2. Descriptive Inference I

DS-GA 3001, Text as Data
Arthur Spirling

February 6, 2018

Housekeeping

Housekeeping

1 Section in full swing!

Housekeeping

1. Section in full swing!
2. OH today

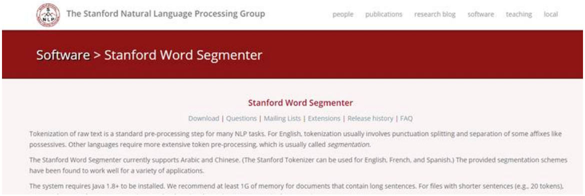
Housekeeping

1. Section in full swing!

2. OH today

3. Speaker series Thursday: Bruno Gonçalves on “Spatio temporal analysis of language use”.

Follow-up: Tokenize Chinese/Arabic



The Stanford Natural Language Processing Group

people publications research blog software teaching local

Software > Stanford Word Segmenter

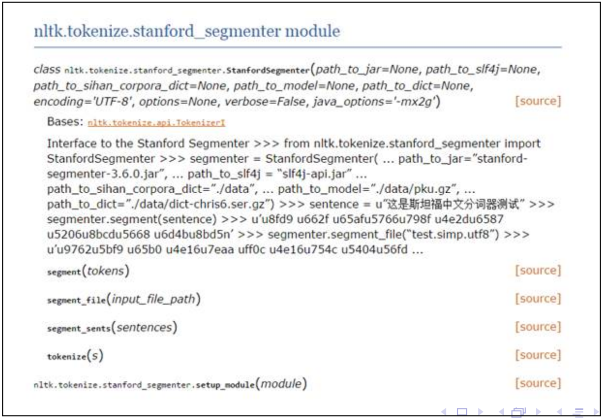
Stanford Word Segmenter

Download | Questions | Mailing Lists | Extensions | Release history | FAQ

Tokenization of raw text is a standard pre-processing step for many NLP tasks. For English, tokenization usually involves punctuation splitting and separation of some affixes like possessives. Other languages require more extensive token pre-processing, which is usually called *segmentation*.

The Stanford Word Segmenter currently supports Arabic and Chinese. (The Stanford Tokenizer can be used for English, French, and Spanish.) The provided segmentation schemes have been found to work well for a variety of applications.

The system requires Java 1.8+ to be installed. We recommend at least 1G of memory for documents that contain long sentences. For files with shorter sentences (e.g., 20 tokens),



nltk.tokenize.stanford_segmenter module

class `nltk.tokenize.stanford_segmenter.StanfordSegmenter`(*path_to_jar=None, path_to_slf4j=None, path_to_sihan_corpora_dict=None, path_to_model=None, path_to_dict=None, encoding='UTF-8', options=None, verbose=False, java_options='-mx2g'*) [source]

Bases: `nltk.tokenize.api.TokenizerI`

Interface to the Stanford Segmenter >>> from nltk.tokenize.stanford_segmenter import StanfordSegmenter >>> segmenter = StanfordSegmenter(... path_to_jar="stanford-segmenter-3.6.0.jar", ... path_to_slf4j = "slf4j-api.jar" ... path_to_sihan_corpora_dict="./data", ... path_to_model="./data/pku.gz", ... path_to_dict="./data/dict-chris6.ser.gz") >>> sentence = u"这是斯坦福中文分词器测试" >>> segmenter.segment(sentence) >>> u'u8fd9 u662f u65afu5766u798f u4e2du6587 u5206u8bcdu5668 u6d4bu8bd5n' >>> segmenter.segment_file("test.simp.utf8") >>> u'u9762u5bf9 u65b0 u4e16u7eaa uff0c u4e16u754c u5404u56fd ...

`segment`(*tokens*) [source]

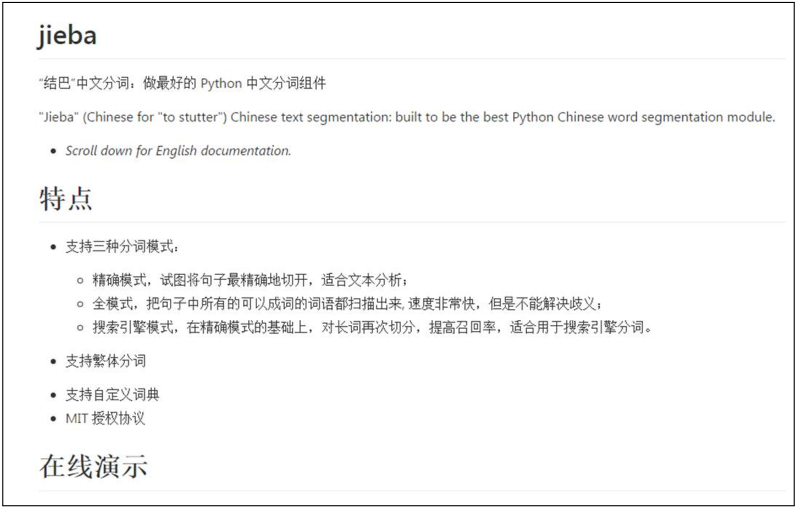
`segment_file`(*input_file_path*) [source]

`segment_sents`(*sentences*) [source]

`tokenize`(*s*) [source]

`nltk.tokenize.stanford_segmenter.setup_module`(*module*) [source]

Follow-up: Tokenize Chinese



jieba

"结巴"中文分词：做最好的 Python 中文分词组件

"Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation module.

