^2), suggests that this word is characteristic of this author.

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

→ for each word a speaker says, compare the frequency with which he uses it (observed), with the frequency that it is used in everyone else's speeches (expected).

If $O > E$ and statistically significantly (via X^2 or G^2), suggests that this word is characteristic of this author. We can rank the words via their (likelihood) statistics (so furthest departures from the null are highest ranked).

From Last week: Identifying characteristic words

Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says, compare the frequency with which he uses it (observed), with the frequency that it is used in everyone else's speeches (expected).

If $O > E$ and statistically significantly (via X^2 or G^2), suggests that this word is characteristic of this author. We can rank the words via their (likelihood) statistics (so furthest departures from the null are highest ranked).

Walker considers transcripts of South Park (pre-processed),

From Last week: Identifying characteristic words

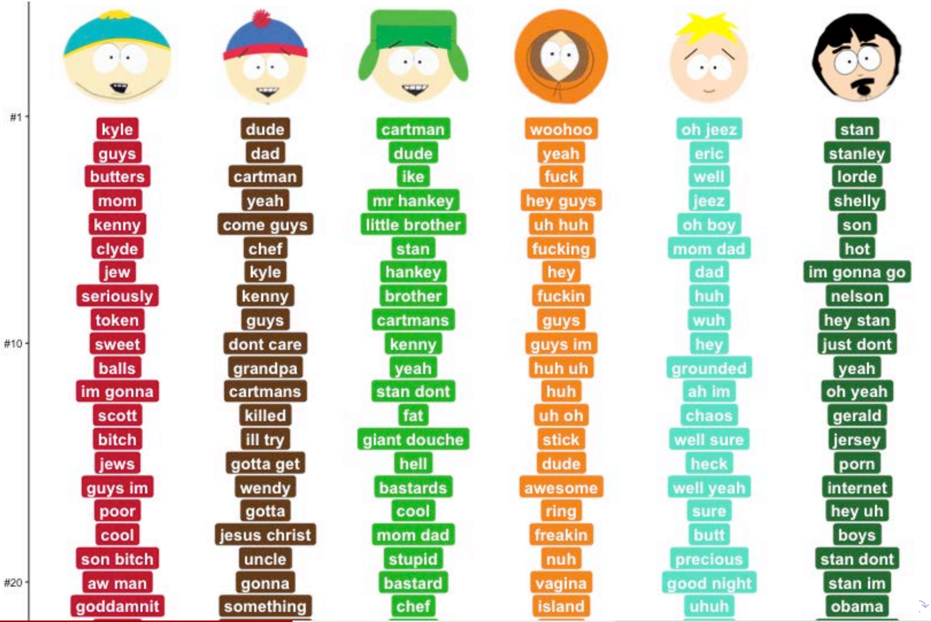
Previously, we used a χ^2 style set up to investigate which bigrams were appearing more than we would expect by chance.
Can use similar idea to find most characteristic words by author or speaker in a given corpus.

$\rightarrow$ for each word a speaker says, compare the frequency with which he uses it (observed), with the frequency that it is used in everyone else's speeches (expected).

If $O > E$ and statistically significantly (via X^2 or G^2), suggests that this word is characteristic of this author. We can rank the words via their (likelihood) statistics (so furthest departures from the null are highest ranked).

Walker considers transcripts of South Park (pre-processed), and collapses on character (treating all other characters' speeches as the corpus when estimating)...

kaylinwalker.com/text-mining-south-park/

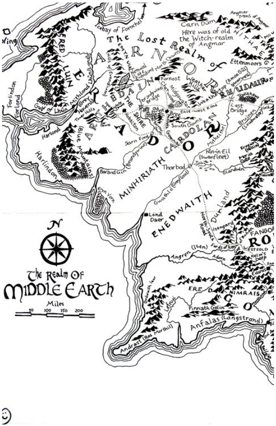


Where Are We?

Where Are We?


The Realm Of
Middle Earth
Miles
50 100 150 200
The Lost Realm of
ARNOR
Here was of old
the Witch-realm
of Angmar
Carn Dûm
Angmar
Ettenmoors
Fornost
Weather Hills
Weathertop
RHUDAUR
Last Bridge
Bree
East-West Road
South Downs
THE SHIRE
Hobbiton
Sarn Ford
CARDOLAN
Tharbad
Nîn-in-Eilph
(Swanfleet)
Mitheithel
Glanduin
MINHIRIATH
Gulf of Lhûn
Grey Havens
Tower Hills
Harlond
Harlindon
Forlindon
Forlond
Blue Mountains
Baranduin
Eryn Vorn
Lond Daer
ENEDWAITH
Dunland
Isengard
FANGORN
Fords of Isen
(Isen)
Adorn
Angren
Helm's Deep
WESTFOLD
ERED NIMRAIS
Pinnath Gelin
Anfalas (Langstrand)
Andrast (Ras Morthil)
Lefnui

Where Are We?



Our fundamental unit of text analysis is the document term matrix.

Where Are We?

Our fundamental unit of text analysis is the document term matrix.

We can compare documents using various distance measures and metrics.

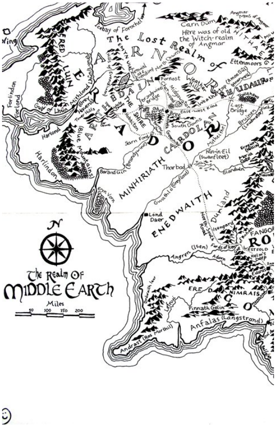
Where Are We?

Our fundamental unit of text analysis is the document term matrix.

We can compare documents using various distance measures and metrics.

now cover some more descriptive measures, dealing with diversity, complexity and style of content.

Where Are We?



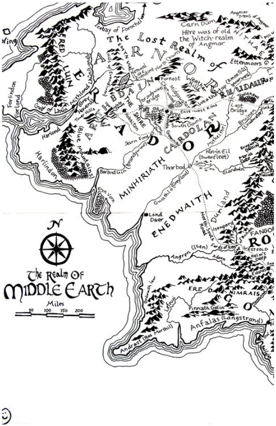
Our fundamental unit of text analysis is the document term matrix.

We can compare documents using various distance measures and metrics.

now cover some more descriptive measures, dealing with diversity, complexity and style of content.

and think seriously about the nature of the sampling process that produces the texts we see,

Where Are We?



Our fundamental unit of text analysis is the document term matrix.

We can compare documents using various distance measures and metrics.

now cover some more descriptive measures, dealing with diversity, complexity and style of content.

and think seriously about the nature of the sampling process that produces the texts we see, and what to do about it.

