University of Tokyo: Text-as-Data
Day 1, Part II

Arthur Spirling

June 3, 2017

Where Are We?

Where Are We?

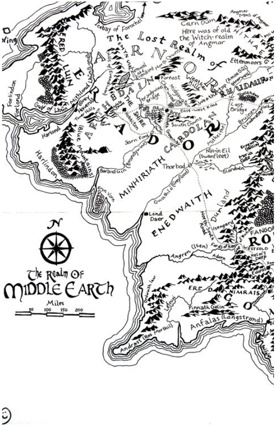
Icebay of Forochel
Carn Dûm
Angmar
Here was of old
the Witch-realm
of Angmar
The Lost Realm of
ARNOR
ERED LUIN
EMYN UIAL
Nenuial
(Lake Evendim)
North Downs
Ettenmoors
Fornost
Weather Hills
Weathertop (Amon Sûl)
RHUDAUR
Hoarwell
Trollshaws
Last Bridge
Bruinen
(Loudwater)
Forlindon
Forlond
Mithlond
(Grey Havens)
Gulf of Lhûn
Lhûn (Lune)
Tower Hills
Harlond
Harlindon
(Blue Mountains)
ERIADOR
ARTHEDAIN
THE SHIRE
Hobbiton
Brandywine Bridge
Old Forest
Bree
East-West Road
Midgewater
South Downs
CARDOLAN
Sarn Ford
(Brandywine)
Baranduin
Mitheithel
Nîn-in-Eilph
(Swanfleet)
Tharbad
Glanduin
Moria Gate
MINHIRIATH
Eryn Vorn
Gwathló (Greyflood)
Lond Daer
ENEDWAITH
Dunland
North-South Road
Isengard
FANGORN
RO
Angren (Isen)
Fords of Isen
Adorn
WESTFOLD
Helm's Deep
Edoras
ERED NIMRAIS
Pinnath Gelin
Anfalas (Langstrand)
Andrast (Ras Morthil)
Lefnui
GO
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.

This is a set of stacked vectors, with each entry in each vector representing the 'amount' of a particular term.

Where Are We?

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

This is a set of stacked vectors, with each entry in each vector representing the 'amount' of a particular term.

This could be (re-)weighted in some way (e.g. tfidf).

Where Are We?

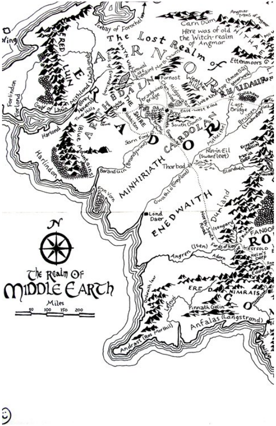
Our fundamental unit of text analysis is the document term matrix.

This is a set of stacked vectors, with each entry in each vector representing the 'amount' of a particular term.

This could be (re-)weighted in some way (e.g. tfidf).

now cover some fundamental statistical properties of text

Where Are We?



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

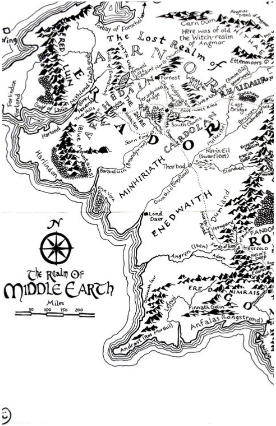
This is a set of stacked vectors, with each entry in each vector representing the 'amount' of a particular term.

This could be (re-)weighted in some way (e.g. tfidf).

now cover some fundamental statistical properties of text

and think about how to compare documents,

Where Are We?



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

This is a set of stacked vectors, with each entry in each vector representing the 'amount' of a particular term.

This could be (re-)weighted in some way (e.g. tfidf).

now cover some fundamental statistical properties of text

and think about how to compare documents, and summarize their content.

Reminder: Quick Note on Terminology

Reminder: Quick Note on Terminology

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

Reminder: Quick Note on Terminology

a type is a unique sequence of characters that are grouped together in some meaningful way. Mostly a word (for us),

Reminder: Quick Note on Terminology

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

Reminder: Quick Note on Terminology

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

Reminder: Quick Note on Terminology

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

e.g. ‘`France`’, ‘`American Revolution`’, ‘`1981`’

Reminder: Quick Note on Terminology

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

e.g. ‘`France`’, ‘`American Revolution`’, ‘`1981`’

a token is a particular *instance* of type.

e.g. “`Dog eat dog world`”,

Reminder: Quick Note on Terminology

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

e.g. ‘`France`’, ‘`American Revolution`’, ‘`1981`’

a token is a particular *instance* of type.

e.g. “`Dog eat dog world`”, contains three types,

Reminder: Quick Note on Terminology

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

e.g. ‘`France`’, ‘`American Revolution`’, ‘`1981`’

a token is a particular *instance* of type.

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

Reminder: Quick Note on Terminology

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

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

a token is a particular *instance* of type.

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

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

Reminder: Quick Note on Terminology

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

e.g. ‘`France`’, ‘`American Revolution`’, ‘`1981`’

a token is a particular *instance* of type.

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

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

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

Lossy Compression

Lossy Compression

- when we use the vector space model we remove some information and throw it away

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

→ this means we cannot restore the original representation of the data:

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

→ this means we cannot restore the original representation of the data: we have a lossy compression.

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

→ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it.

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

→ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

e.g. Reuters Corpus Volume 1 (RCV1) (2004) is a benchmark text collection of $\sim$ 800000 manually coded news stories.

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

e.g. Reuters Corpus Volume 1 (RCV1) (2004) is a benchmark text collection of $\sim$ 800000 manually coded news stories.

RCV1 has $484,494$ types and $197,879,290$ tokens (MR&S book, Table 5.1).

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

e.g. Reuters Corpus Volume 1 (RCV1) (2004) is a benchmark text collection of $\sim$ 800000 manually coded news stories.

RCV1 has $484,494$ types and $197,879,290$ tokens (MR&S book, Table 5.1).

rm numbers	473,723	179,158,204

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

e.g. Reuters Corpus Volume 1 (RCV1) (2004) is a benchmark text collection of $\sim$ 800000 manually coded news stories.

RCV1 has $484,494$ types and $197,879,290$ tokens (MR&S book, Table 5.1).

rm numbers	473,723	179,158,204
lowercase	391,523	179,158,204

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

e.g. Reuters Corpus Volume 1 (RCV1) (2004) is a benchmark text collection of $\sim$ 800000 manually coded news stories.

RCV1 has $484,494$ types and $197,879,290$ tokens (MR&S book, Table 5.1).

rm numbers	473,723	179,158,204
lowercase	391,523	179,158,204
rm 150 stopwords	391,373	94,516,599

Lossy Compression

- when we use the vector space model we remove some information and throw it away (e.g. word order, numbers, capitals, stop words).

$\rightarrow$ this means we cannot restore the original representation of the data: we have a lossy compression.

but presumably, life becomes a lot simpler and the tradeoff is worth it. How much simpler?

e.g. Reuters Corpus Volume 1 (RCV1) (2004) is a benchmark text collection of $\sim$ 800000 manually coded news stories.

RCV1 has $484,494$ types and $197,879,290$ tokens (MR&S book, Table 5.1).

rm numbers	473,723	179,158,204
lowercase	391,523	179,158,204
rm 150 stopwords	391,373	94,516,599
stemming	322,383	94,516,599

Heap's Law: Type-Token relationship

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law:

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: $\boxed{M = kT^b}$

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: $\boxed{M = kT^b}$

where $k \in (30, 100)$ and $b \in (0.4, 0.6)$ for English.

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: $\boxed{M = kT^b}$

where $k \in (30, 100)$ and $b \in (0.4, 0.6)$ for English.

if we pre-process in different ways,

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: $\boxed{M = kT^b}$

where $k \in (30, 100)$ and $b \in (0.4, 0.6)$ for English.

if we pre-process in different ways, we cause k to be different.

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: $\boxed{M = kT^b}$

where $k \in (30, 100)$ and $b \in (0.4, 0.6)$ for English.

if we pre-process in different ways, we cause k to be different.

NB number of types increases rapidly at first,

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

$$\text{Heap's Law: } \boxed{M = kT^b}$$

where $k \in (30, 100)$ and $b \in (0.4, 0.6)$ for English.

if we pre-process in different ways, we cause k to be different.

NB number of types increases rapidly at first, then less rapidly.

