1. Representing Text (Flipped)

DS-GA 1015, Text as Data
Arthur Spirling

February 9, 2021

Housekeeping

Housekeeping

1 Section has began! Make sure you attend.

Housekeeping

1 Section has began! Make sure you attend.

2 Materials—recordings, slides—now on Classes website.

Housekeeping

1. Section has began! Make sure you attend.
2. Materials—recordings, slides—now on Classes website.
3. Federal engagement requirement: do the online form in lab (or email). Or we have to report you are "unengaged".
4. Cohort B next week.

Goal of Text Analysis

Goal of Text Analysis

In many (most?) social science applications of text as data, we are trying to make an inference about a *latent variable*.

Goal of Text Analysis

In many (most?) social science applications of text as data, we are trying to make an inference about a *latent variable*.

$\boxed{\text{What is a latent variable?}}$

Goal of Text Analysis

In many (most?) social science applications of text as data, we are trying to make an inference about a *latent variable*.

$\boxed{\text{What is a latent variable?}}$

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

Goal of Text Analysis

In many (most?) social science applications of text as data, we are trying to make an inference about a *latent variable*.

$\boxed{\text{What is a latent variable?}}$

$\rightarrow$ something which we cannot observe directly but which we can make inferences about from things we can observe. Examples include ideology, ambition, narcissism, propensity to vote etc.

Goal of Text Analysis

In many (most?) social science applications of text as data, we are trying to make an inference about a *latent variable*.

$\boxed{\text{What is a latent variable?}}$

$\rightarrow$ something which we cannot observe directly but which we can make inferences about from things we can observe. Examples include ideology, ambition, narcissism, propensity to vote etc.

Sampling

Sampling

The corpus is made up of the documents within it,

Sampling

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

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets,

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often,

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases,

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties,

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

$\longrightarrow$ depending on your philosophical position,

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim$ 1% of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

$\longrightarrow$ depending on your philosophical position, you still need to think about sampling error.

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

$\longrightarrow$ depending on your philosophical position, you still need to think about sampling error. This is because there exists a superpopulation of populations from which the universe you observed came from.

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

$\longrightarrow$ depending on your philosophical position, you still need to think about sampling error. This is because there exists a superpopulation of populations from which the universe you observed came from.

Random error may not be the only concern:

Sampling

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

We sample for reasons of time, resources or (legal) necessity.

e.g. Twitter gives you $\sim 1\%$ of all their tweets, but it would presumably be prohibitively expensive to store 100%.

Often, authors claim to have the universe of cases in their corpus: *all* press releases, *all* treaties, *all* debate speeches.

$\longrightarrow$ depending on your philosophical position, you still need to think about sampling error. This is because there exists a superpopulation of populations from which the universe you observed came from.

Random error may not be the only concern: corpus should be representative in some well defined sense for inferences to be meaningful.

Exercise

Exercise

Exercise

You are consulting for a company who want to know what the world thinks of their product, a shampoo that slows balding in men.

Exercise

You are consulting for a company who want to know what the world thinks of their product, a shampoo that slows balding in men. They tell you to scrape Facebook data (timelines) as your corpus, and to analyze who is using it, and what they think of it.

Exercise

You are consulting for a company who want to know what the world thinks of their product, a shampoo that slows balding in men. They tell you to scrape Facebook data (timelines) as your corpus, and to analyze who is using it, and what they think of it.

Q Excluding any technical issues with the scraping,

Exercise

You are consulting for a company who want to know what the world thinks of their product, a shampoo that slows balding in men. They tell you to scrape Facebook data (timelines) as your corpus, and to analyze who is using it, and what they think of it.

Q Excluding any technical issues with the scraping, give three concerns about the validity of inferences from such a project.

II. Reducing Complexity

II. Reducing Complexity

- language is extraordinarily complex,

II. Reducing Complexity

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

II. Reducing Complexity

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

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

II. Reducing Complexity

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

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

Q $\boxed{\text{What do we mean by this?}}$

II. Reducing Complexity

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

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

Q $\boxed{\text{What do we mean by this?}}$

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

II. Reducing Complexity

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

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

Q $\boxed{\text{What do we mean by this?}}$

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

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

II. Reducing Complexity

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

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

Q $\boxed{\text{What do we mean by this?}}$

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

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

II. Reducing Complexity

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

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

Q $\boxed{\text{What do we mean by this?}}$

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

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

II. Reducing Complexity

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

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

Q $\boxed{\text{What do we mean by this?}}$

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

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

From Texts to Numeric Data

From Texts to Numeric Data

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

From Texts to Numeric Data

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

2. strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.
2. strip away ‘superfluous’ material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.
2. strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.
4. add descriptive annotations that preserve context: tagging.