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.

Tabloid vs Broadsheet

Tabloid vs Broadsheet

RCH NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

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

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

Photo: Getty Images

MORE ON: ISIS

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

Tabloid vs Broadsheet

NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

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

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

Photo: Getty Images

MORE ON: ISIS

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

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

Tabloid vs Broadsheet

RCH NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

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

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

MORE ON: ISIS

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

The New York Times

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

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

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

MIDDLE EAST

Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC. 28, 2015

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

Email

Share

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

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

Tabloid vs Broadsheet

NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

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

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

Photo: Getty Images

MORE ON: ISIS

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

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

The New York Times

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

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

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

MIDDLE EAST

Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC. 28, 2015

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

Email

Share

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

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

Hmm. . .

Hmm. . .

Unexpected, and mostly product of different text lengths:

Hmm. . .

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

Hmm. . .

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

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

Hmm...

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

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

so make denominator non-linear:

Hmm. . .

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

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

so make denominator non-linear:

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

Hmm...

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

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

so make denominator non-linear:

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

so NY Post: $\frac{250}{\sqrt{491}} = 11.28$;

Hmm...

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

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

so make denominator non-linear:

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

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

Hmm...

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

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

so make denominator non-linear:

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

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

$\rightarrow$ has been augmented

Hmm...

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

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

so make denominator non-linear:

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

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

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

Other Ideas

Other Ideas

Malvern and Richards (1997), D.

Other Ideas

Malvern and Richards (1997), *D*. In essence:

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times

Other Ideas

Malvern and Richards (1997), *D*. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times

3. take mean of all sample TTR's and record (bootstrap style)

Other Ideas

Malvern and Richards (1997), *D*. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.
5. estimate relationship between number of tokens N in sample and (mean) TTR where

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.
5. estimate relationship between number of tokens N in sample and (mean) TTR where D is chosen such that the curve estimated fits the empirically observed curve via: $TTR = \frac{D}{N}\left[\left(1 + 2\frac{N}{D}\right)^{\frac{1}{2}} - 1\right]$

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.
5. estimate relationship between number of tokens N in sample and (mean) TTR where D is chosen such that the curve estimated fits the empirically observed curve via: $TTR = \frac{D}{N}\left[\left(1 + 2\frac{N}{D}\right)^{\frac{1}{2}} - 1\right]$

→ typically gives a value of D for a given text somewhere between 10 and 100.

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.
5. estimate relationship between number of tokens N in sample and (mean) TTR where D is chosen such that the curve estimated fits the empirically observed curve via: $TTR = \frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]$

$\rightarrow$ typically gives a value of D for a given text somewhere between 10 and 100. Lowest possible is 0 (i.e. $\frac{1}{N}$), if saying same word repeatedly.

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.
5. estimate relationship between number of tokens N in sample and (mean) TTR where D is chosen such that the curve estimated fits the empirically observed curve via: $TTR = \frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]$

$\rightarrow$ typically gives a value of D for a given text somewhere between 10 and 100. Lowest possible is 0 (i.e. $\frac{1}{N}$), if saying same word repeatedly.

NB subsequent authors suggest 'better' ways to proceed that require less computational effort

Other Ideas

Malvern and Richards (1997), D. In essence:

1. randomly sample 35 tokens (without replacement) and calculate TTR.
2. repeat (1), say 100 times
3. take mean of all sample TTR's and record (bootstrap style)
4. repeat (1)–(3) for 36–50 tokens.
5. estimate relationship between number of tokens N in sample and (mean) TTR where D is chosen such that the curve estimated fits the empirically observed curve via: $TTR = \frac{D}{N}\left[\left(1+2\frac{N}{D}\right)^{\frac{1}{2}}-1\right]$

→ typically gives a value of D for a given text somewhere between 10 and 100. Lowest possible is 0 (i.e. $\frac{1}{N}$), if saying same word repeatedly.

NB subsequent authors suggest 'better' ways to proceed that require less computational effort and less noise.

Transcripts of Interviews (from Duran et al, 2004)

Transcripts of Interviews (from Duran et al, 2004)

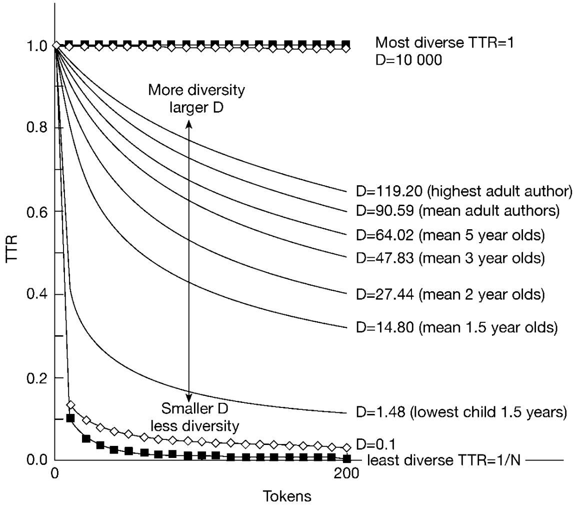


Figure 1: Model TTR plotted against samples of increasing length for different values of D

Other Ideas II: MTLD

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

and in practice, choose that given TTR value to be 0.72.

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

and in practice, choose that given TTR value to be 0.72.

so starting at beginning of text, go word-by-word and record number of words before hitting TTR$= 0.72$ or below.

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

and in practice, choose that given TTR value to be 0.72.

so starting at beginning of text, go word-by-word and record number of words before hitting TTR$=0.72$ or below. Once reached,

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

and in practice, choose that given TTR value to be 0.72.

so starting at beginning of text, go word-by-word and record number of words before hitting TTR$=0.72$ or below. Once reached, consider new segment and record number of words required to obtain $TTR \leq 0.72$ again.

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

and in practice, choose that given TTR value to be 0.72.

so starting at beginning of text, go word-by-word and record number of words before hitting TTR$= 0.72$ or below. Once reached, consider new segment and record number of words required to obtain $TTR \leq 0.72$ again. Repeat until end of text.

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def "the mean length of sequential word strings in a text that maintain a given TTR value"

and in practice, choose that given TTR value to be 0.72.

so starting at beginning of text, go word-by-word and record number of words before hitting TTR$=0.72$ or below. Once reached, consider new segment and record number of words required to obtain $TTR \leq 0.72$ again. Repeat until end of text. Allowances made for various very short segments and remainders.