- *Scroll down for English documentation.*

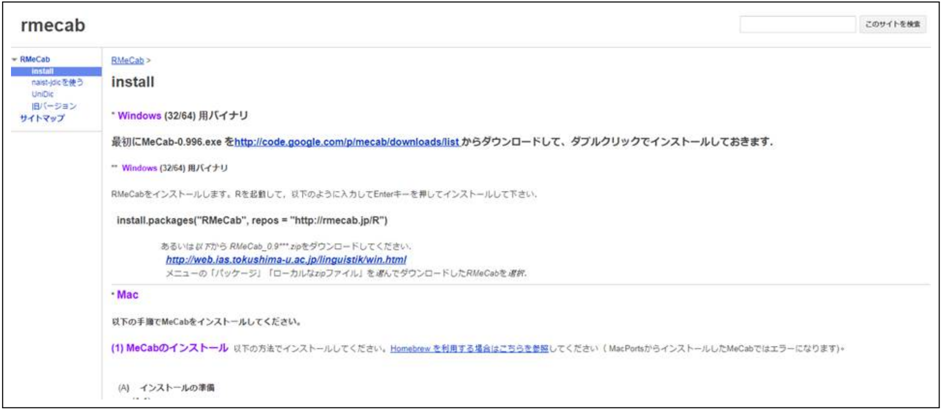
特点

- 支持三种分词模式：
  - 精确模式，试图将句子最精确地切开，适合文本分析；
  - 全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；
  - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。
- 支持繁体分词
- 支持自定义词典
- MIT 授权协议

在线演示

`https://github.com/fxsjy/jieba`

Follow-up: Tokenize Japanese



rmecab

install

* Windows (32/64) 用バイナリ

最初にMeCab-0.996.exe をhttp://code.google.com/p/mecab/downloads/list からダウンロードして、ダブルクリックでインストールしておきます.

** Windows (32/64) 用バイナリ

RMeCabをインストールします。Rを起動して，以下のように入力してEnterキーを押してインストールして下さい.

```
install.packages("RMeCab", repos = "http://rmecab.jp/R")
```

あるいは以下から RMeCab_0.9***.zipをダウンロードしてください.
http://web.ias.tokushima-u.ac.jp/linguistik/win.html
メニューの「パッケージ」「ローカルなzipファイル」を選んでダウンロードしたRMeCabを選択.

· Mac

以下の手順でMeCabをインストールしてください。

(1) MeCabのインストール 以下の方法でインストールしてください。Homebrew を利用する場合はこちらを参照してください（MacPortsからインストールしたMeCabではエラーになります)。

(A) インストールの準備

Follow-up: Accents

Follow-up: Accents

Rarely an issue in English,

Follow-up: Accents

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

Follow-up: Accents

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

Follow-up: 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:

Follow-up: 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 `pẽna`.

Follow-up: 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 `pẽna`. Perhaps map back to non-accented words (look-up table), or make use of specific unicode (if available)?

Follow-up: 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 `pẽna`. 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),

Follow-up: 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 `pẽna`. 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!

Follow-up: 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 `pẽna`. 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.

Where Are We?

Where Are We?

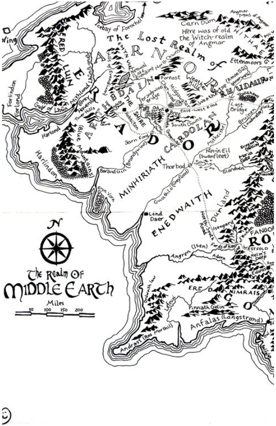
The Realm of
MIDDLE EARTH
Miles
50
100
150
200
N
The Lost Realm of ARNOR
Here was of old the Witch-realm of Angmar
Carn Dûm
Angmar
Mtns of Angmar
Icebay of Forochel
ERED LUIN
ERIADOR
ARTHEDAIN
CARDOLAN
RHUDAUR
Forlindon
Forlond
Gulf of Lhûn
Mithlond (Grey Havens)
Lhûn (Lune)
Harlond
Harlindon
Blue Mountains
Tower Hills
Hills of Evendim
EMYN UIAL
Nenuial (Lake Evendim)
North Downs
Fornost
Weather Hills
Weathertop (Amon Sûl)
Hoarwell
Ettenmoors
Trollshaws
Last Bridge
Bruinen (Loudwater)
Ford
THE SHIRE
White Downs
Hobbiton
Brandywine Bridge
Old Forest
Bree
Chetwood
Midgewater
East-West Road
South Downs
Greenway
Sarn Ford
Mitheithel
Baranduin (Brandywine)
MINHIRIATH
Tharbad
Nîn-in-Eilph (Swanfleet)
Glanduin
Eregion (Hollin)
Moria Gate
Gwathlo (Greyflood)
Eryn Vorn
Lond Daer
ENEDWAITH
Dunland
North-South Road
Isengard
FANGORN
ROHAN
Fords of Isen
Gap of Rohan
WESTFOLD
Helm's Deep
Edoras
Angren (Isen)
Adorn
ERED NIMRAIS
GONDOR
Druwaith Iaur
Lefnui
Pinnath Gelin
Anfalas (Langstrand)
Andrast (Ras Morthil)

☺

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

Where Are We?