Heap's Law: Type-Token relationship

So pre-processing 'works' in the sense that it serves to simplify the problem.

but how does the total number of types M, change as total number of tokens T increases ?

Heap's Law: $\boxed{M = kT^b}$

where $k \in (30, 100)$ and $b \in (0.4, 0.6)$ for English.

if we pre-process in different ways, we cause k to be different.

NB number of types increases rapidly at first, then less rapidly. Need to pre-process, especially for long collections!

Zipf's Law

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term?

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

→ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth. . .

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

and third most common term is one third as common as most common,

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

and third most common term is one third as common as most common,

and fourth most common term is one quarter as common as most common,

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

and third most common term is one third as common as most common,

and fourth most common term is one quarter as common as most common,

etc Can rewrite as:

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

and third most common term is one third as common as most common,

and fourth most common term is one quarter as common as most common,

etc Can rewrite as: corpus frequency of $i = ci^k$ or

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

and third most common term is one third as common as most common,

and fourth most common term is one quarter as common as most common,

etc Can rewrite as: corpus frequency of $i = ci^k$ or $\log(\text{corpus frequency}) = \log c + k \log i$,

Zipf's Law

Heap's Law tells us about the relationship between tokens and terms.

but what about the relationship between the relative frequency of terms in the corpus?

$\rightarrow$ how much more common is the most common term relative to the second common term? What about relative to the the third most common term? And the fourth...

Zipf's Law: corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

so second most common term is half as common as most common,

and third most common term is one third as common as most common,

and fourth most common term is one quarter as common as most common,

etc Can rewrite as: corpus frequency of $i = ci^k$ or $\log(\text{corpus frequency}) = \log c + k \log i$, where i is the rank, $k = -1$.

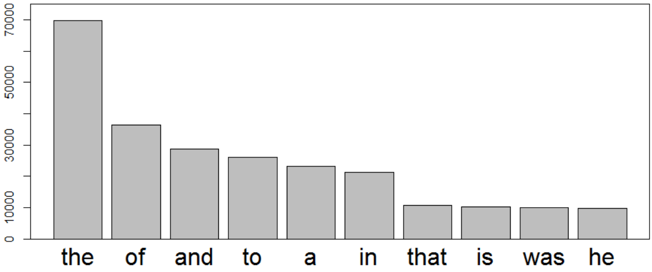
Brown Corpus (1961)

Brown Corpus (1961)

term	freq
the	69836
of	36365
and	28826
to	26126
a	23157
in	21314
that	10777
is	10182
was	9968
he	9801

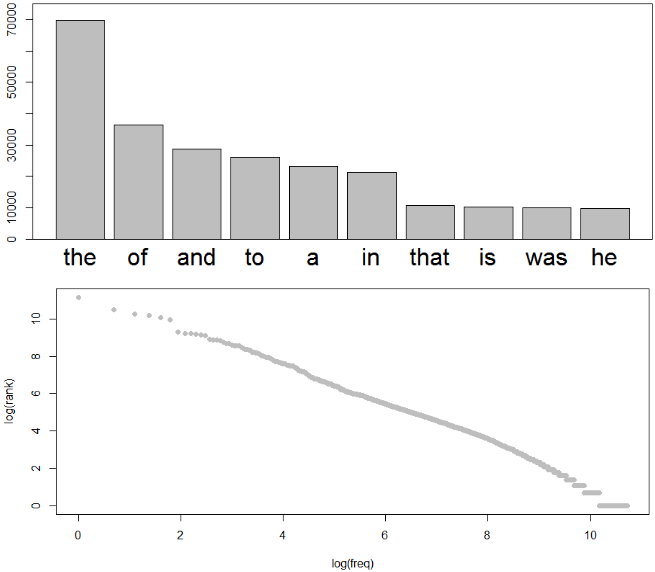
Brown Corpus (1961)

term	freq
the	69836
of	36365
and	28826
to	26126
a	23157
in	21314
that	10777
is	10182
was	9968
he	9801



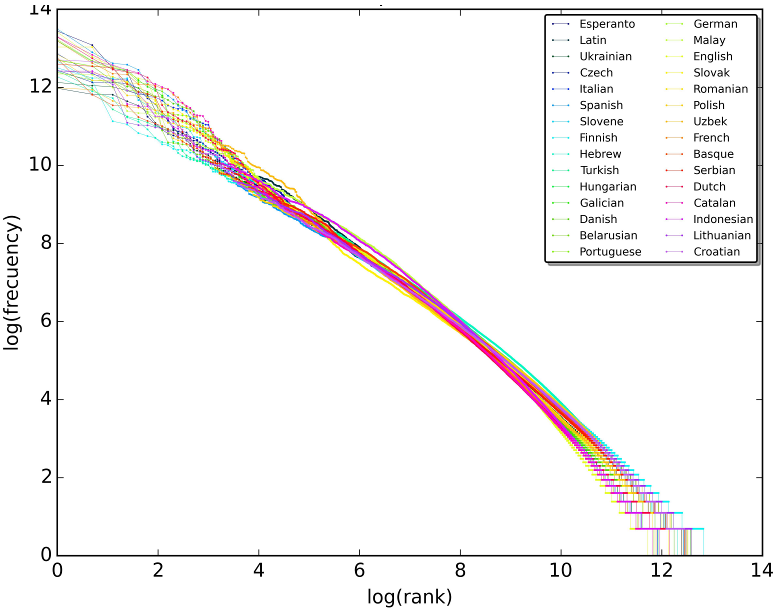
Brown Corpus (1961)

term	freq
the	69836
of	36365
and	28826
to	26126
a	23157
in	21314
that	10777
is	10182
was	9968
he	9801



Other Languages (Wikipedia)

Other Languages (Wikipedia)



Distance Metrics and Measures

Comparing Texts: Distance

Comparing Texts: Distance

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

Comparing Texts: Distance

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

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

Comparing Texts: Distance

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

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

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

Comparing Texts: Distance

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

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

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

$\rightarrow$ tells us about similarity of documents

Comparing Texts: Distance

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

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

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

$\rightarrow$ tells us about similarity of documents

and is typically required for application of multivariate techniques, anyway

Comparing Texts: Distance

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

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

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

$\rightarrow$ tells us about similarity of documents

and is typically required for application of multivariate techniques, anyway

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

Metrics vs Measures

Metrics vs Measures

NB not all measures of distance or similarity are metrics.

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1 no negative distances: $s_{ij} \geq 0$

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1 no negative distances: $s_{ij} \geq 0$

2 distance between documents is zero $\iff$ documents are identical

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1 no negative distances: $s_{ij} \geq 0$

2 distance between documents is zero $\iff$ documents are identical

3 distance between documents is symmetric:

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1. no negative distances: $s_{ij} \geq 0$
2. distance between documents is zero $\iff$ documents are identical
3. distance between documents is symmetric: $s_{ij} = s_{ji}$

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1. no negative distances: $s_{ij} \geq 0$
2. distance between documents is zero $\iff$ documents are identical
3. distance between documents is symmetric: $s_{ij} = s_{ji}$
4. measures satisfy triangle inequality. $s_{ik} \leq s_{ij} + s_{jk}$

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1. no negative distances: $s_{ij} \geq 0$
2. distance between documents is zero $\iff$ documents are identical
3. distance between documents is symmetric: $s_{ij} = s_{ji}$
4. measures satisfy triangle inequality. $s_{ik} \leq s_{ij} + s_{jk}$

i.e. if doc i is similar to doc j *and* doc j is similar to doc k,

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1. no negative distances: $s_{ij} \geq 0$
2. distance between documents is zero $\iff$ documents are identical
3. distance between documents is symmetric: $s_{ij} = s_{ji}$
4. measures satisfy triangle inequality. $s_{ik} \leq s_{ij} + s_{jk}$

i.e. if doc i is similar to doc j *and* doc j is similar to doc k, then doc i is similar to doc k

Metrics vs Measures

NB not all measures of distance or similarity are metrics. To be a metric, the measure of *distance* between documents i and j, s_{ij} must have certain properties:

1 no negative distances: $s_{ij} \geq 0$

2 distance between documents is zero $\iff$ documents are identical

3 distance between documents is symmetric: $s_{ij} = s_{ji}$

4 measures satisfy triangle inequality. $s_{ik} \leq s_{ij} + s_{jk}$

i.e. if doc i is similar to doc j *and* doc j is similar to doc k, then doc i is similar to doc k (we have an upper bound on how far apart they can be)

Euclidean Distance

Euclidean Distance

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

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

Euclidean Distance

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

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

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

Euclidean Distance

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

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

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

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

Euclidean Distance

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

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

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

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

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

Euclidean Distance

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

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

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

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

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

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

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space. Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

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

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

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

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

Euclidean Distance

The 'ordinary', 'straight line' distance between two points in space. Recall that $\mathbf{y}_i$ and $\mathbf{y}_j$ are documents,

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

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

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

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

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

larger distances imply lower similarity.