Other Ideas II: MTLD

Measure of Textual Lexical Diversity, MTLD (McCarthy and Jarvis, 2010).

def “the mean length of sequential word strings in a text that maintain a given TTR value”

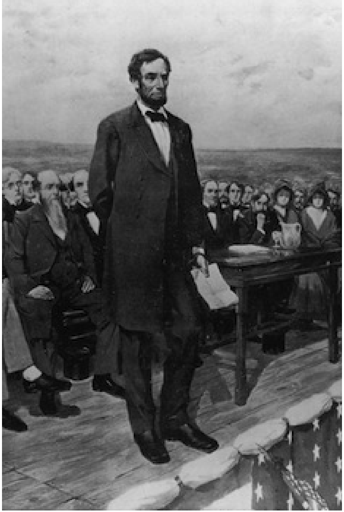
and in practice, choose that given TTR value to be 0.72.

so starting at beginning of text, go word-by-word and record number of words before hitting TTR$= 0.72$ or below. Once reached, consider new segment and record number of words required to obtain $TTR \leq 0.72$ again. Repeat until end of text. Allowances made for various very short segments and remainders.

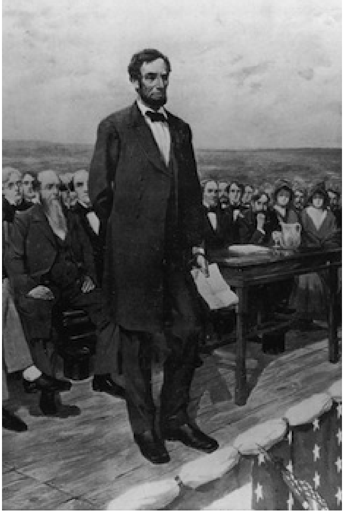
$\rightarrow$ if text is highly diverse, be able to maintain given threshold for longer (on average) and thus mean number of words will be higher.

Lincoln Example

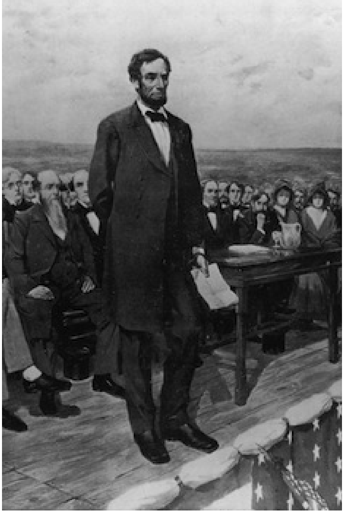
Lincoln Example



Lincoln Example



. . . that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.



. . . that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

$\boxed{\texttt{of the people, by the people, for the people,}}$

$\boxed{\texttt{of the people, by the people, for the people,}}$

of (TTR=1.00 because this type has appeared once)

$\boxed{\texttt{of the people, by the people, for the people,}}$

of (TTR=1.00 because this type has appeared once) the (1.00)

`of the people, by the people, for the people,`

of (TTR=1.00 because this type has appeared once) the (1.00) people (1.00)

$\boxed{\texttt{of the people, by the people, for the people,}}$

of (TTR=1.00 because this type has appeared once) the (1.00) people (1.00) by (1.00)

`of the people, by the people, for the people,`

of (TTR=1.00 because this type has appeared once) the (1.00) people (1.00) by (1.00) the ($\frac{4}{5} = 0.80$)

$\boxed{\texttt{of the people, by the people, for the people,}}$

of (TTR=1.00 because this type has appeared once) the (1.00) people (1.00) by (1.00) the ($\frac{4}{5} = 0.80$) people ($\frac{4}{6} = 0.67$) for (0.714) the (0.625) people (0.556)

`of the people, by the people, for the people,`

of (TTR=1.00 because this type has appeared once) the (1.00) people (1.00) by (1.00) the ($\frac{4}{5} = 0.80$) people ($\frac{4}{6} = 0.67$) for (0.714) the (0.625) people (0.556)

of (1.00) the (1.00) people (1.00) by (1.00) the (0.80) people (0.67)
|| for (0.714) the (.625) people (0.556)

`of the people, by the people, for the people,`

of (TTR=1.00 because this type has appeared once) the (1.00) people (1.00) by (1.00) the ($\frac{4}{5} = 0.80$) people ($\frac{4}{6} = 0.67$) for (0.714) the (0.625) people (0.556)

of (1.00) the (1.00) people (1.00) by (1.00) the (0.80) people (0.67)
|| for (0.714) the (.625) people (0.556)

|| for (1.00) the (1.00) people (1.00)...

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 its 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 its 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 its too small. I respectfully disagree.

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

Measurement of Linguistic Complexity

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.

Notes

Notes

Flesch scoring only uses syllable information:

Notes

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

Notes

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

e.g. “`Indeed, the shoemaker was frightened`” would score similarly to

Notes

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

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

Notes

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

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

Widely used because it 'works',

Notes

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

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

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

Notes

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

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

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

One of many such indices:

Notes

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

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

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

One of many such indices: Gunning-Fog,

Notes

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

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

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

One of many such indices: Gunning-Fog, Dale-Chall,

Notes

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

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

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

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index,

Notes

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

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

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

One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG.

Notes

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

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

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

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

Notes

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

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

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

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

Surprisingly little effort to describe statistical behavior of estimator:

Notes

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

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

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

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

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

Aside: Syllables

Aside: Syllables

how to count syllables?

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English,

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice.

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Count number of vowels...

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Count number of vowels. . .

$+1$ if y makes sound of vowel

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Count number of vowels...

+1 if `y` makes sound of vowel (e.g. `colony`)

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Count number of vowels. . .

+1 if `y` makes sound of vowel (e.g. `colony`)

-1 for silent vowels (e.g. `e` in `invisible`)

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Count number of vowels...

- $+1$ if y makes sound of vowel (e.g. `colony`)
- -1 for silent vowels (e.g. `e` in `invisible`)
- $\pm$ rules for diphthongs and triphthongs:

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

Count number of vowels...