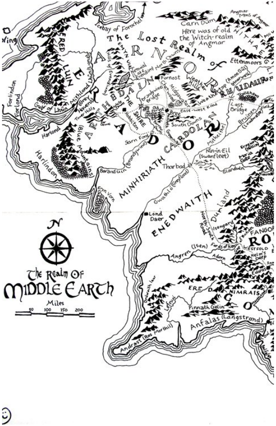
Our fundamental unit of text analysis is the document term matrix.

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

This is 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.

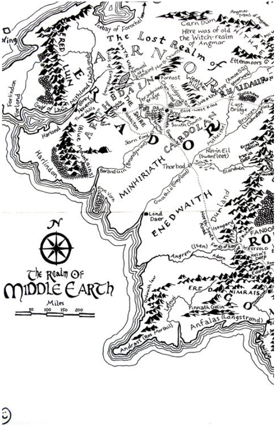
This is a set of stacked vectors, with each entry in each vector representing the ‘amount’ of a particular term.

This is 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 is 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: From Texts to Numeric Data

Reminder: From Texts to Numeric Data

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

Reminder: 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.

Reminder: 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.

Reminder: 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.

Reminder: 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.

Reminder: From Texts to Numeric Data

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

2. strip away 'superfluous' material: HTML tags, capitalization, punctuation, stop words etc.
3. cut document up into useful elementary pieces: tokenization.
4. add descriptive annotations that preserve context: tagging.
5. map tokens back to common form: lemmatization, stemming.

6. operate/model.

Reminder: From Texts to Numeric Data

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

“PREPROCESSING”

6. operate/model.

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).

→ 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).

→ 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).

→ 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).

→ 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).

→ 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).

→ 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 ∼ 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).

→ 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 ~ 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 preprocess 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 preprocess 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 preprocess 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 ?

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

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

if we preprocess 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.

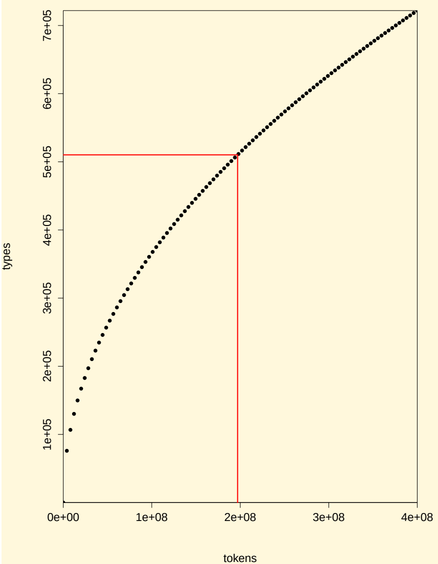
if we preprocess in different ways, we cause k to be different.

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

$k = 44, b = 0.49, T = 400,000$

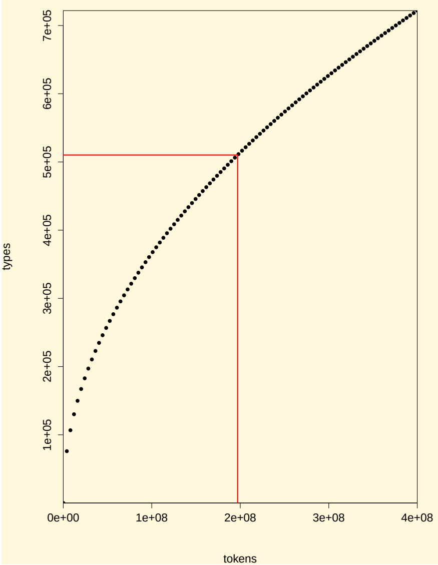
$k = 44, b = 0.49, T = 400,000$

RCV1.

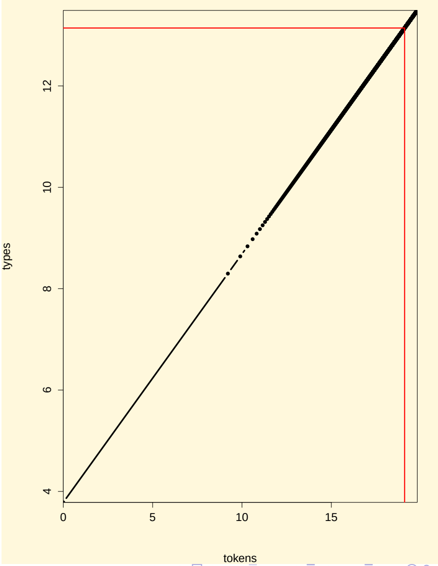


$k = 44, b = 0.49, T = 400,000$

RCV1.