Partner exercise

Partner exercise

1. consider three documents in term frequency space:

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

   Which documents will Euclidean distance place closest together?

Partner exercise

1. consider three documents in term frequency space:

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

   Which documents will Euclidean distance place closest together? Why?

Partner exercise

1. consider three documents in term frequency space:

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

   Which documents will Euclidean distance place closest together? Why?

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

Partner exercise

1. consider three documents in term frequency space:

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

   Which documents will Euclidean distance place closest together? Why?

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

Better Approach

Better Approach

Euclidean distance rewards magnitude, rather than direction.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length,

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

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

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

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

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

Better Approach

Euclidean distance rewards magnitude, rather than direction.

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

We can do better by normalizing document length, and rewarding relatively similar uses of terms.

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

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

Cosine Similarity

Cosine Similarity

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

Cosine Similarity

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

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

Cosine Similarity

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

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

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

Cosine Similarity

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

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

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

and document length is controlled for.

Cosine Similarity

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

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

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

and document length is controlled for.

so intuitively,

Cosine Similarity

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

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

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

and document length is controlled for.

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

Cosine Similarity

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

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

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

and document length is controlled for.

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

Cosine Similarity

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

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

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

and document length is controlled for.

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

Example

Example

$\mathbf{y}_i = [2.3, 4.3]$; $\mathbf{y}_j = [3.9, 2.1]$

Example

$$\mathbf{y}_i = [2.3, 4.3];\ \mathbf{y}_j = [3.9, 2.1]$$

then $\mathbf{y}_i \cdot \mathbf{y}_j = [2.3 \times 3.9] + [4.3 \times 2.1] = 18.$

Example

$$\mathbf{y}_i = [2.3, 4.3];\ \mathbf{y}_j = [3.9, 2.1]$$

then $\mathbf{y}_i \cdot \mathbf{y}_j = [2.3 \times 3.9] + [4.3 \times 2.1] = 18.$

and $||\mathbf{y}_i|| = \sqrt{2.3^2 + 4.3^2} = 4.88;\ ||\mathbf{y}_j|| = \sqrt{3.9^2 + 2.1^2} = 4.43$

Example

$\mathbf{y}_i = [2.3, 4.3]$; $\mathbf{y}_j = [3.9, 2.1]$

then $\mathbf{y}_i \cdot \mathbf{y}_j = [2.3 \times 3.9] + [4.3 \times 2.1] = 18.$

and $||\mathbf{y}_i|| = \sqrt{2.3^2 + 4.3^2} = 4.88$; $||\mathbf{y}_j|| = \sqrt{3.9^2 + 2.1^2} = 4.43$

so $c_{ij} = \frac{18}{4.88 \times 4.43} =$

Example

$$\mathbf{y}_i = [2.3, 4.3];\ \mathbf{y}_j = [3.9, 2.1]$$

then $\mathbf{y}_i \cdot \mathbf{y}_j = [2.3 \times 3.9] + [4.3 \times 2.1] = 18.$

and $||\mathbf{y}_i|| = \sqrt{2.3^2 + 4.3^2} = 4.88;\ ||\mathbf{y}_j|| = \sqrt{3.9^2 + 2.1^2} = 4.43$