+1 if y makes sound of vowel (e.g. `colony`)

-1 for silent vowels (e.g. `e` in `invisible`)

± rules for diphthongs and triphthongs: i.e. multiple vowels that sound like one (e.g. `book`)

Aside: Syllables

how to count syllables? (other than lookup dictionary words)

well initially done by hand for sample of text.

but in English, counting clusters of vowels works well enough in practice. In essence:

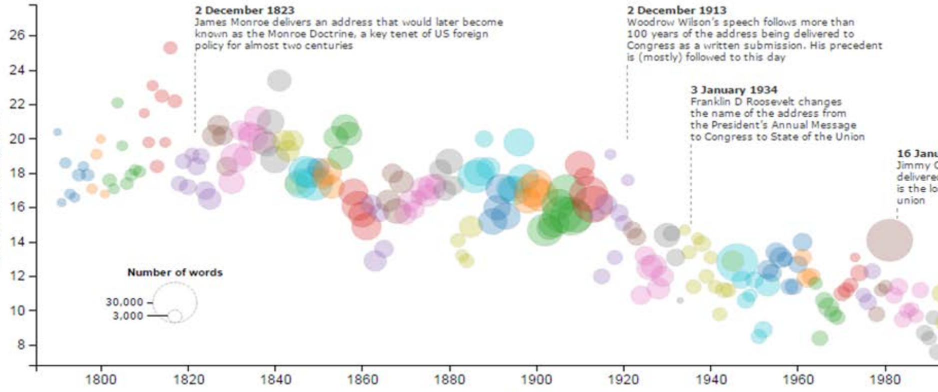
Count number of vowels. . .

- $+1$ if y makes sound of vowel (e.g. `colony`)
- -1 for silent vowels (e.g. `e` in `invisible`)
- $\pm$ rules for diphthongs and triphthongs: i.e. multiple vowels that sound like one (e.g. `book`)
- $\pm$ rules for special endings and use of consonents (e.g. `hassle` vs `mule`)

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

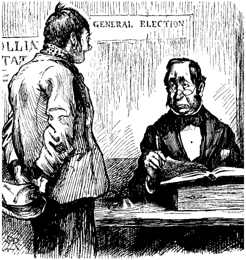
Leaders and their incentives

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



C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

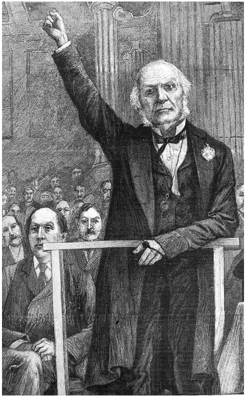
Leaders and their incentives

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

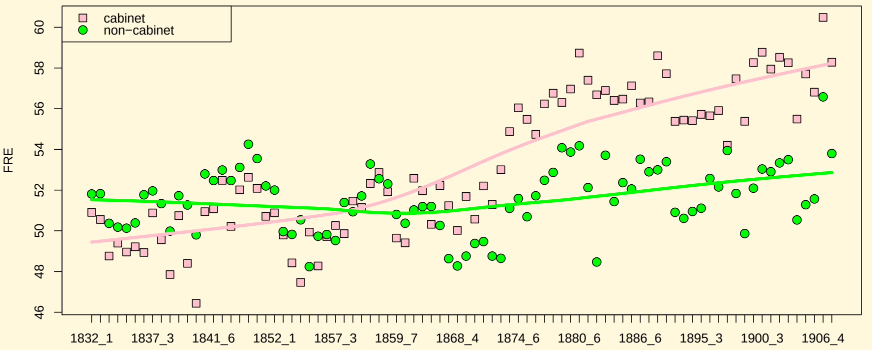
↓ local, clientalistic appeals via bribery...

↑ '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



Dale-Chall, 1948

Dale-Chall, 1948

yields grade level of text sample.

Dale-Chall, 1948

yields grade level of text sample.

DC

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

Dale-Chall, 1948

yields grade level of text sample.

DC

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

where PDW is percentage of difficult words,

Dale-Chall, 1948

yields grade level of text sample.

DC

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

where PDW is percentage of difficult words,

and a 'difficult' word is one that does not appear on Dale & Chall's list of 763

Dale-Chall, 1948

yields grade level of text sample.

DC

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

where PDW is percentage of difficult words,

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

Dale-Chall, 1948

yields grade level of text sample.

DC

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

where PDW is percentage of difficult words,

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

Dale-Chall, 1948

yields grade level of text sample.

DC

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

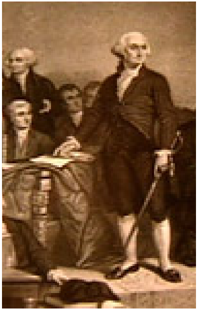
where PDW is percentage of difficult words,

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

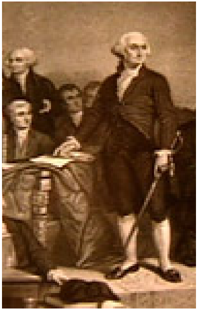
e.g. `about`, `back`, `call`, etc.

Partner Exercise

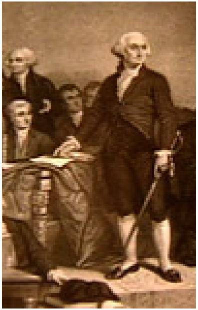
Partner Exercise



Partner Exercise

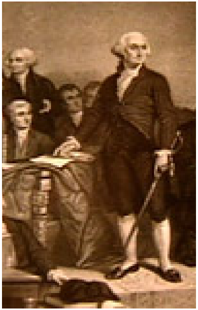


The FRE of SOTU speeches is increasing. Why might it be difficult to make readability comparisons over time?



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

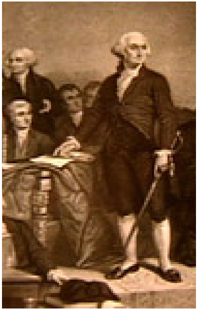
Partner Exercise



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

Does the nature of the decline suggest that speeches are becoming simpler for demand (i.e. voter) or supply (i.e. leader) incentive reasons?

Partner Exercise



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

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

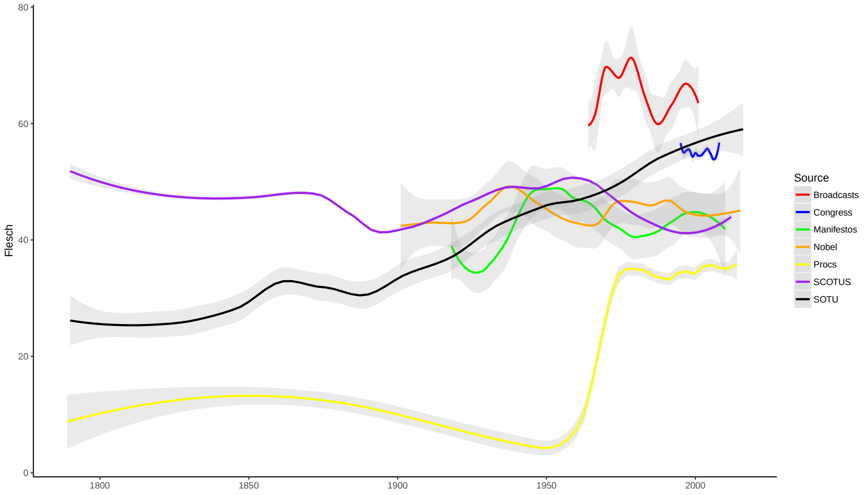
The Great Sentence Length Shift (Benoit, Munger & Spirling)

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not.