RCV1, log-log.



Zipf's Law

Zipf's Law

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

Zipf's Law

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

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 types.

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?

Zipf's Law

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

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

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

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 types.

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 *i*th 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 types.

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 $\boxed{\propto \frac{1}{i}}$

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 types.

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 $\boxed{\propto \frac{1}{i}}$

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 types.

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 $\boxed{\propto \frac{1}{i}}$

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 types.

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 $\boxed{\propto \frac{1}{i}}$

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 types.

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 $\boxed{\propto \frac{1}{i}}$

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 types.

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 $\boxed{\propto \frac{1}{i}}$

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, $k = -1$.

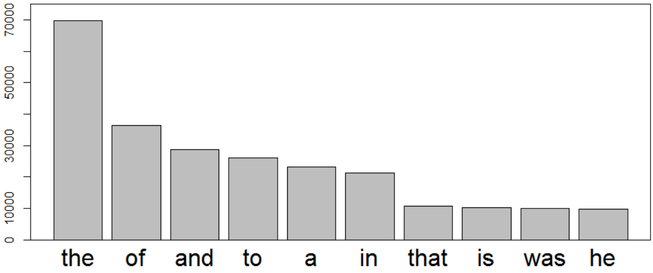
Brown Corpus (1961)

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

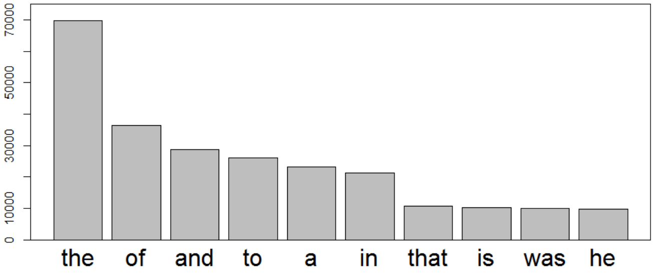
Brown Corpus (1961)

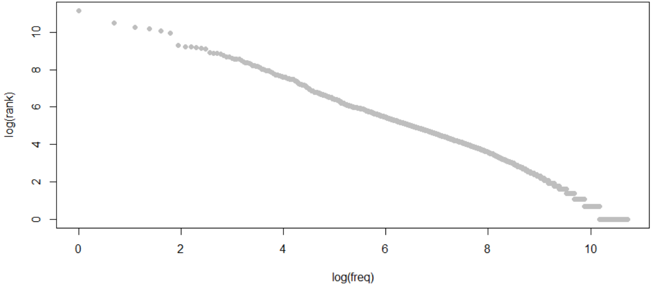
term	freq
the	69836
of	36365
and	28826
to	26126
a	23157
in	21314
that	10777
is	10182
was	9968
he	9801



Brown Corpus (1961)

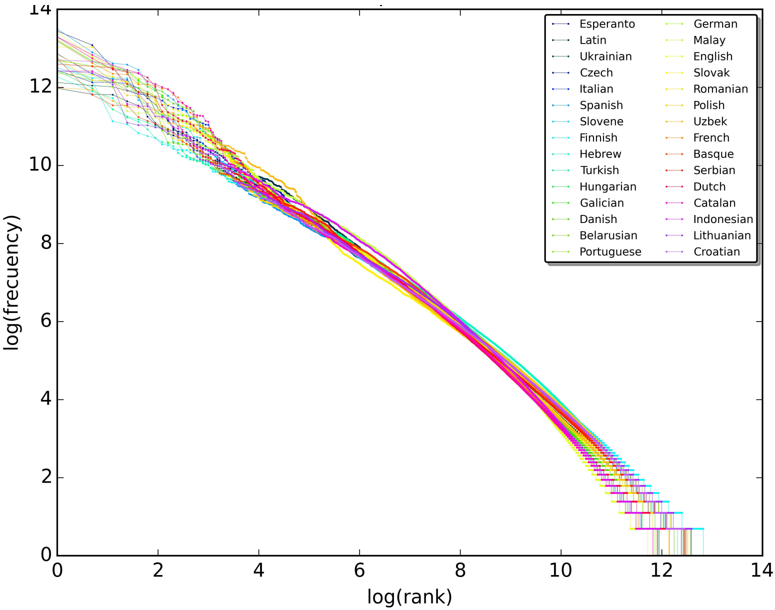
term	freq
the	69836
of	36365
and	28826
to	26126
a	23157
in	21314
that	10777
is	10182
was	9968
he	9801