so $c_{ij} = \frac{18}{4.88 \times 4.43} = 0.83.$

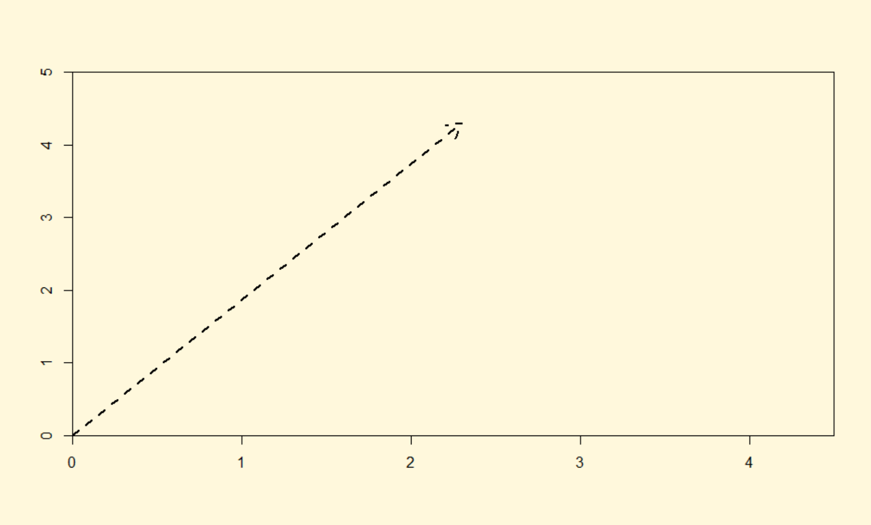
Graphically

Graphically

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

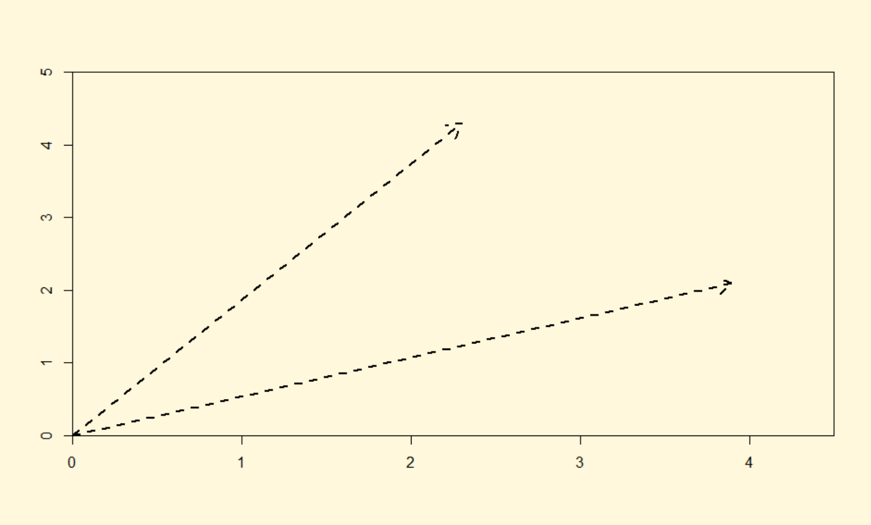
Graphically

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



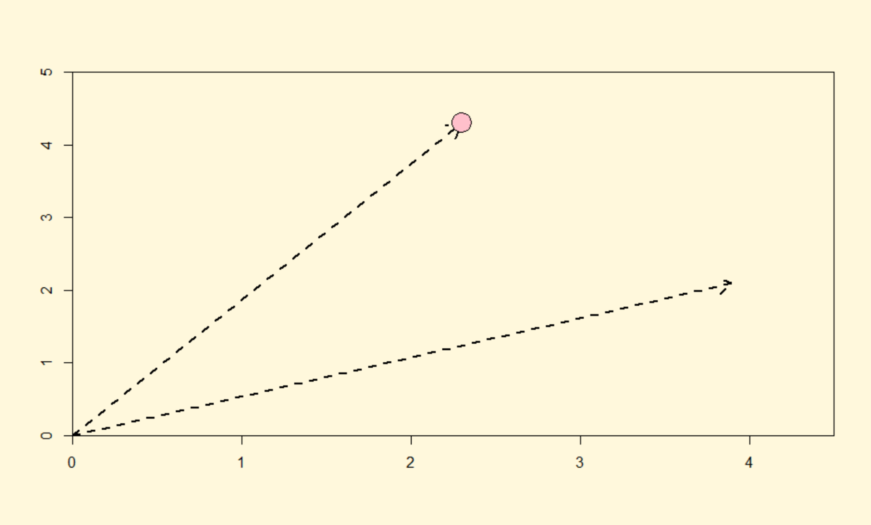
Graphically

$y_i = [2.3, 4.3]$; $y_j = [3.9, 2.1]$



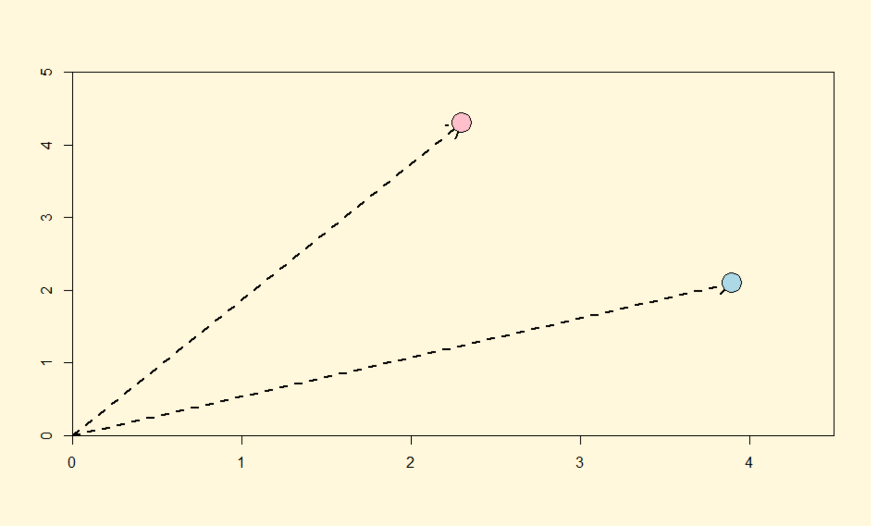
Graphically

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



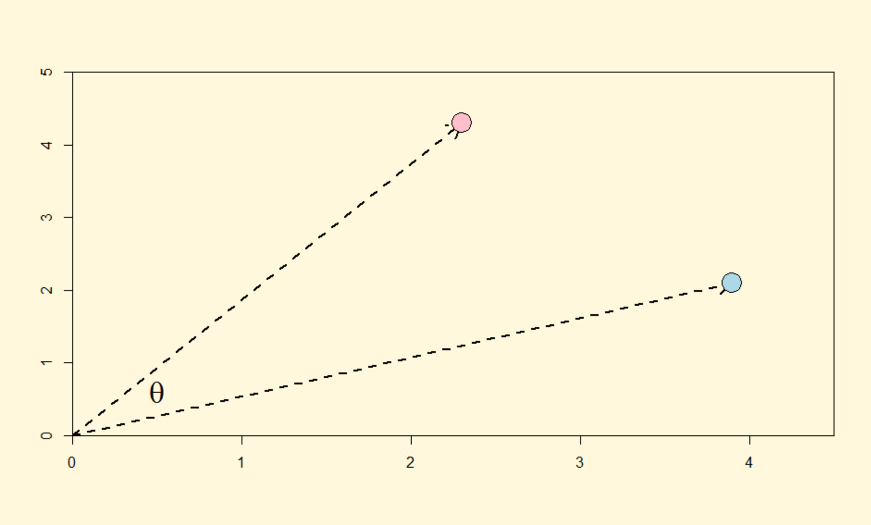
Graphically

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



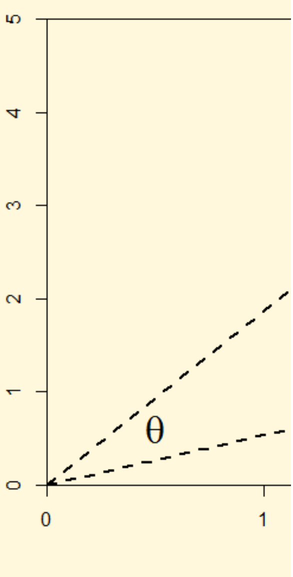
Graphically

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

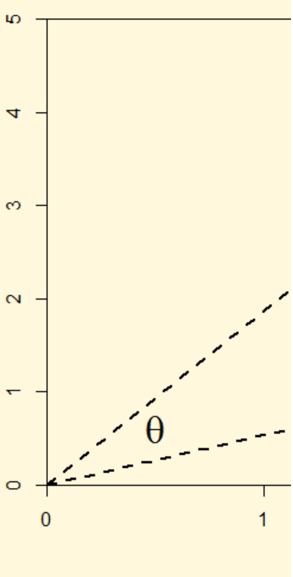


Algebra

Algebra

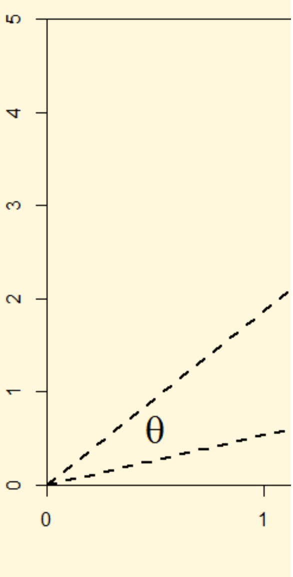


Algebra



know dot product of vectors:

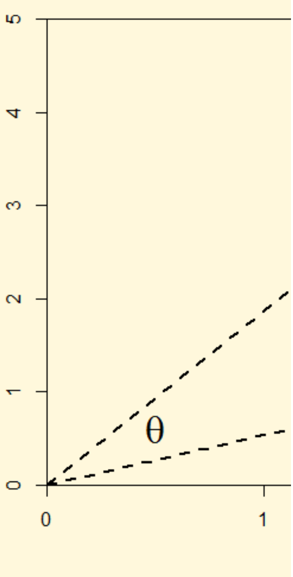
Algebra



know dot product of vectors:

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

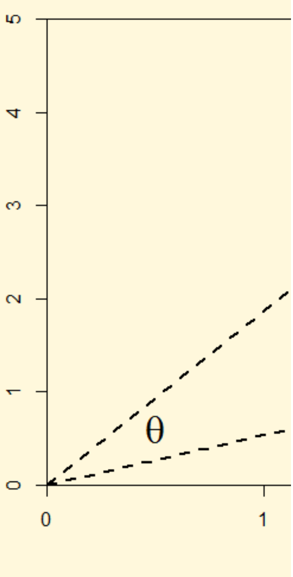
Algebra



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

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

Algebra

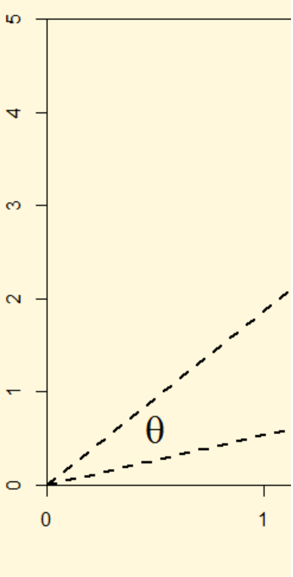


know dot product of vectors:
$$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$$

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

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

Algebra



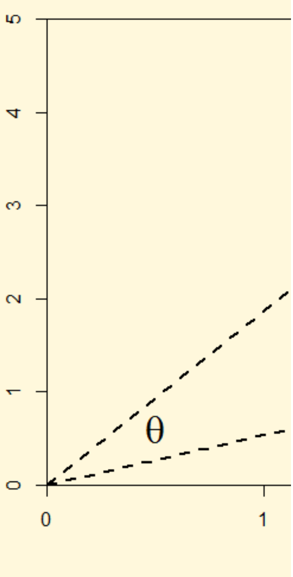
know dot product of vectors:
$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i|| ||\mathbf{y}_j|| \cos\theta$

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

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

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

Algebra



know dot product of vectors:

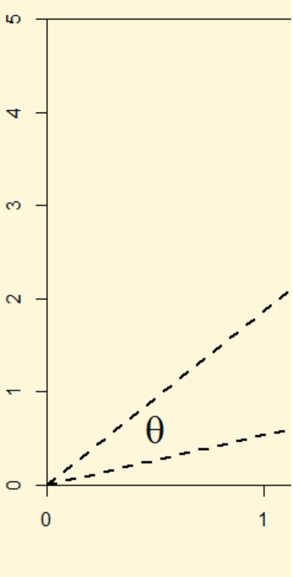
$$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i|| ||\mathbf{y}_j|| \cos\theta$$

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

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

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

Algebra



know dot product of vectors:

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

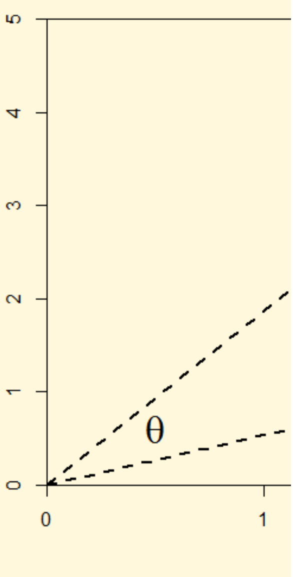
then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}$

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

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

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

Algebra



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

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

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

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

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

Looks about right.