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.
2. strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.
4. add descriptive annotations that preserve context: tagging.
5. map tokens back to common form: lemmatization, stemming.

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.
2. strip away ‘superfluous’ material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.
4. add descriptive annotations that preserve context: tagging.
5. map tokens back to common form: lemmatization, stemming.

operate/model.

From Texts to Numeric Data

1. collect raw text in machine readable/electronic form. Decide what constitutes a document.
2. strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.
4. add descriptive annotations that preserve context: tagging.
5. map tokens back to common form: lemmatization, stemming.

operate/model.

Q $\boxed{\text{What do we call the creation/curation of features before we model?}}$

Quick Note on Terminology

Quick Note on Terminology

$\boxed{\text{What is a type?}}$

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ a unique sequence of characters that are grouped together in some meaningful way.

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ a unique sequence of characters that are grouped together in some meaningful way. Mostly a word (for us),

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ a unique sequence of characters that are grouped together in some meaningful way. Mostly a word (for us), but might also be a word plus punctuation,

Quick Note on Terminology

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

Quick Note on Terminology

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

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

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`'

$\boxed{\text{What is a token ?}}$

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`’

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`'

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

e.g. "`Dog eat dog world`",

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`’

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

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

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`'

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

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

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`'

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

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

$\boxed{\text{What is a term ?}}$

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`'

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

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

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

Quick Note on Terminology

$\boxed{\text{What is a type?}}$ 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`'

$\boxed{\text{What is a token ?}}$ particular *instance* of type.

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

$\boxed{\text{What is a term ?}}$ 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.

Exercise

Exercise

Q Mostly we use whitespace to define subunits for tokenization, but this doesn't work in some applications and some languages. Explain why, give an example.

Exercise

Q Mostly we use whitespace to define subunits for tokenization, but this doesn't work in some applications and some languages. Explain why, give an example.

Q We talked about some common stop words.

Exercise

Q Mostly we use whitespace to define subunits for tokenization, but this doesn't work in some applications and some languages. Explain why, give an example.

Q We talked about some common stop words. Give an example of a stop word you would add to the list for an application in your field.

Exercise

Exercise

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

Exercise

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

Exercise

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

Exercise

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

1 `The mountains are beautiful in Ore. and Wash.`

2 `http://www.wsj.com/articles/son-of-saul-not-about-the-survivors-1449590175`

3 `I can't go with him to Beijing.`

Exercise: Word Order

Exercise: Word Order

What is bag of words?

Exercise: Word Order

What is bag of words? Why does it make storing the information in our documents easy?

Exercise: Word Order

What is bag of words? Why does it make storing the information in our documents easy?

How could we retain word order if we wanted it?

Exercise: Word Order

What is bag of words? Why does it make storing the information in our documents easy?

How could we retain word order if we wanted it?

What is a bigram?

Exercise: Word Order

What is bag of words? Why does it make storing the information in our documents easy?

How could we retain word order if we wanted it?

What is a bigram? What is a trigram ? What would they be for the following passage?

Exercise: Word Order

What is bag of words? Why does it make storing the information in our documents easy?

How could we retain word order if we wanted it?

What is a bigram? What is a trigram ? What would they be for the following passage? Do any of them have a frequency > 1?

Exercise: Word Order

What is bag of words? Why does it make storing the information in our documents easy?

How could we retain word order if we wanted it?

What is a bigram? What is a trigram ? What would they be for the following passage? Do any of them have a frequency > 1?

`This is America's day. This is democracy's day. A day of history and hope.`

Exercise: Vector Space Model

Exercise: Vector Space Model

Removing punctuation (leave `'s`) but no other preprocessing, what would be the vector space representation of this passage?

Exercise: Vector Space Model

Removing punctuation (leave `'s`) but no other preprocessing, what would be the vector space representation of this passage?