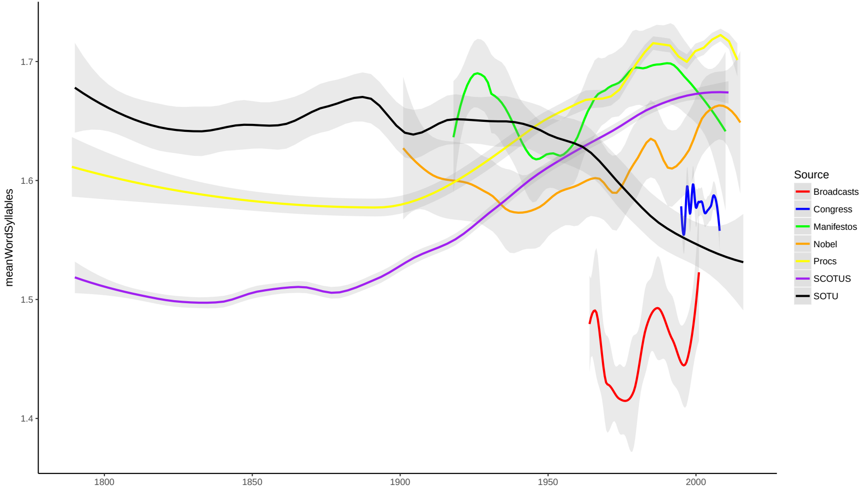


The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns?

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables?

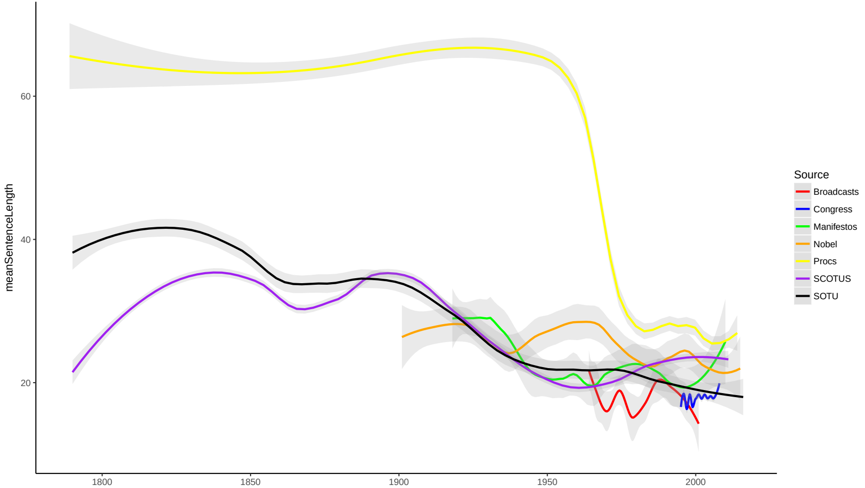


The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables? Sentence length?

The Great Sentence Length Shift (Benoit, Munger & Spirling)

SOTU is simpler: is public discourse becoming 'dumbed down'? Probably not. What's driving these patterns? Syllables? Sentence length?



Can we do better?

Can we do better?

i.e. have I, personally, written a paper about this?

Can we do better?

i.e. have I, personally, written a paper about this?

YES!

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult,

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE,

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Key innovation:

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Key innovation: rarity is from google books corpus,

Can we do better?

i.e. have I, personally, written a paper about this?

YES! BMS crowdsource thousands of snippet comparisons: ask raters which is more difficult, make that a function of covariates.

Everything wrapped into well-known statistical model (Bradley-Terry) for pairwise comparisons: can make probabilistic statements about difficultly.

Model performance not hugely better than FRE, but note we are at least able to make the comparison!

Can ask: what is Pr(Eisenhower easier than Bush)? (is ~ 0.43)

Key innovation: rarity is from google books corpus, and fitted to local decade and domain (adults) that you care about.

Paper and Software

Paper and Software


Download This Paper
Open PDF in Browser
Share:

Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

Kenneth Benoit
London School of Economics & Political Science (LSE); Trinity College Dublin

Kevin Munger
New York University (NYU)

Arthur Spirling
New York University

Date Written: October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a

Paper and Software

Download This Paper | Open PDF in Browser | Share:

Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

Kenneth Benoit
London School of Economics & Political Science (LSE); Trinity College Dublin

Kevin Munger
New York University (NYU)

Arthur Spirling
New York University

Date Written: October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a

CRAN not published | build passing | build passing | coverage 27%

Code for use in measuring the sophistication of political text

"Measuring and Explaining Political Sophistication Through Textual Complexity" by Kenneth Benoit, Kevin Munger, and Arthur Spirling. This package is built on quanteda.

How to install

Using the **devtools** package:

```
devtools::install_github("kbenoit/sophistication")
```

Included Data

new name	original name	description
`data_corpus_fifthgrade`	`fifthCorpus`	Fifth-grade reading texts
`data_corpus_crimson`	`crimsonCorpus`	Editorials from the Harvard *Crimson*
`data_corpus_partybroadcast`	`partybcastCorpus`	UK political party broadcasts
`data_corpus_presdebates`	`presDebateCorpus`	US presidential debates 2016

How to use

Paper and Software

Download This Paper | Open PDF in Browser | Share:

Measuring and Explaining Political Sophistication Through Textual Complexity

42 Pages • Posted: 1 Nov 2017

Kenneth Benoit
London School of Economics & Political Science (LSE); Trinity College Dublin

Kevin Munger
New York University (NYU)

Arthur Spirling
New York University

Date Written: October 30, 2017

Abstract

The sophistication of political communication has been measured using "readability" scores developed from other contexts, but their application out of domain is problematic. We systematically review the shortcomings of previous measures, before developing a new approach, with software, better suited to the task. We use the crowd to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of comprehension. We include previously excluded features such as parts of speech, and a

CRAN not published | build passing | build passing | coverage 27%

Code for use in measuring the sophistication of political text

"Measuring and Explaining Political Sophistication Through Textual Complexity" by Kenneth Benoit, Kevin Munger, and Arthur Spirling. This package is built on quanteda.