1983 General Election Manifestos

1983 General Election Manifestos

1983 General Election Manifestos

- Labour manifesto as ‘longest suicide note in history’

1983 General Election Manifestos

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

1983 General Election Manifestos

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

1983 General Election Manifestos

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

- Conservative manifesto promised trade union curbs, deflation etc.

1983 General Election Manifestos

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

- Conservative manifesto promised trade union curbs, deflation etc.

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

1997 General Election Manifestos

1997 General Election Manifestos

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.
- ‘New Labour’ and ‘Third Way’

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

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

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

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

1997 General Election Manifestos

- Conservative manifesto promised continuation of moderate Major years.

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

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

Animals at the Zoo

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

> Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan:

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance.

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$.

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$,

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as ‘taxicab’ distance or ‘city block’ distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

Canberra:

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

Canberra: weighted version of Manhattan distance.

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

> Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).
>
> Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.
>
> Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

Minowski: generalized version of Euclidean and Manhattan.

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

Minowski: generalized version of Euclidean and Manhattan. $(\sum |\mathbf{y}_i - \mathbf{y}_j|^c)^{\frac{1}{c}}$.

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

Minowski: generalized version of Euclidean and Manhattan. $(\sum |\mathbf{y}_i - \mathbf{y}_j|^c)^{\frac{1}{c}}$. If c is 1, this is Manhattan. If c is 2, this is Euclidean.

Animals at the Zoo

- we can produce a cosine dissimilarity measure via $1 - c_{ij}$ (though not a metric)

but there are a large number of other distance measures on offer:

Jaccard: size of the intersection of the two documents (number of common words between the documents) divided by the size of the union of the two documents (total number of unique words in docs).

Manhattan: known as 'taxicab' distance or 'city block' distance. Absolute difference between coordinates: $||\mathbf{y}_i - \mathbf{y}_j|| = \sum |\mathbf{y}_i - \mathbf{y}_j|$. As we go from $\mathbf{y}_i$ to $\mathbf{y}_j$, have to do so at right angles: travel along, turn 90° and then up (or down), then turn 90° and go along, turn 90° etc.

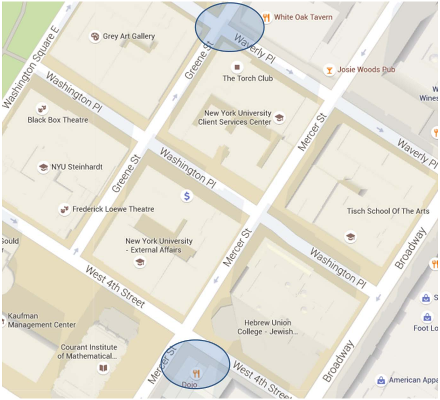
Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

Minowski: generalized version of Euclidean and Manhattan. $(\sum |\mathbf{y}_i - \mathbf{y}_j|^c)^{\frac{1}{c}}$. If c is 1, this is Manhattan. If c is 2, this is Euclidean.

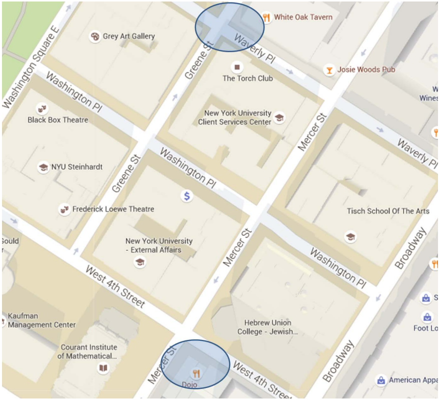
etc

Partner Exercise

Partner Exercise

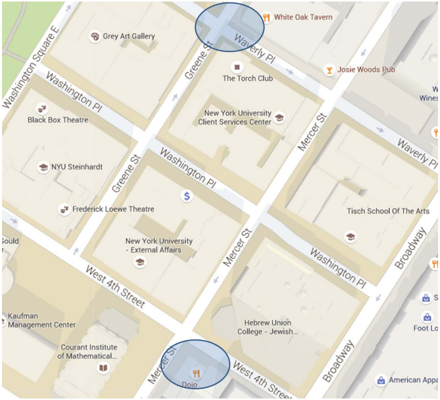


Partner Exercise



Suppose a block is one unit long and one unit wide.

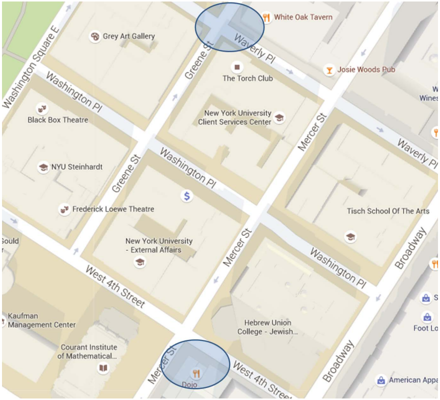
Partner Exercise



Suppose a block is one unit long and one unit wide.

- what is Euclidean distance between Dojo and White Oak Tavern?

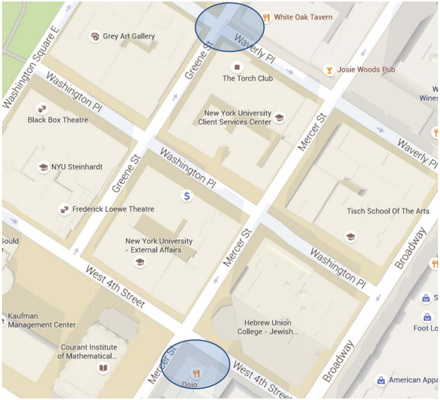
Partner Exercise



Suppose a block is one unit long and one unit wide.

- what is Euclidean distance between Dojo and White Oak Tavern?
- what is Manhattan distance between Dojo and White Oak Tavern?

Partner Exercise

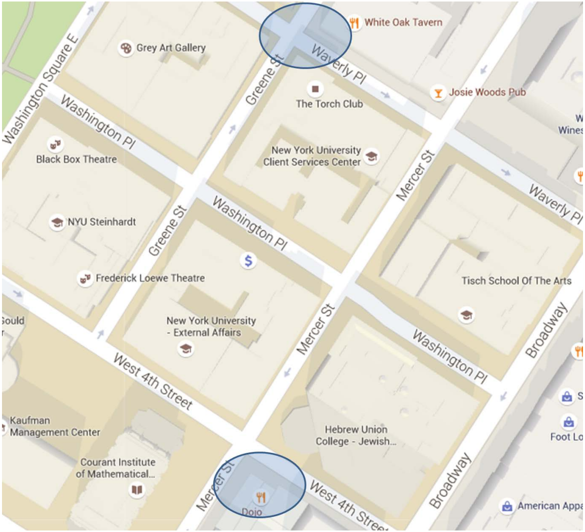


Suppose a block is one unit long and one unit wide.

- what is Euclidean distance between Dojo and White Oak Tavern?
- what is Manhattan distance between Dojo and White Oak Tavern?

