3. Descriptive Inference II (Flipped)

DS-GA 1015, Text as Data
Arthur Spirling

Feb 23, 2021

Housekeeping

Housekeeping

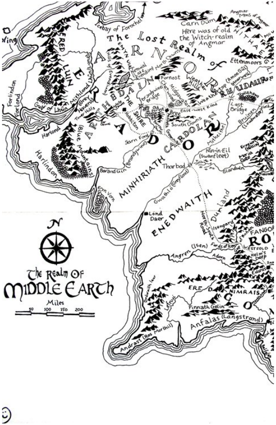
HW 1 out tonight (deadline: two weeks). Turn in an RMarkdown book. Note the academic honesty policy!

Where Are We?

Where Are We?

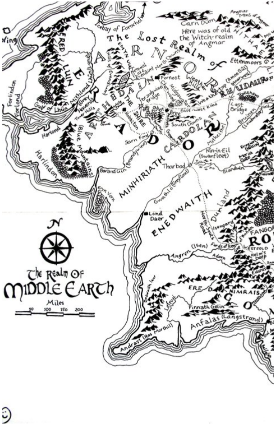
The Lost Realm of
ARNOR
Carn Dûm
Here was of old the Witch-realm of Angmar
Ettenmoors
Fornost
Weather Hills
Weathertop
RHUDAUR
Last Bridge
Bree
East-West Road
South Downs
THE SHIRE
Hobbiton
Sarn Ford
CARDOLAN
Tharbad
Nîn-in-Eilph (Swanfleet)
Glanduin
MINHIRIATH
ENEDWAITH
Lond Daer
Dunland
Isengard
FANGORN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

Where Are We?



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

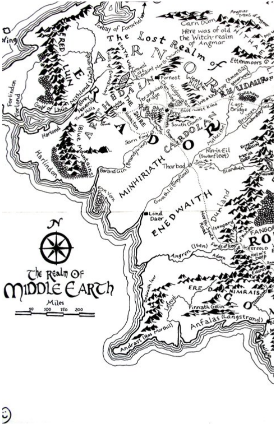
Where Are We?



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

We can compare documents using various distance measures and metrics.

Where Are We?

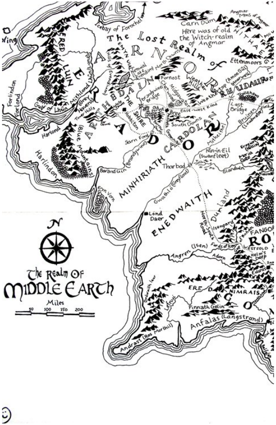


Our fundamental unit of text analysis is the document term matrix.

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?



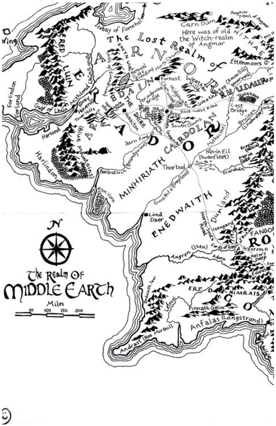
Our fundamental unit of text analysis is the document term matrix.

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.

Distinctive terms (χ^2): Democratic Debates

Distinctive terms (χ^2): Democratic Debates

Laura Bronner
@laurabronner

Some of the more distinctive words and phrases this #DemDebate

docs	turn_the_page	billionaires	diverse	mr_trump	zero	busted_my_neck
BIDEN	0	1	0	0	1	1
BUTTIGIEG	5	0	0	0	0	0
KLOBUCHAR	0	0	0	0	0	0
SANDERS	0	4	0	1	0	0
STEYER	0	0	5	8	0	0
WARREN	0	6	0	0	0	0
YANG	0	0	0	0	7	0

>

11:04 PM · Feb 7, 2020 · Twitter Web App

Recap

Recap

What is the TTR? What does it tell us?

Recap

What is the TTR? What does it tell us?

There is some evidence that it (initially) *falls* as babies learn to speak. Why?

Other Ideas MTLD

Other Ideas MTLD

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

Other Ideas 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 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 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 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 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 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 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 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 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"

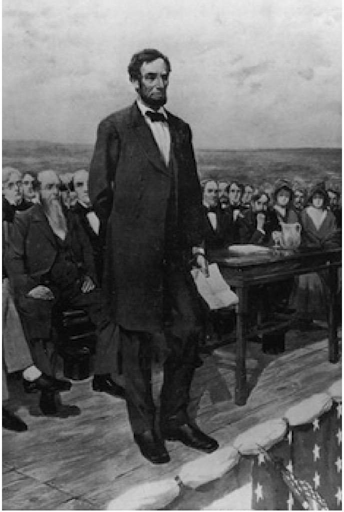
and in practice, choose that given TTR value to be 0.72.