How to install

Using the **devtools** package:

```
devtools::install_github("kbenoit/sophistication")
```

Included Data

new name	original name	description
`data_corpus_fifthgrade`	`fifthCorpus`	Fifth-grade reading texts
`data_corpus_crimson`	`crimsonCorpus`	Editorials from the Harvard *Crimson*
`data_corpus_partybroadcast`	`partybcastCorpus`	UK political party broadcasts
`data_corpus_presdebates`	`presDebateCorpus`	US presidential debates 2016

How to use

`github.com/kbenoit/sophistication`

Style and Stylometrics

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.

Mystery of *The Federalist Papers*

85 essays published anonymously in 1787 and 1788

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

That leaves 12 that are disputed.

Mosteller and Wallace, 1963/4

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).

$\rightarrow$ 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)

The Model (Airoldi et al, 2007)

The Model (Airoldi et al, 2007)

let X_w be number of times word w appears in a document.

The Model (Airoldi et al, 2007)

let X_w be number of times word w appears in a document. Expected rate of occurrence is $\Theta_w = \omega\mu$

The Model (Airoldi et al, 2007)

let X_w be number of times word w appears in a document. Expected rate of occurrence is $\Theta_w = \omega\mu$ where μ is expected rate of occurrence in a 'reference length' document (say 1000 words)

The Model (Airoldi et al, 2007)

let X_w be number of times word w appears in a document. Expected rate of occurrence is $\Theta_w = \omega\mu$ where μ is expected rate of occurrence in a 'reference length' document (say 1000 words) and ω is length of given document as multiple of reference document.

The Model (Airoldi et al, 2007)

let X_w be number of times word w appears in a document. Expected rate of occurrence is $\Theta_w = \omega\mu$ where μ is expected rate of occurrence in a 'reference length' document (say 1000 words) and ω is length of given document as multiple of reference document.

so Poisson:

$$P(X_w = x|\Theta_w = (\omega, \mu)) = \frac{e^{\omega\mu}(\omega\mu)^x}{x!}$$

The Model (Airoldi et al, 2007)

let X_w be number of times word w appears in a document. Expected rate of occurrence is $\Theta_w = \omega\mu$ where μ is expected rate of occurrence in a 'reference length' document (say 1000 words) and ω is length of given document as multiple of reference document.

so Poisson:

$$P(X_w = x|\Theta_w = (\omega, \mu)) = \frac{e^{\omega\mu}(\omega\mu)^x}{x!}$$

and Negative Binomial (which adds a gamma distributed random effect, δ):

$$NB(X_w = x|\Theta_w = (\omega, \mu, \delta)) = \frac{\gamma(x+k)}{x!\gamma(k)}(\omega\delta)^x(1+\omega\delta)^{-(x+k)}$$

Estimation and Inference

Estimation and Inference

1 using non-informative priors for the parameters of the distributions,

Estimation and Inference

1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote.

Estimation and Inference

1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors:

Estimation and Inference

1. using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

Estimation and Inference

1. using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

2. begin with equal odds of authorship,

Estimation and Inference

1. using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

2. begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

Estimation and Inference

1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection,

Estimation and Inference

1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection, and some sensitivity checking (allow priors to vary somewhat)

Estimation and Inference

1 using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

2 begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection, and some sensitivity checking (allow priors to vary somewhat)

$\rightarrow$ evidence for Madison is overwhelming for most of the disputed papers: $\sim$ a million to one!

Estimation and Inference

1. using non-informative priors for the parameters of the distributions, obtain the posteriors on the parameters for the authors on the texts we know they wrote. Record the mean of those posteriors: they become the estimates of the parameters in the next step.

2. begin with equal odds of authorship, and take each of the words combined with the parameter estimates to produce posterior odds of authorship (i.e. odds updated by the data we observed).

NB this requires careful (groups of) word selection, and some sensitivity checking (allow priors to vary somewhat)

$\rightarrow$ evidence for Madison is overwhelming for most of the disputed papers: $\sim$ a million to one!

$+$ confirmed by many subsequent analyses (via e.g. machine learning)

Other Work on Attribution

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery,

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

$\rightarrow$ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

$\rightarrow$ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

2007 Airoldi et al consider authorship of Ronald Reagan radio addresses in the 1970s:

Other Work on Attribution



1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

2007 Airoldi et al consider authorship of Ronald Reagan radio addresses in the 1970s: show prescient moves towards end of MAD policy on nuclear weapons,

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

2007 Airoldi et al consider authorship of Ronald Reagan radio addresses in the 1970s: show prescient moves towards end of MAD policy on nuclear weapons, and beginning of Strategic Defence Initiative.

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

2007 Airoldi et al consider authorship of Ronald Reagan radio addresses in the 1970s: show prescient moves towards end of MAD policy on nuclear weapons, and beginning of Strategic Defence Initiative.

→ strong evidence that Reagan composed them himself (not aides),

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

2007 Airoldi et al consider authorship of Ronald Reagan radio addresses in the 1970s: show prescient moves towards end of MAD policy on nuclear weapons, and beginning of Strategic Defence Initiative.

→ strong evidence that Reagan composed them himself (not aides), and thus that he was already planning such things prior to presidency.

Other Work on Attribution

1493 Lorenzo Valla demonstrates that *Donation of Constantine* (giving political authority to the Pope) was a forgery, and not of the period 315AD.

→ contained anachronistic vernacular

1991 Elliot & Valenza find Shakespeare had very consistent style in poems (and not consistent with other candidates)

2007 Airoldi et al consider authorship of Ronald Reagan radio addresses in the 1970s: show prescient moves towards end of MAD policy on nuclear weapons, and beginning of Strategic Defence Initiative.

→ strong evidence that Reagan composed them himself (not aides), and thus that he was already planning such things prior to presidency.

Pushing ‘Stylometry’ Further

Pushing ‘Stylometry’ Further

Pushing 'Stylometry' Further

Pushing 'Stylometry' Further

Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

Pushing 'Stylometry' Further

Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Pushing ‘Stylometry’ Further

Danescu-Niculescu-Mizil et al (“Mark My Words!”) show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al (“No Country for Old Members”) study participants in online communities (e.g. `BeerAdvocate`).