Solution

Solution

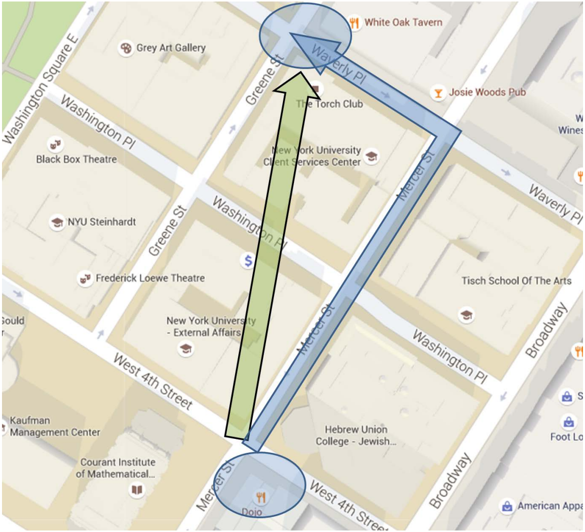


White Oak Tavern
Grey Art Gallery
Washington Square E
Waverly Pl
Greene St
The Torch Club
Josie Woods Pub
Washington Pl
Black Box Theatre
New York University
Client Services Center
Mercer St
Waverly Pl
NYU Steinhardt
Greene St
Washington Pl
Frederick Loewe Theatre
Tisch School Of The Arts
New York University
- External Affairs
Mercer St
Broadway
Washington Pl
West 4th Street
Kaufman
Management Center
Hebrew Union
College - Jewish...
Courant Institute
of Mathematical...
Mercer St
Dojo
West 4th Street
Broadway
American Appa

Solution

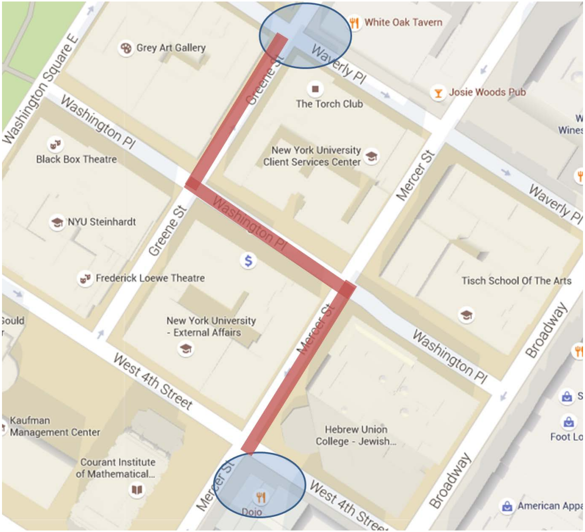
- Euclidean ($\sqrt{5}$)

Solution



- Euclidean ($\sqrt{5}$)
- Manhattan (3)

Solution



- Euclidean ($\sqrt{5}$)
- Manhattan (3)
- Manhattan (3)

Descriptive Statistics: Key Words in Context

Key Words in Context

Key Words in Context

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

Key Words in Context

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

Key Words in Context

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

$\rightarrow$ quick overview of general use,

Key Words in Context

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

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

Key Words in Context

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

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

also true in social science applications where we might want to understand how a given concept appears,

Key Words in Context

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

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

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

Key Words in Context

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

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

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

1 keyword of interest.

Key Words in Context

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

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

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

1. keyword of interest.
2. context —typically the sentence in which it appears.

Key Words in Context

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

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

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

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

Example: ‘democratic’ and the Second Reform Act

Example: ‘democratic’ and the Second Reform Act

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

A LEAP IN THE DARK.

Example: ‘democratic’ and the Second Reform Act

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

A LEAP IN THE DARK.

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

Example: ‘democratic’ and the Second Reform Act

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

A LEAP IN THE DARK.

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

Example: ‘democratic’ and the Second Reform Act

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

A LEAP IN THE DARK.

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

$\rightarrow$ represents approximate doubling of electorate.

Example: ‘democratic’ and the Second Reform Act

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

A LEAP IN THE DARK.

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

→ represents approximate doubling of electorate.

Debates of the time are lively and long.

Example: ‘democratic’ and the Second Reform Act

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

A LEAP IN THE DARK.

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

→ represents approximate doubling of electorate.

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

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

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

A LEAP IN THE DARK.

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

→ represents approximate doubling of electorate.

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

q What role did ‘`democratic`’ play in the debate?

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

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

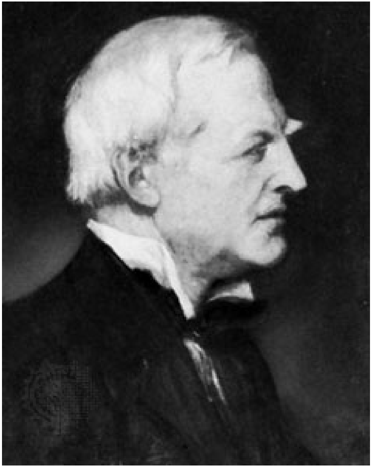
Detail: s267738.txt

Detail: s267738.txt

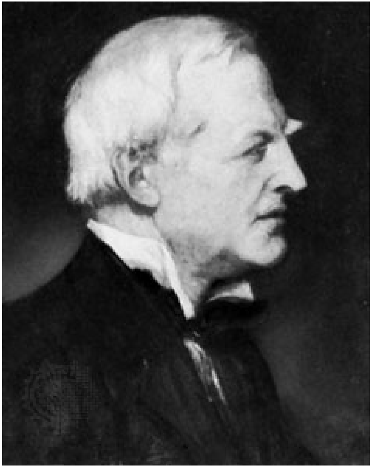
preword	word	postword
swept away in some further	democratic	change. And it is for
throne. When you get a	democratic	basis for your institutions, you
differences between ourselves and other	democratic	legislatures? Where is the democratic
democratic legislatures? Where is the	democratic	legislature which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

The Original Speaker and Speech

The Original Speaker and Speech

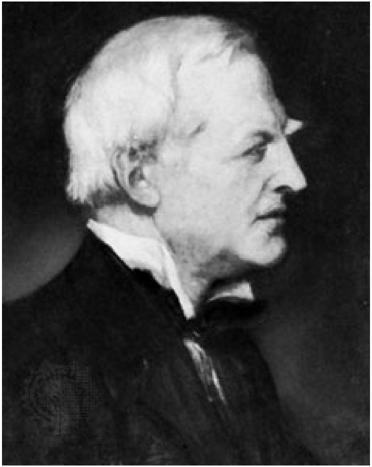


The Original Speaker and Speech



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

The Original Speaker and Speech



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

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

Partner Exercise

Partner Exercise

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

Partner Exercise

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

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

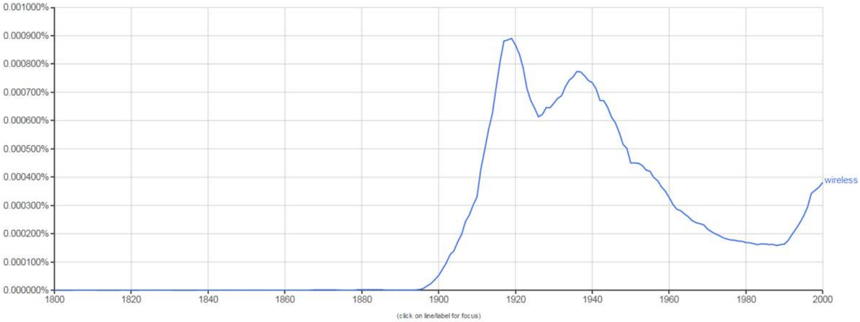
Partner Exercise

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

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

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

Use of ‘Wireless’



Descriptive Statistics: Diversity and Complexity

Lexical Diversity

Lexical Diversity

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

Lexical Diversity

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

Lexical Diversity

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

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

Lexical Diversity

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

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

thus we typically have many more tokens than types,

Lexical Diversity

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

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

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

Lexical Diversity

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

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

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

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

Lexical Diversity

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

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

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

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

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

Lexical Diversity

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

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

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

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

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

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

Tabloid vs Broadsheet

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

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

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

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


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.

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

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

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.

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

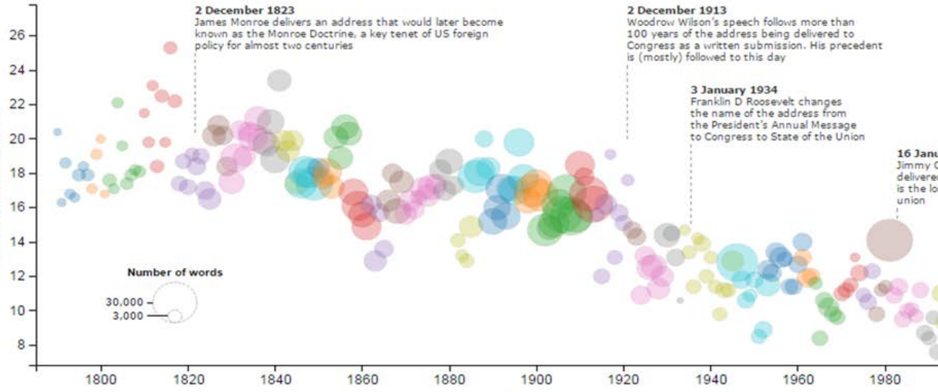
One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated (at text level).