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.

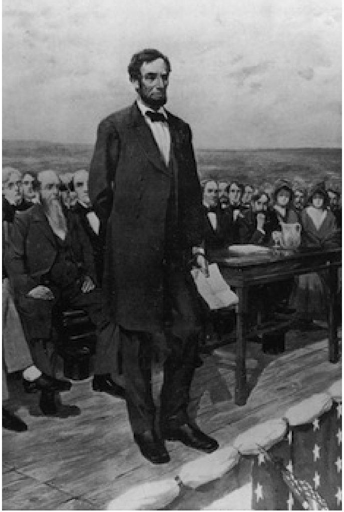
$\rightarrow$ if text is highly diverse, be able to maintain given threshold for longer (on average) and thus mean number of words will be higher.

Lincoln Example

Lincoln Example

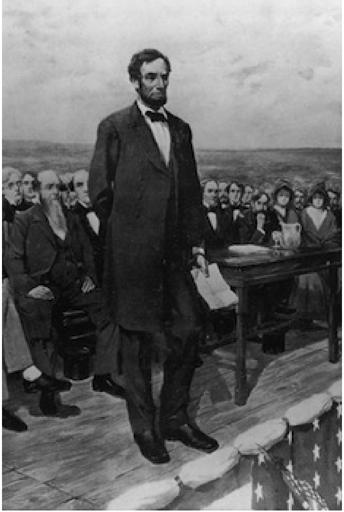


Lincoln Example



. . . 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.

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.

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

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

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)