`This is America's day. This is democracy's day. A day of history and hope.`

Exercise: Vector Space Model

Removing punctuation (leave `'s`) but no other preprocessing, what would be the vector space representation of this passage?

`This is America's day. This is democracy's day. A day of history and hope.`

Solution

Exercise: Vector Space Model

Removing punctuation (leave `’s`) but no other preprocessing, what would be the vector space representation of this passage?

`This is America’s day. This is democracy’s day. A day of history and hope.`

Solution

sort `A America’s and day day day democracy’s history hope is is of This This`

then (1,1,1,3,1,1,1,2,1,2)

Introducing tf-idf

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w.

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w. When the word is common in the corpus,

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w. When the word is common in the corpus, this will be a small number. When the word is rare,

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w. When the word is common in the corpus, this will be a small number. When the word is rare, this will be a large number.

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w. When the word is common in the corpus, this will be a small number. When the word is rare, this will be a large number.

$\boxed{tf_{dw} \cdot \ln \frac{|D|}{df_w}}$, term frequency-inverse document frequency: tf-idf.

Introducing tf-idf

- tf_{dw}, term frequency: number of times word w appears in document d
- df_w, document frequency: number of documents in the collection of documents that contain word w

- $\ln \frac{|D|}{df_w}$, inverse document frequency: (natural) log of the total size of the corpus $|D|$ divided by the number of documents in the collection of documents that contain word w. When the word is common in the corpus, this will be a small number. When the word is rare, this will be a large number.

$\boxed{tf_{dw} \cdot \ln \frac{|D|}{df_w}}$, term frequency-inverse document frequency: tf-idf.

Some data from the 1980s

Some data from the 1980s

Suppose

	features							
	senator	hatfield	mr	chief	justice	president	vice	bush
1981	2	1	3	1	1	5	2	1
1985	4	0	0	1	1	3	1	1
1989	2	0	6	1	2	6	1	0

Some data from the 1980s

Suppose

	features							
	senator	hatfield	mr	chief	justice	president	vice	bush
1981	2	1	3	1	1	5	2	1
1985	4	0	0	1	1	3	1	1
1989	2	0	6	1	2	6	1	0

What is tf-idf for senator in 1981?

Some data from the 1980s

Suppose

	features							
	senator	hatfield	mr	chief	justice	president	vice	bush
1981	2	1	3	1	1	5	2	1
1985	4	0	0	1	1	3	1	1
1989	2	0	6	1	2	6	1	0

What is tf-idf for `senator` in 1981?

What is tf-idf for `mr` in 1989?

Solution

Solution

1981 '`senator`' is used 2 times.

Solution

1981 ‘`senator`’ is used 2 times. So, $tf = 2$.

Solution

1981 ‘`senator`’ is used 2 times. So, *tf* $=2$.

and in the 3 speeches (our corpus),

Solution

1981 ‘`senator`’ is used 2 times. So, tf=2.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

Solution

1981 ‘`senator`’ is used 2 times. So, $tf = 2$.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the *idf* is $\ln \frac{|D|}{df} = \ln \left(\frac{3}{3}\right)$

Solution

1981 '`senator`' is used 2 times. So, $tf = 2$.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{3}\right) = 0$

Solution

1981 ‘`senator`’ is used 2 times. So, *tf* =2.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the *idf* is $\ln \frac{|D|}{df} = \ln \left(\frac{3}{3}\right) = 0$

$\rightarrow$ tf-idf=0 for ‘`senator`’ in 1981.

Solution

1981 ‘`senator`’ is used 2 times. So, $tf = 2$.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the *idf* is $\ln \frac{|D|}{df} = \ln \left(\frac{3}{3}\right) = 0$

$\rightarrow$ tf-idf=0 for ‘`senator`’ in 1981.

but ‘`mr`’ is used 6 times in 1989, 3 times in 1981 and not at all in 1985.