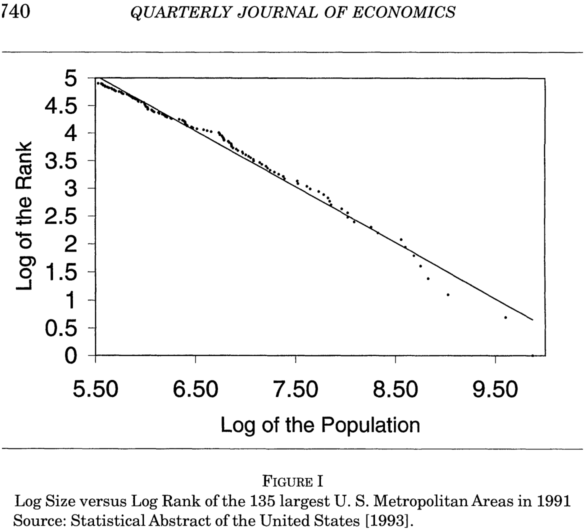
Other Languages (Wikipedia)

Other Languages (Wikipedia)



City Populations in US (Gabaix, 1999)

City Populations in US (Gabaix, 1999)



740 *QUARTERLY JOURNAL OF ECONOMICS*

FIGURE I

Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991

Source: Statistical Abstract of the United States [1993].

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)?

→ 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)?

→ 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)?

→ 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.

→ 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.

→ 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.

→ 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.

→ 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.

→ 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.