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

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

`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 ($\frac{4}{5} = 0.80$)

```
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)...

Defining 'Complexity'

Defining ‘Complexity’

Gibson, 1998 “Linguistic complexity: locality of syntactic dependencies” ($\sim$ 3000 cites)

Defining 'Complexity'

Gibson, 1998 "Linguistic complexity: locality of syntactic dependencies" ($\sim$ 3000 cites)

$\rightarrow$ complexity is about "memory cost associated with keeping track of obligatory syntactic requirements":

Defining 'Complexity'

Gibson, 1998 "Linguistic complexity: locality of syntactic dependencies" ($\sim$ 3000 cites)

$\rightarrow$ complexity is about "memory cost associated with keeping track of obligatory syntactic requirements": so, longer reader has to keep idea/entity in mind before confirming its relationship to another, the harder the text.

Defining ‘Complexity’

Gibson, 1998 “Linguistic complexity: locality of syntactic dependencies” ($\sim$ 3000 cites)

$\rightarrow$ complexity is about “memory cost associated with keeping track of obligatory syntactic requirements”: so, longer reader has to keep idea/entity in mind before confirming its relationship to another, the harder the text.

“`The reporter who the senator attacked admitted the error`” is harder than “`The reporter who attacked the senator admitted the error`” because less obvious to whom ‘who’ refers.

Exercise

Exercise

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

Some say my tax plan is too big. Others say it's too small. I respectfully disagree.

Exercise

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

Some say my tax plan is too big. Others say it's too small. I respectfully disagree.

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

Exercise

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

Some say my tax plan is too big. Others say it's too small. I respectfully disagree.

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

What's the Matter with FRE?

What's the Matter with FRE?

- Flesch (1948) suggests *Flesch Reading Ease* statistic

What's the Matter with FRE?

- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

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

What's the Matter with FRE?

- 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.

What's the Matter with FRE?

- 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).

What's the Matter with FRE?

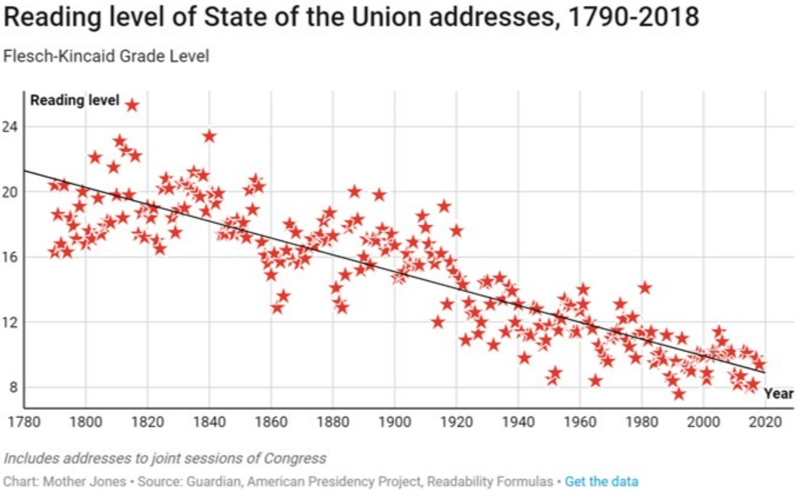
- Flesch (1948) suggests *Flesch Reading Ease* statistic

FRE

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

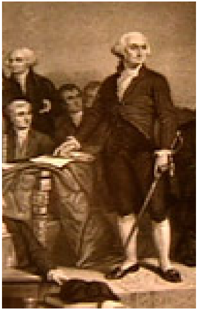
based on $\hat{\beta}$s from linear model where $y =$ average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

What's 'wrong' with this measurement approach?

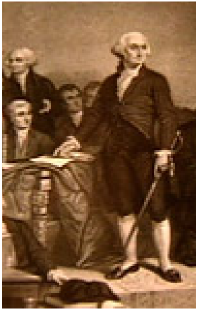

Reading level of State of the Union addresses, 1790-2018
Flesch-Kincaid Grade Level