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

The state of our union is ... dumber:

How the linguistic standard of the presidential address has declined

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



Leaders and their incentives

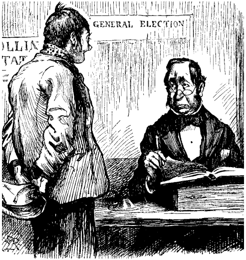
Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate



Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

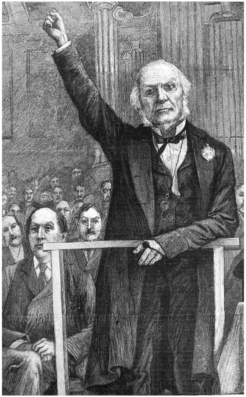
Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery. . .

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



Leaders and their incentives

- C19th Britain is notable for fast expansion of suffrage.
- new voters tended to be poorer and less literate
- ↓ local, clientalistic appeals via bribery. . .
- ↑ 'party orientated electorate', with national policies and national leaders
- Q how did these leaders respond to new voters?

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

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

Q how did these leaders respond to new voters?

A by changing nature of their speech:

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

↓ local, clientalistic appeals via bribery...

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

Q how did these leaders respond to new voters?

A by changing nature of their speech: simpler,

Leaders and their incentives

C19th Britain is notable for fast expansion of suffrage.

new voters tended to be poorer and less literate

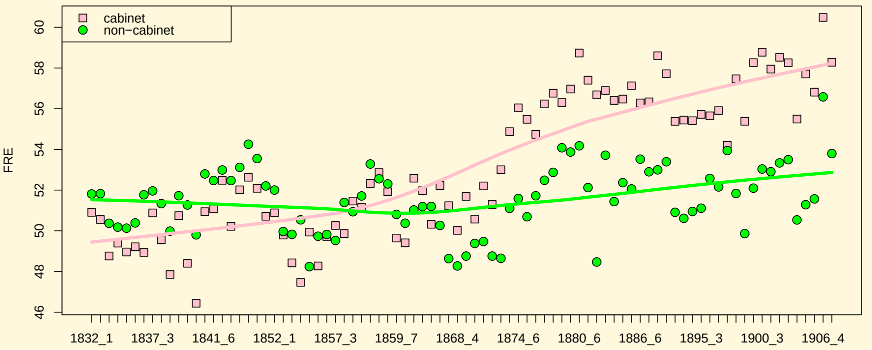
↓ local, clientalistic appeals via bribery...

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

Q how did these leaders respond to new voters?

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

Flesch overtime plot



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

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

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

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

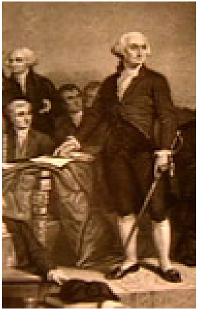
where PDW is percentage of difficult words,

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

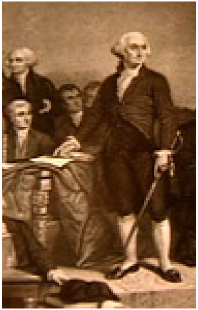
e.g. `about`, `back`, `call`, etc.

Partner Exercise

Partner Exercise

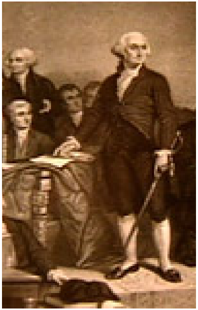


Partner Exercise



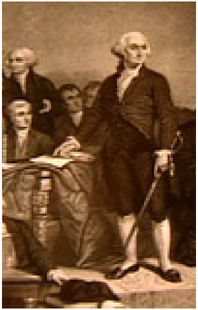
The FRE of SOTU speeches is declining. Why might it be difficult to make readability comparisons over time?

Partner Exercise



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

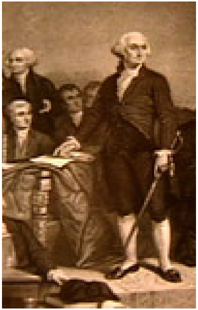
Partner Exercise



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

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

Partner Exercise



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

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

Descriptive Statistics: 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).

→ 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:

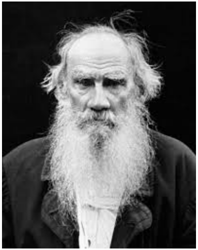
1. authors use them unconsciously
2. therefore, don't vary much by topic.

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).

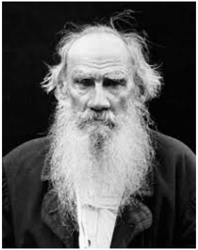
→ wrong, but models relying on these assns discriminate well (see Peng & Hengartner on e.g. Austin v Shakespeare)

Partner Exercise

Partner Exercise

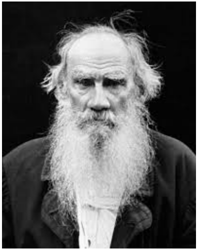


Partner Exercise



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

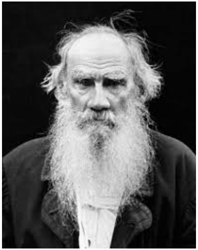
Partner Exercise



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

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

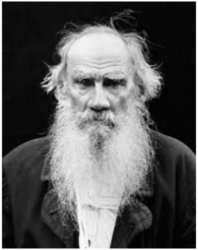
Partner Exercise



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

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

Partner Exercise

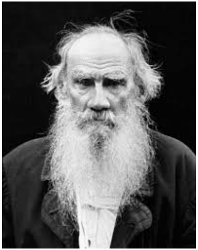


Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

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

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

Partner Exercise

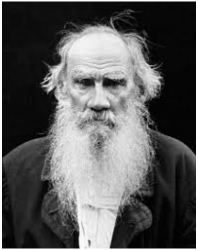


Suppose you wanted to compare the novels (and only the novels) of Leo Tolstoy and J.D. Salinger.

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

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

Partner Exercise



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

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

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

Sampling and Uncertainty

Sampling and Uncertainty

Sampling and Uncertainty

To now,

Sampling and Uncertainty

To now, we've been concerned with point estimates

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

But this is unsatisfying:

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

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

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

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

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

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

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

Sampling and Uncertainty

To now, we've been concerned with point estimates

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

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

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

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

Sampling Distributions: Reminder

Sampling Distributions: Reminder

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

Sampling Distributions: Reminder

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

Sampling Distributions: Reminder

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

We want the sampling distribution of sample mean,

Sampling Distributions: Reminder

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

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

Sampling Distributions: Reminder

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

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

$\rightarrow$ important,

Sampling Distributions: Reminder

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

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

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

Sampling Distributions: Reminder

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

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

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

Sampling Distributions: Reminder

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

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

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

Sampling Distributions: Reminder

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

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

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

Sampling Distributions: Reminder

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

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

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

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)— $\bar{x} \sim \left(\mu, \frac{\sigma^2}{n}\right)$.

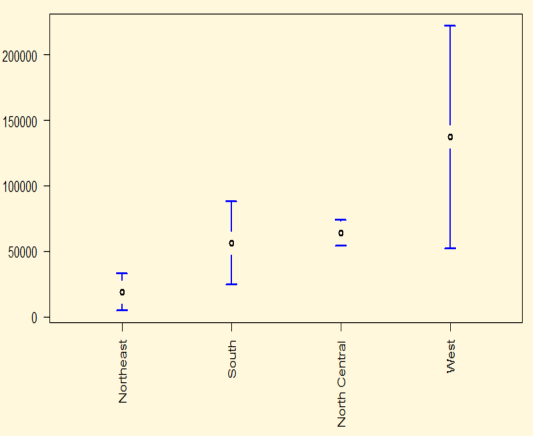
NB We call the standard deviation of the sampling distribution the standard error of the statistic.

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?

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

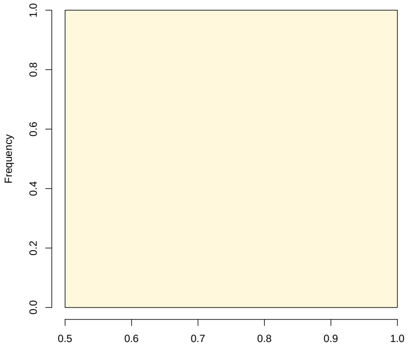
so resample data ($n = 20$ every time),
and record R^2—then plot...