→ 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

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

→ 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

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

→ 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

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

→ 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

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

→ 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

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

→ 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

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

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

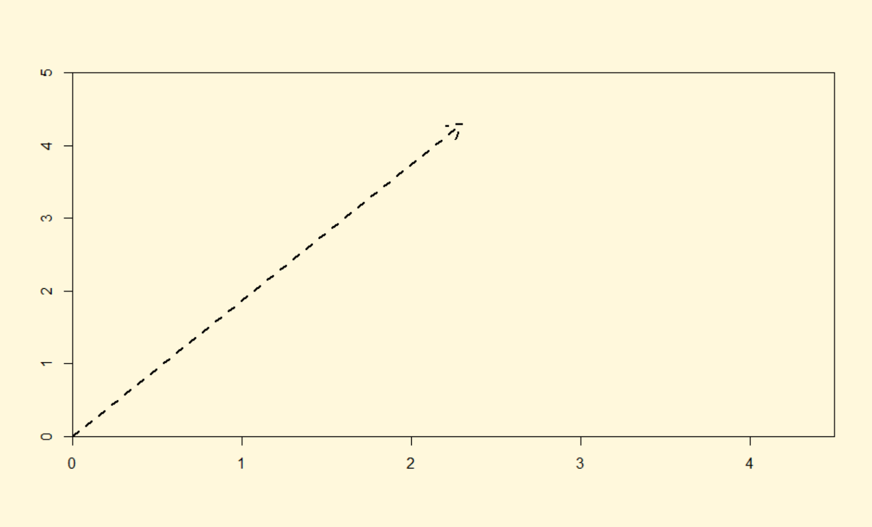
Graphically

Graphically

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

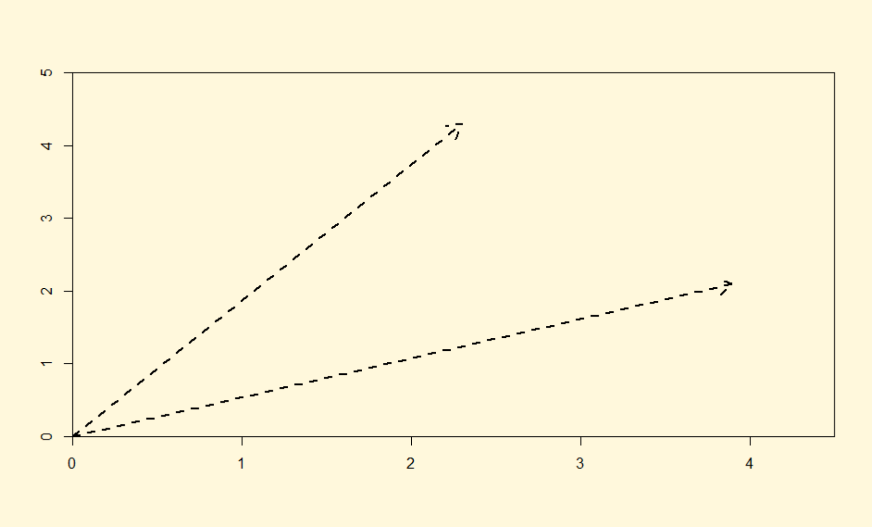
Graphically

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



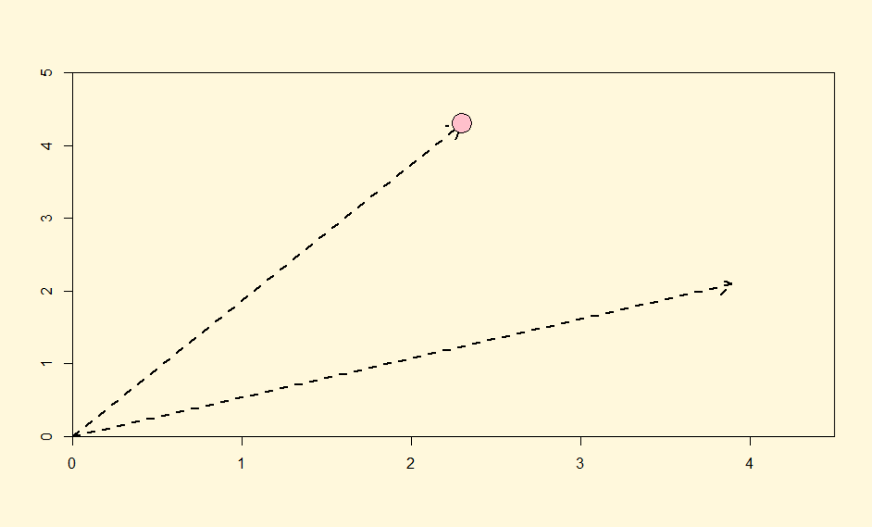
Graphically

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



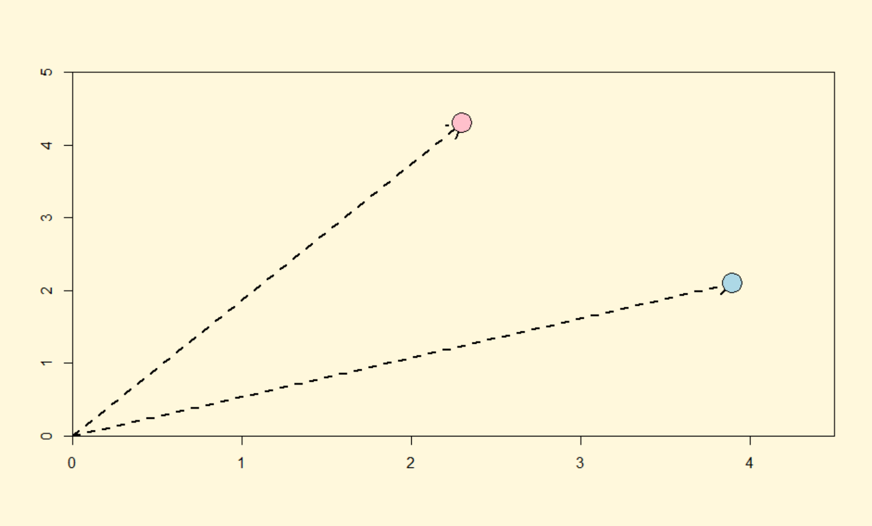
Graphically

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



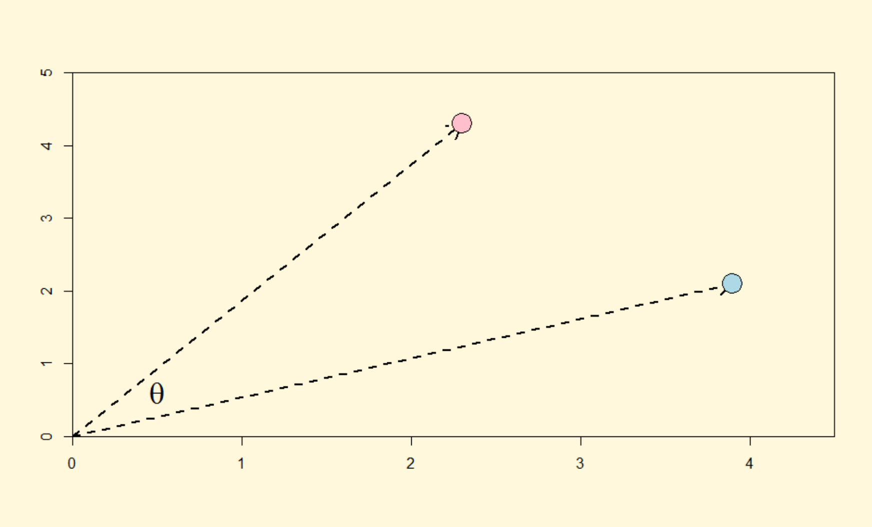
Graphically

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



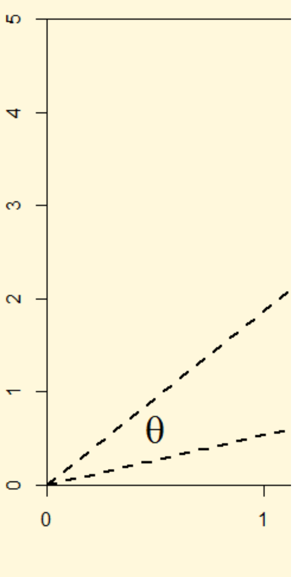
Graphically

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

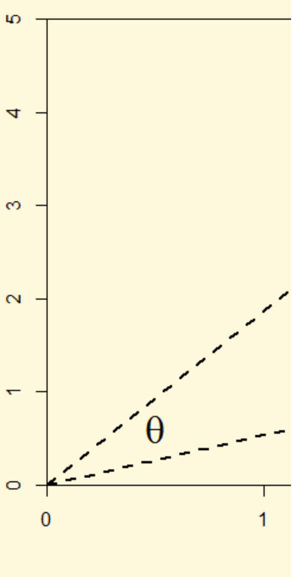


Algebra

Algebra

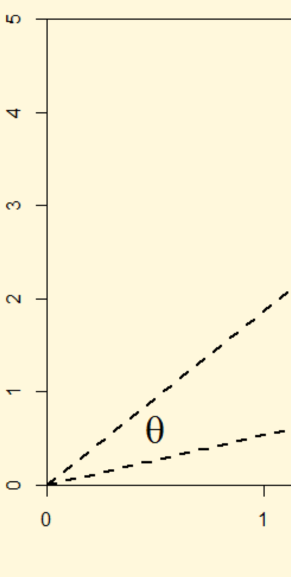


Algebra



know dot product of vectors:

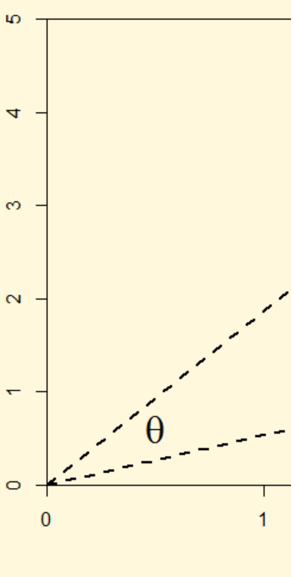
Algebra



know dot product of vectors:

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

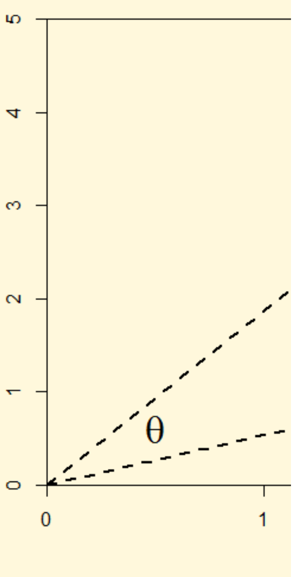
Algebra



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

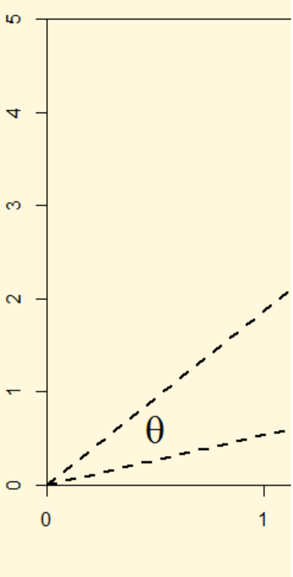


know dot product of vectors:
$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$

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:

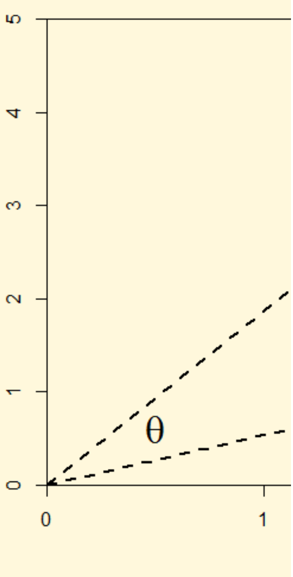
$$\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$$

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



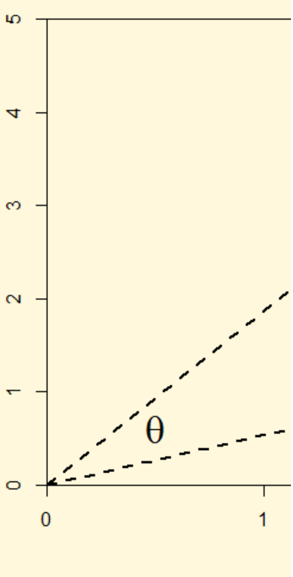
know dot product of vectors: $\mathbf{y}_i \cdot \mathbf{y}_j = ||\mathbf{y}_i||||\mathbf{y}_j|| \cos\theta$

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:

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

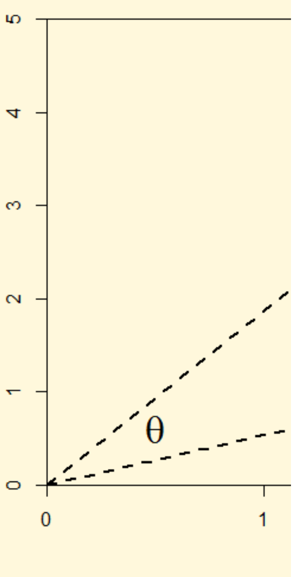
then $\cos\theta = \frac{\mathbf{y}_i \cdot \mathbf{y}_j}{||\mathbf{y}_i||||\mathbf{y}_j||}$

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

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

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

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

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

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.

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. $(\sum |\mathbf{y}_i - \mathbf{y}_j|^c)^{\frac{1}{c}}$.

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.

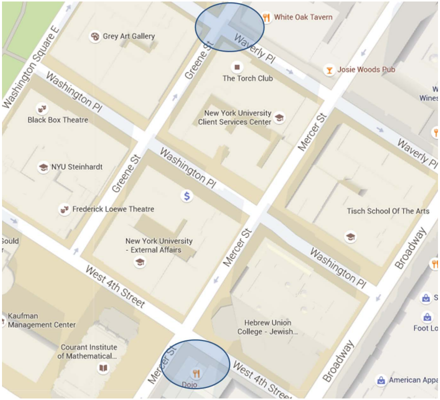
Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

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.