Reading level
24
20
16
12
8
Year
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
2020
Includes addresses to joint sessions of Congress
Chart: Mother Jones • Source: Guardian, American Presidency Project, Readability Formulas • Get the data

Exercise

Exercise

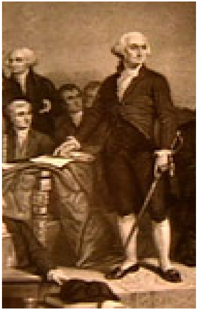


Exercise



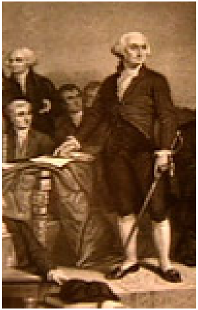
The FRE of SOTU speeches is increasing. Why might it be difficult to make readability comparisons over time?

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?)

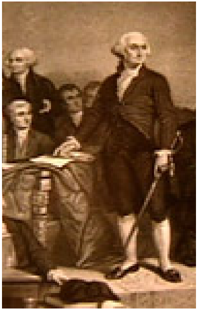
Exercise



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?

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)

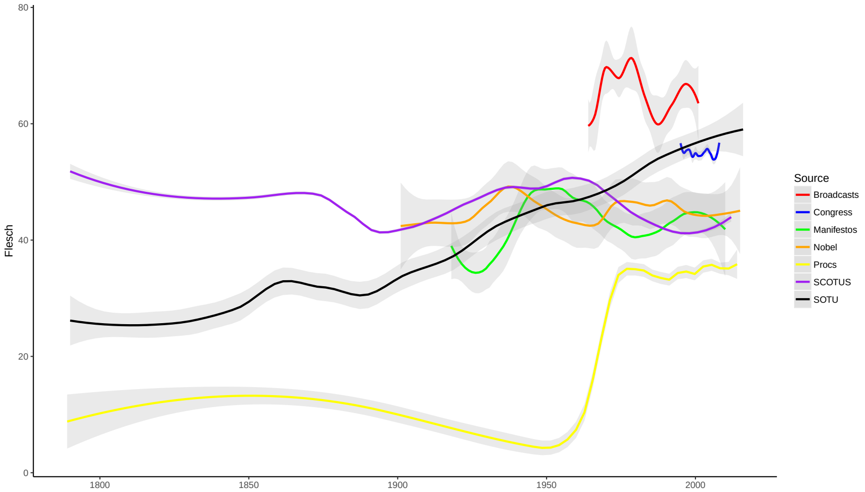
The Great Sentence Length Shift (Benoit, Munger & Spirling)

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.

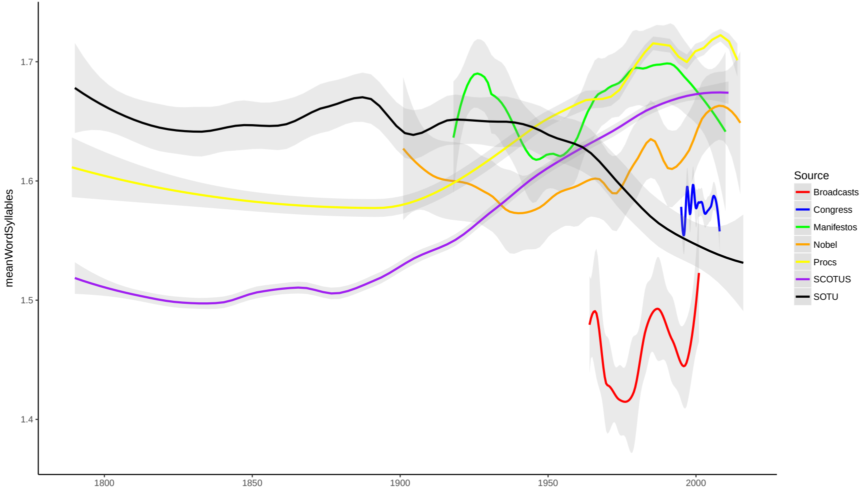


The Great Sentence Length Shift (Benoit, Munger & Spirling)

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?

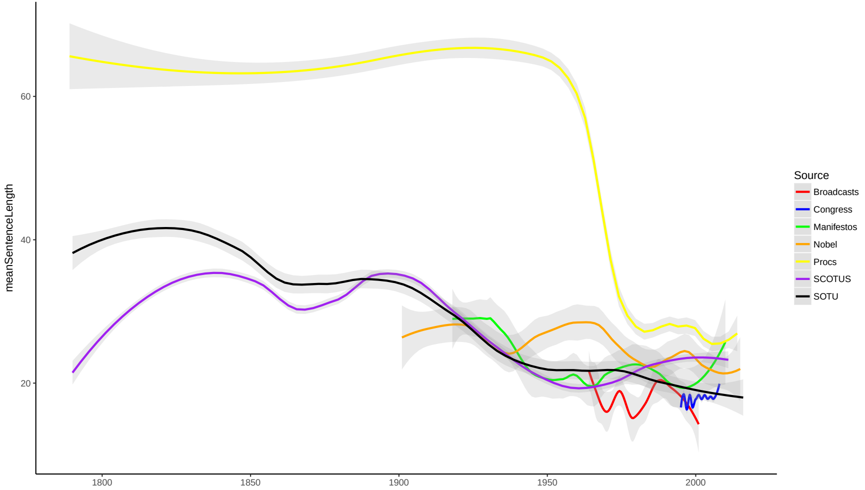


The Great Sentence Length Shift (Benoit, Munger & Spirling)

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?



'Fixing' FRE: Sophistication

‘Fixing’ FRE: Sophistication

1. Ask adults to make pairwise comparisons between documents: crowdsource thousands of such contests.
2. Use elementary statistical model (GLM) for contest outcomes, plus ML to reduce large number of highly correlated covariates on ‘right hand side’.
3. Incorporate rarity in systematic way via Google Books Corpus.

'Fixing' FRE: Sophistication

1. Ask adults to make pairwise comparisons between documents: crowdsource thousands of such contests.

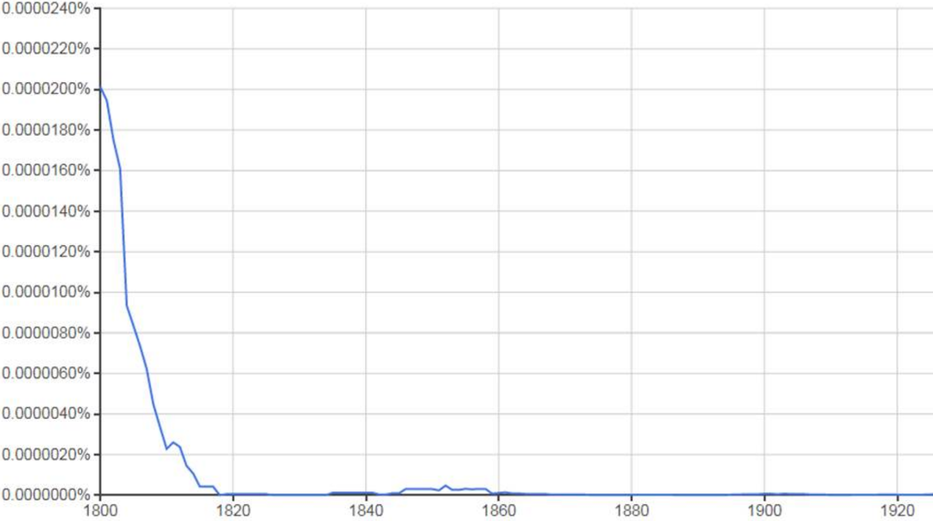
2. Use elementary statistical model (GLM) for contest outcomes, plus ML to reduce large number of highly correlated covariates on 'right hand side'.

3. Incorporate rarity in systematic way via Google Books Corpus.

4. Provide meaningful uncertainty inference estimates via bootstrapping of document-level estimates.

$\rightarrow$ *provide better measure of political sophistication*

Cleaning `ftupid`: What Could Possibly Go Wrong?



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

Q why use function words?

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

Q why use function words? what is the motivation?

A More General Model: The Backbencher's Dilemma...

A More General Model: The Backbencher's Dilemma. . .

Rise of the '*professional*' politician: salaried, ambitious, no non-political experience, dependent on party elites.

A More General Model: The Backbencher's Dilemma. . .

Rise of the 'professional' politician: salaried, ambitious, no non-political experience, dependent on party elites.

But also know partisan voting is on decline: MPs try to develop personal brands to improve Pr(re-election)

A More General Model: The Backbencher's Dilemma. . .

Rise of the 'professional' politician: salaried, ambitious, no non-political experience, dependent on party elites.