Pushing 'Stylometry' Further

Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al ("No Country for Old Members") study participants in online communities (e.g. `BeerAdvocate`). Two stages: users adopt community terminology,

Pushing 'Stylometry' Further

Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al ("No Country for Old Members") study participants in online communities (e.g. `BeerAdvocate`). Two stages: users adopt community terminology, and then norms pass them by.

Pushing ‘Stylometry’ Further

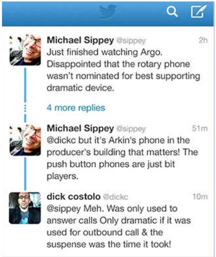
Danescu-Niculescu-Mizil et al (“Mark My Words!”) show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al (“No Country for Old Members”) study participants in online communities (e.g. `BeerAdvocate`). Two stages: users adopt community terminology, and then norms pass them by.

Eisenstein (“Rhetorical Patterns in Legislative Speech”) models discourse relations—

Pushing 'Stylometry' Further



Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al ("No Country for Old Members") study participants in online communities (e.g. `BeerAdvocate`). Two stages: users adopt community terminology, and then norms pass them by.

Eisenstein ("Rhetorical Patterns in Legislative Speech") models discourse relations—conceptual links between units of text,

Pushing 'Stylometry' Further

Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al ("No Country for Old Members") study participants in online communities (e.g. `BeerAdvocate`). Two stages: users adopt community terminology, and then norms pass them by.

Eisenstein ("Rhetorical Patterns in Legislative Speech") models discourse relations—conceptual links between units of text, like 'so', 'however'—

Pushing 'Stylometry' Further

Danescu-Niculescu-Mizil et al ("Mark My Words!") show that twitter users in conversations stylistically accommodate each other (beyond topic and homophily).

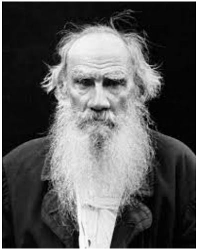
e.g. tone of tentativeness is contagious.

Danescu-Niculescu-Mizil et al ("No Country for Old Members") study participants in online communities (e.g. `BeerAdvocate`). Two stages: users adopt community terminology, and then norms pass them by.

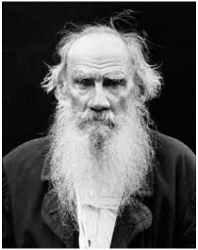
Eisenstein ("Rhetorical Patterns in Legislative Speech") models discourse relations—conceptual links between units of text, like 'so', 'however'—as function of covariates (e.g. ideology of member)

Partner Exercise

Partner Exercise

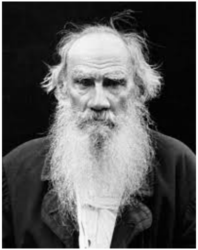


Partner Exercise



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

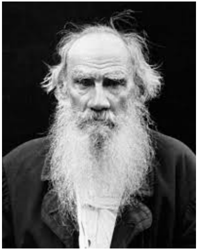
Partner Exercise



Suppose you wanted to compare 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?

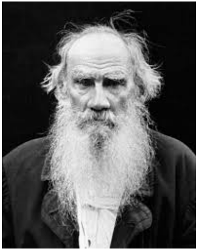
Partner Exercise



Suppose you wanted to compare the novels 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

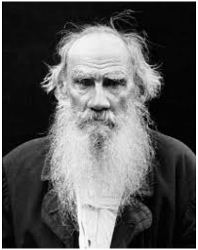


Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.

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?

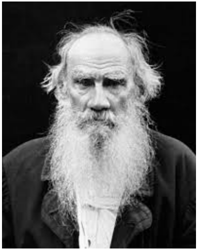
Partner Exercise



Suppose you wanted to compare the novels 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 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

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

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.

Normal Case

Normal Case

For a large enough number of samples of sufficient size,

Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is normally distributed (via the Central Limit Theorem)—

Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is normally distributed (via the Central Limit Theorem)— $\bar{x} \sim \left(\mu, \frac{\sigma^2}{n}\right)$

Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is normally distributed (via the Central Limit Theorem)— $\bar{x} \sim \left(\mu, \frac{\sigma^2}{n}\right)$.

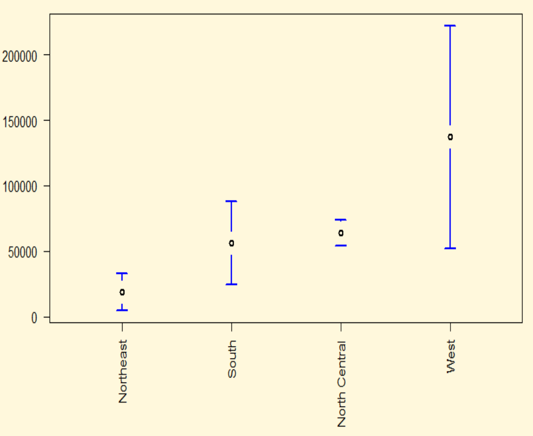
NB We call the standard deviation of the sampling distribution the standard error of the statistic.

Normal Case

For a large enough number of samples of sufficient size, that sampling distribution is normally distributed (via the Central Limit Theorem)—$\bar{x} \sim \left(\mu, \frac{\sigma^2}{n}\right)$.

NB We call the standard deviation of the sampling distribution the standard error of the statistic.

Very helpful!



Sampling Distributions for Text

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

btw in politics, for strategic reasons,

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

btw in politics, for strategic reasons, π might not even be the author's true position/complexity/diversity on an issue

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

btw in politics, for strategic reasons, π might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage?

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

btw in politics, for strategic reasons, π might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage? Is it normal?

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

btw in politics, for strategic reasons, π might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage? Is it normal? Maybe, maybe not.

Sampling Distributions for Text

Need to first think about data generating process by which author intent or characteristic π becomes realized message τ.

This mapping is assumed to be *stochastic* in the sense that it would not be the same 'every time' (even if π were constant).

btw in politics, for strategic reasons, π might not even be the author's true position/complexity/diversity on an issue

But what is the sampling distribution of FRE for a passage? Is it normal? Maybe, maybe not.

$\rightarrow$ difficult to know how we should calculate the sampling distribution and thus the standard error.

Bootstrapping

Bootstrapping

Bootstrapping is a method to obtain the properties of an estimator

Bootstrapping

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

Bootstrapping

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

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

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

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

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

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

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

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

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

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

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

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 Example

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2,

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2,
via bootstrap

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2,
via bootstrap

so resample data ($n = 20$ every time),

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2,
via bootstrap

so resample data ($n = 20$ every time),
and record R^2

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2,
via bootstrap

so resample data ($n = 20$ every time),
and record R^2—then plot...