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. . .

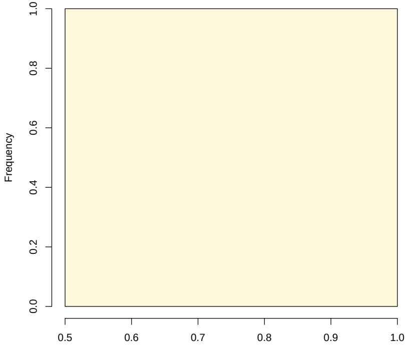


Bootstrap Example

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?

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

so tokens? paragraphs?

Bootstrap Unit

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

so tokens? paragraphs? sentences?

Bootstrap Unit

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

so tokens? paragraphs? sentences?

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

Bootstrap Unit

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

so tokens? paragraphs? sentences?

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

Bootstrap Unit

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

so tokens? paragraphs? sentences?

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

Bootstrap Unit

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

so tokens? paragraphs? sentences?

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

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

Bootstrap Unit

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

so tokens? paragraphs? sentences?

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

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

Bootstrap Unit

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

so tokens? paragraphs? sentences?

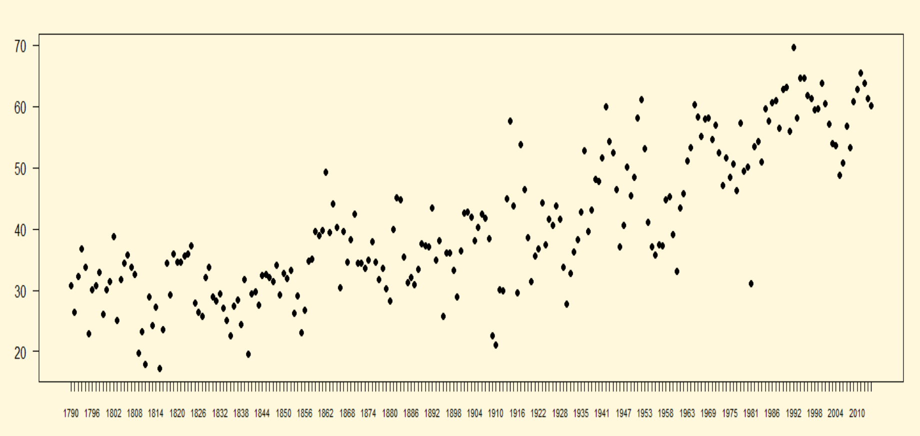
Benoit, Laver and Mikhalov (2009) use (quasi-)sentence-level bootstrap, which makes sense for manifestos: natural unit in which they are written.

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

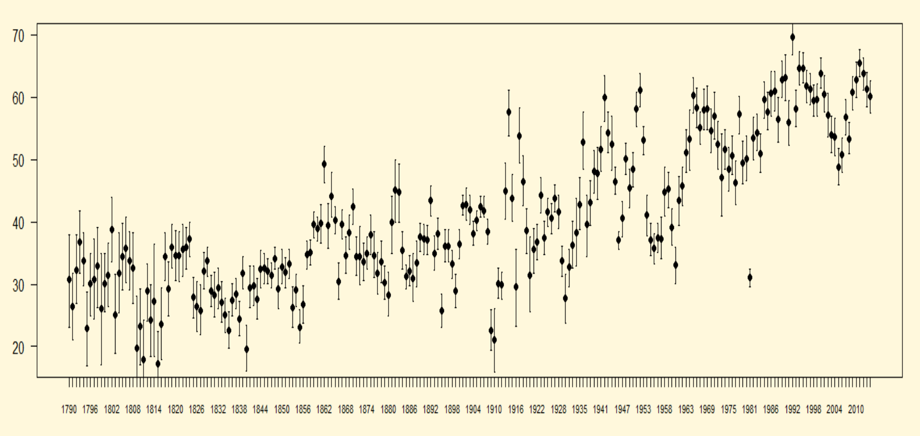
btw long texts give rise to smaller SEs than short ones, which makes sense!

SOU: 1000 bootstrap samples

SOU: 1000 bootstrap samples



SOU: 1000 bootstrap samples



Descriptive Statistics: Burstiness

Burstiness

Burstiness

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

Burstiness

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

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

Burstiness

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

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

Burstiness

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

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

Idea is to model arrival times of words.

Burstiness

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

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

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

Burstiness

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

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

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

Burstiness

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

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

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

Surges must be long,

Burstiness

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

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

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

Surges must be long, and/or intense,

Burstiness

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

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

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

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

Burstiness

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

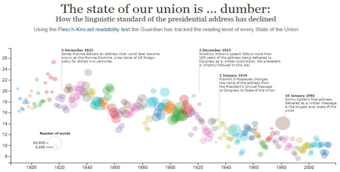
Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

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

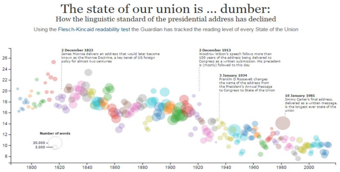
Surges must be long, and/or intense, depending on specification of model.

Applying to SOTU, 1790–2002

Applying to SOTU, 1790–2002



Applying to SOTU, 1790–2002



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

Burstiness: more details

Details

Details

denote the gap ‘times’ as x.

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

Details

denote the gap ‘times’ as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters,

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large,

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

Details

denote the gap ‘times’ as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down),

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down), we have evidence that a burst is occurring (ending).

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down), we have evidence that a burst is occurring (ending).

In principle,

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down), we have evidence that a burst is occurring (ending).

In principle, s could be estimated, but typically set to 2.

Details II

Details II

Here gaps are number of speeches between uses of a term.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches,

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word,

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

i.e. if $s = 2$, arrival rate has to increase (wrt to $\frac{1}{\alpha_0}$): $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

i.e. if $s = 2$, arrival rate has to increase (wrt to $\frac{1}{\alpha_0}$): $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

$\rightarrow$ This is a geometric process: unless wait time halves (somewhere), we never leave level $i = 0$.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

i.e. if $s = 2$, arrival rate has to increase (wrt to $\frac{1}{\alpha_0}$): $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

$\rightarrow$ This is a geometric process: unless wait time halves (somewhere), we never leave level $i = 0$.

$\Leftrightarrow$ this is an infinite state hidden Markov model.

Partner Exercise

Partner Exercise

1. Do we need to remove stop words when using calculating burstiness of given tokens? Why or why not?

Partner Exercise

1. Do we need to remove stop words when using calculating burstiness of given tokens? Why or why not?

2. Should we stem the words in the texts?

Partner Exercise

1. Do we need to remove stop words when using calculating burstiness of given tokens? Why or why not?

2. Should we stem the words in the texts?

3. How do models of the burstiness of words differ from 'topic models'? Which would you use to study changing subjects of debate over time? Which would you use to study conceptual change?

Cost Term

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter.

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter. Cost of moving from state i to state j is:

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter. Cost of moving from state i to state j is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where n is the total number of mentions;

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter. Cost of moving from state i to state j is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where n is the total number of mentions; γ is the parameter (set to 1).

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter. Cost of moving from state i to state j is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where n is the total number of mentions; γ is the parameter (set to 1).

so large changes to state 'upwards' will be costly; changing down is 'free'

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter. Cost of moving from state i to state j is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where n is the total number of mentions; γ is the parameter (set to 1).

so large changes to state 'upwards' will be costly; changing down is 'free'

and finding 'best' state sequence $\mathbf{q} = (q_{i1}, \ldots, q_{in})$ involves minimizing $c(\mathbf{q}|\mathbf{x}) = [\sum \tau(i,j)] + [\sum - \ln f_{it}(x_t)]$

Cost Term

Relying only on the α term would tend to produce a 'lot' of state changes

so use a second, cost parameter. Cost of moving from state i to state j is:

$$\tau(i,j) = \begin{cases} (j-i)\gamma \ln n & \text{if } j > i; \\ 0 & \text{if } j < i. \end{cases}$$

where n is the total number of mentions; γ is the parameter (set to 1).

so large changes to state 'upwards' will be costly; changing down is 'free'

and finding 'best' state sequence $\mathbf{q} = (q_{i1}, \ldots, q_{in})$ involves minimizing $c(\mathbf{q}|\mathbf{x}) = [\sum \tau(i,j)] + [\sum -\ln f_{it}(x_t)]$