But also know partisan voting is on decline: MPs try to develop personal brands to improve Pr(re-election)

Related: unclear how seniority affects this.

‘Interestingness’ as a Measurement Problem

'Interestingness' as a Measurement Problem

All data is labeled:

‘Interestingness’ as a Measurement Problem

All data is labeled: know who said what.

'Interestingness' as a Measurement Problem

All data is labeled: know who said what. Question is whether machine can pinpoint you as speaker of your speech(es)

'Interestingness' as a Measurement Problem

All data is labeled: know who said what. Question is whether machine can pinpoint you as speaker of your speech(es) .

Intuition: you are 'interesting' if we can determine you were the author/speaker of a speech with relative high probability (on average).

‘Interestingness’ as a Measurement Problem

All data is labeled: know who said what. Question is whether machine can pinpoint you as speaker of your speech(es) .

Intuition: you are ‘interesting’ if we can determine you were the author/speaker of a speech with relative high probability (on average). You are ‘boring’ if we can’t.

‘Interestingness’ as a Measurement Problem

All data is labeled: know who said what. Question is whether machine can pinpoint you as speaker of your speech(es) .

Intuition: you are ‘interesting’ if we can determine you were the author/speaker of a speech with relative high probability (on average). You are ‘boring’ if we can’t.

Generalize: two directions, across all speeches, across all speakers, take average pairwise differences.

Formally. . .

Formally...

Consider posterior log-odds of authorship for speech i for speaker t vs s ($\sim$ M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Formally...

Consider posterior log-odds of authorship for speech i for speaker t vs s ($\sim$ M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where x_{iv} is the incidence of token v in speech i.

Formally...

Consider posterior log-odds of authorship for speech i for speaker t vs s ($\sim$ M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where x_{iv} is the incidence of token v in speech i.

Then, think about average log-odds per token (let n_i be tokens in speech i)

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Formally...

Consider posterior log-odds of authorship for speech i for speaker t vs s ($\sim$ M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where x_{iv} is the incidence of token v in speech i.

Then, think about average log-odds per token (let n_i be tokens in speech i)

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Then, average over all speakers and speeches (by t).

Formally...

Consider posterior log-odds of authorship for speech i for speaker t vs s ($\sim$ M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where x_{iv} is the incidence of token v in speech i.

Then, think about average log-odds per token (let n_i be tokens in speech i)

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