Solution

1981 ‘`senator`’ is used 2 times. So, tf=2.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the idf is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{3}\right) = 0$

$\rightarrow$ tf-idf=0 for ‘`senator`’ in 1981.

but ‘`mr`’ is used 6 times in 1989, 3 times in 1981 and not at all in 1985.

so idf is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{2}\right)$

Solution

1981 ‘`senator`’ is used 2 times. So, $tf = 2$.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{3}\right) = 0$

$\rightarrow$ tf-idf=0 for ‘`senator`’ in 1981.

but ‘`mr`’ is used 6 times in 1989, 3 times in 1981 and not at all in 1985.

so *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{2}\right) = 0.41$

Solution

1981 ‘`senator`’ is used 2 times. So, $tf=2$.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{3}\right) = 0$

$\rightarrow$ tf-idf=0 for ‘`senator`’ in 1981.

but ‘`mr`’ is used 6 times in 1989, 3 times in 1981 and not at all in 1985.

so *idf* is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{2}\right) = 0.41$

$\rightarrow$ tf-idf=$0.41 \times 6 = 2.46$ for ‘`mr`’ in 1989.

Solution

1981 '`senator`' is used 2 times. So, tf=2.

and in the 3 speeches (our corpus), it is used (at least once) in *every* speech. So, $|D| = 3$ and $df = 3$

so the idf is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{3}\right) = 0$

$\rightarrow$ tf-idf=0 for '`senator`' in 1981.

but '`mr`' is used 6 times in 1989, 3 times in 1981 and not at all in 1985.

so idf is $\ln \frac{|D|}{df} = \ln\left(\frac{3}{2}\right) = 0.41$

$\rightarrow$ tf-idf=$0.41 \times 6 = 2.46$ for '`mr`' in 1989.

Exercise

Exercise

- Why do we log the idf part in tf-idf?

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)

- Suppose the goal is *rank* terms by their tf-idf: does the base of the logarithm matter?

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)
- Suppose the goal is *rank* terms by their tf-idf: does the base of the logarithm matter?
- Consider comparing two novels from Tolstoy in terms of the common (weighted) terms they contain. Now consider comparing two tweets.

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)

- Suppose the goal is *rank* terms by their tf-idf: does the base of the logarithm matter?

- Consider comparing two novels from Tolstoy in terms of the common (weighted) terms they contain. Now consider comparing two tweets. Which set of documents will, on average, have more elements in common?

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)
- Suppose the goal is *rank* terms by their tf-idf: does the base of the logarithm matter?
- Consider comparing two novels from Tolstoy in terms of the common (weighted) terms they contain. Now consider comparing two tweets. Which set of documents will, on average, have more elements in common? Why?

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)
- Suppose the goal is *rank* terms by their tf-idf: does the base of the logarithm matter?
- Consider comparing two novels from Tolstoy in terms of the common (weighted) terms they contain. Now consider comparing two tweets. Which set of documents will, on average, have more elements in common? Why? What should we do about this?

Exercise

- Why do we log the idf part in tf-idf? (hint: think about how we'd like idf to react to very rare vs fairly rare words)
- Suppose the goal is *rank* terms by their tf-idf: does the base of the logarithm matter?
- Consider comparing two novels from Tolstoy in terms of the common (weighted) terms they contain. Now consider comparing two tweets. Which set of documents will, on average, have more elements in common? Why? What should we do about this?

Extra: Accents

Extra: Accents

Rarely an issue in English,

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`.

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

More of a concern in other languages, but mostly when accent completely changes meaning:

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

More of a concern in other languages, but mostly when accent completely changes meaning: `pena` vs `peña`.

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

More of a concern in other languages, but mostly when accent completely changes meaning: `pena` vs `peña`. Perhaps map back to non-accented words (look-up table), or make use of specific unicode (if available)?

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

More of a concern in other languages, but mostly when accent completely changes meaning: `pena` vs `peña`. Perhaps map back to non-accented words (look-up table), or make use of specific unicode (if available)?

In practice, often written same way in casual communication (emails, search queries),

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

More of a concern in other languages, but mostly when accent completely changes meaning: `pena` vs `peña`. Perhaps map back to non-accented words (look-up table), or make use of specific unicode (if available)?

In practice, often written same way in casual communication (emails, search queries), and disambiguation can be hard!

Extra: Accents

Rarely an issue in English, though we might want to make sure `cliché` is treated as `cliche`. Generally, preprocessing gets rid of accents.

More of a concern in other languages, but mostly when accent completely changes meaning: `pena` vs `peña`. Perhaps map back to non-accented words (look-up table), or make use of specific unicode (if available)?

In practice, often written same way in casual communication (emails, search queries), and disambiguation can be hard!

Grammatical gender often removed via stopping.