Bootstrap Example

Have simple linear model, $n = 20$
of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2,
via bootstrap

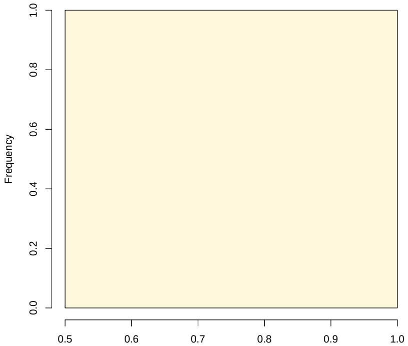
so resample data ($n = 20$ every time),
and record R^2—then plot...

Bootstrap Example

Have simple linear model, $n = 20$ of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2, via bootstrap

so resample data ($n = 20$ every time), and record R^2—then plot...

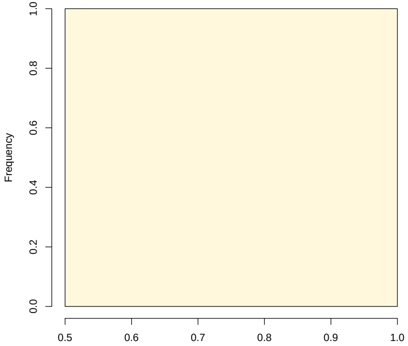


Bootstrap Example

Have simple linear model, $n = 20$ of form $y_i = \beta_0 + \beta_1 X_1 + \beta X_2 + \epsilon_i$

Want to know distribution of R^2, via bootstrap

so resample data ($n = 20$ every time), and record R^2—then plot...



Bootstrap Unit

Bootstrap Unit

When we have a document,

Bootstrap Unit

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?

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?

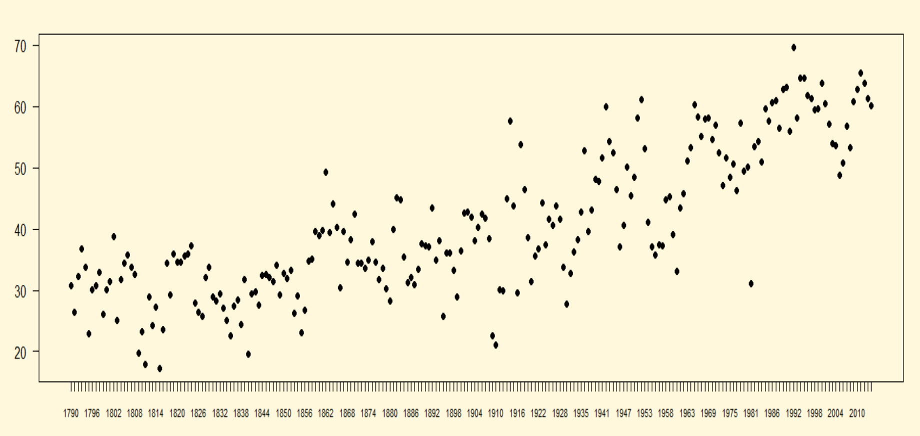
Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

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.

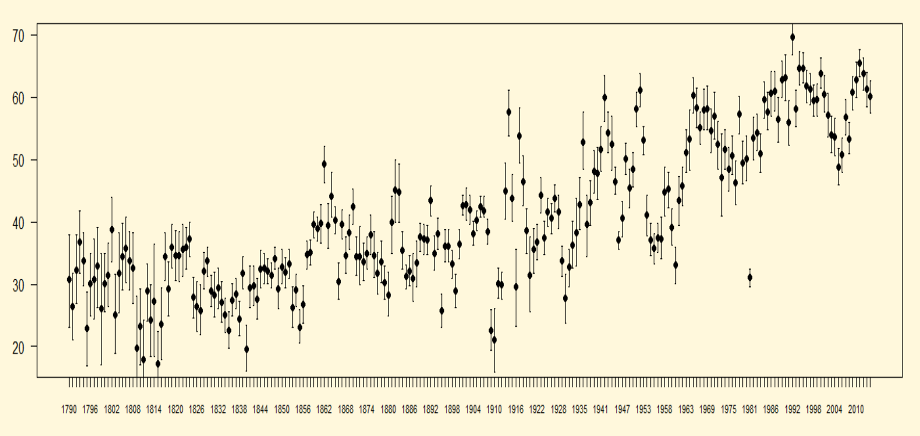
btw long texts give rise to smaller SEs than short ones, which makes sense!

SOU: 1000 bootstrap samples

SOU: 1000 bootstrap samples



SOU: 1000 bootstrap samples



Related Issues

Related Issues

Often, we are dealing with something that is estimated, rather than known.

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text,

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text, or its position in ideological space

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text, or its position in ideological space

Y we use it as a dependent variable which induces heteroskedasticity .

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text, or its position in ideological space

Y we use it as a dependent variable which induces heteroskedasticity .

$\rightarrow$ White or Efron correction of the SEs,

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text, or its position in ideological space

Y we use it as a dependent variable which induces heteroskedasticity .

$\rightarrow$ White or Efron correction of the SEs, or some FGLS variant.

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text, or its position in ideological space

Y we use it as a dependent variable which induces heteroskedasticity .

$\rightarrow$ White or Efron correction of the SEs, or some FGLS variant.

X if we use it as an independent variable, we may have issues of measurement error.

Related Issues

Often, we are dealing with something that is estimated, rather than known.

e.g. the difficulty of a text, or its position in ideological space

Y we use it as a dependent variable which induces heteroskedasticity .

$\rightarrow$ White or Efron correction of the SEs, or some FGLS variant.

X if we use it as an independent variable, we may have issues of measurement error.

$\rightarrow$ SIMEX (simulation-extrapolation) or MO (multiple overimputation) might be called for.

Partner Exercise

Partner Exercise

Suppose you are in a simple linear regression context and you have estimated FRE scores.

Partner Exercise

Suppose you are in a simple linear regression context and you have estimated FRE scores.

1 What is a larger threat to (causal) inference: (random) noise in the dependent variable, or (random) noise in the independent variable? Why?

Partner Exercise

Suppose you are in a simple linear regression context and you have estimated FRE scores.

1. What is a larger threat to (causal) inference: (random) noise in the dependent variable, or (random) noise in the independent variable? Why?
2. What if the goal is prediction of the expected value of Y only?