Then, average over all speakers and speeches (by t).

Variance has closed form analytical expression.

Formally...

Consider posterior log-odds of authorship for speech i for speaker t vs s ($\sim$ M&W):

$$\sum_{v \in V} x_{iv} \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

where x_{iv} is the incidence of token v in speech i.

Then, think about average log-odds per token (let n_i be tokens in speech i)

$$\sum_{v \in V} (x_{iv}/n_i) \log \left\{ \frac{\Pr(v|t)}{\Pr(v|s)} \right\}$$

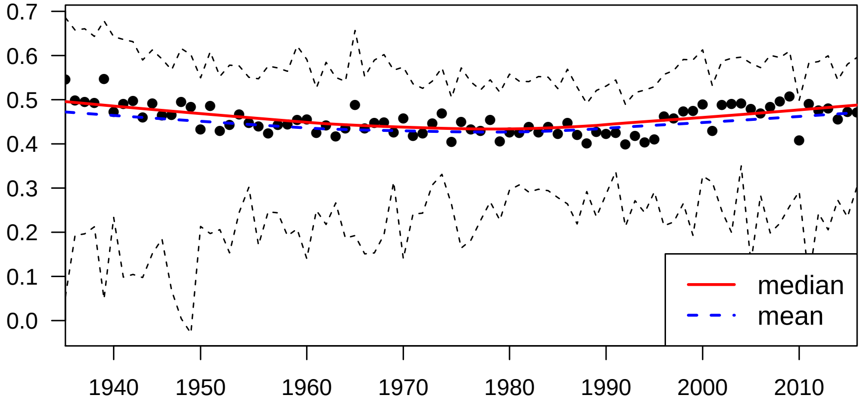
Then, average over all speakers and speeches (by t).

Variance has closed form analytical expression.

Estimation/fitting generally fast.

Average Level of Boringness is Constant!

Average Level of Boringness is Constant!



Software etc



Paper:
`http://nyu.edu/projects/spirling/documents/VeryBoring.pdf`

Software:
`https://github.com/leslie-huang/stylest`

Vignette:
`https://leslie-huang.github.io/stylest/`

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,

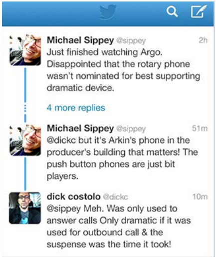
Pushing 'Stylometry' Further

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).

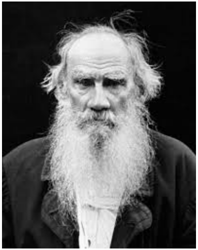
e.g. tone of tentativeness is contagious.

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.

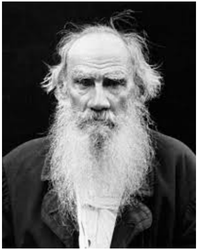
Eisenstein (“Rhetorical Patterns in Legislative Speech”) models discourse relations—conceptual links between units of text, like ‘so’, ‘however’—as function of covariates (e.g. ideology of member)

Exercise

Exercise

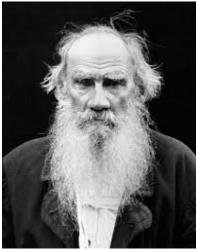


Exercise



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

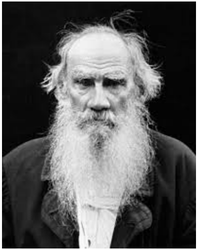
Exercise



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?

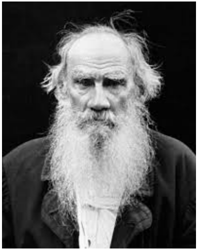
Exercise



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?

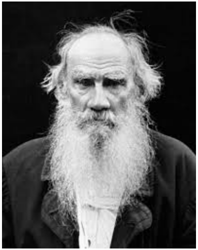
Exercise



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?

Exercise

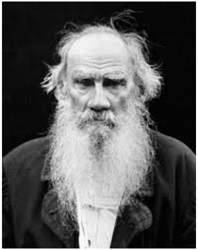


Suppose you wanted to compare the novels of Leo Tolstoy and J.D. Salinger.

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?

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?

Recap

Recap

What is a standard error ?

Recap

What is a standard error ? Why is it hard to obtain for something estimated on text?

Recap

What is a standard error ? Why is it hard to obtain for something estimated on text?

What is bootstrapping? What does it give us?

Recap

What is a standard error ? Why is it hard to obtain for something estimated on text?

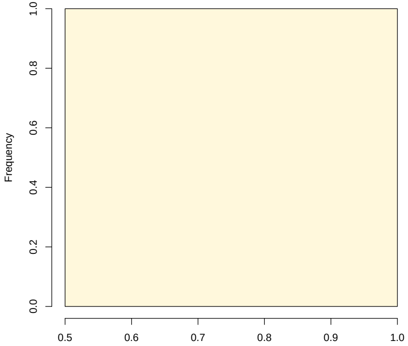
What is bootstrapping? What does it give us?

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...

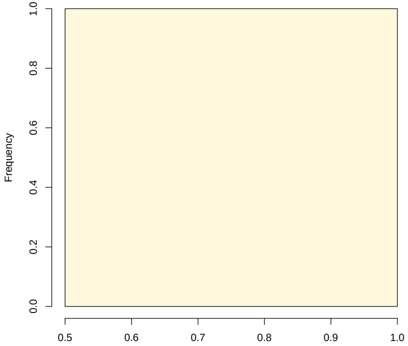


Bootstrap Example

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...



Exercise

Exercise

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

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?

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?