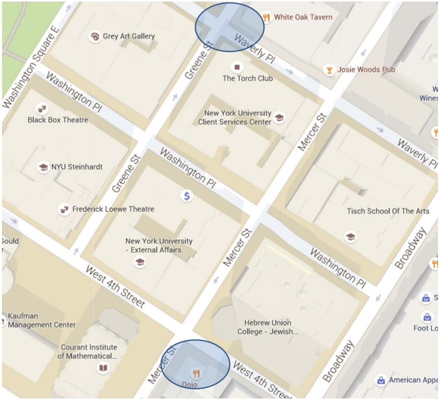
etc

Partner Exercise

Partner Exercise

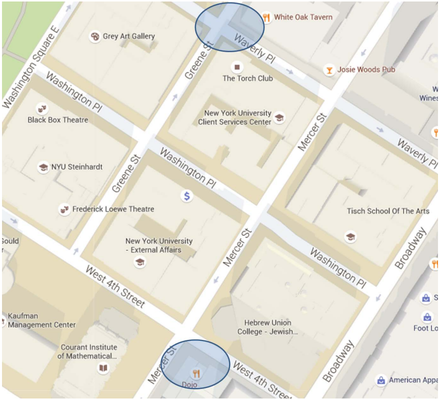


Partner Exercise



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

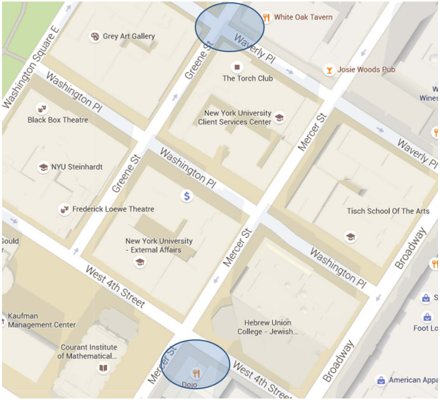
Partner Exercise



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

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

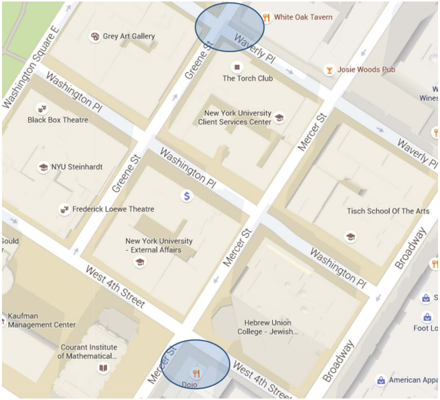
Partner Exercise



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

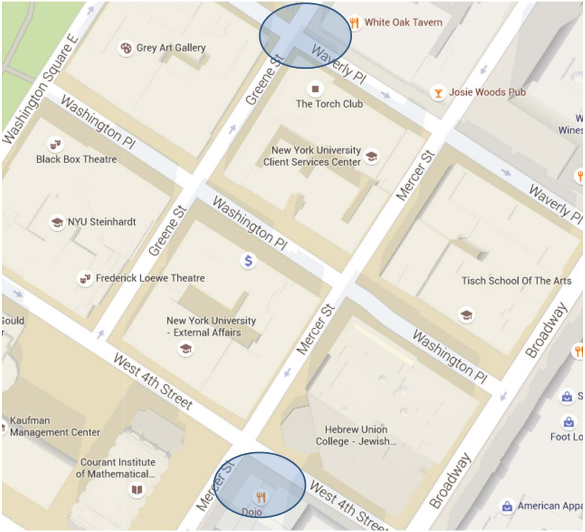


Suppose a block is one unit long and one unit wide.

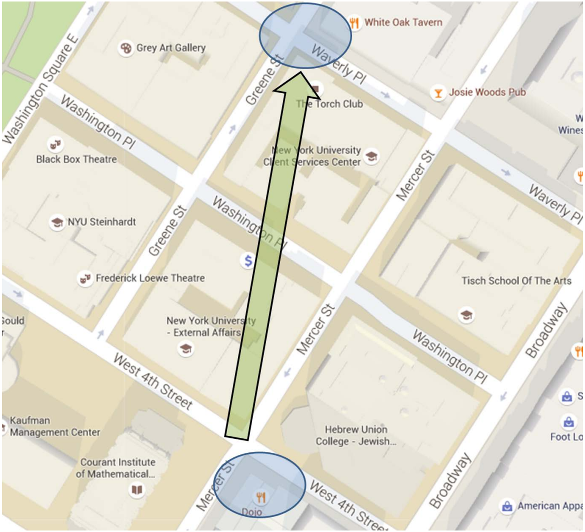
- what is Euclidean distance between Dojo and White Oak Tavern?
- what is Manhattan distance between Dojo and White Oak Tavern?

Solution

Solution

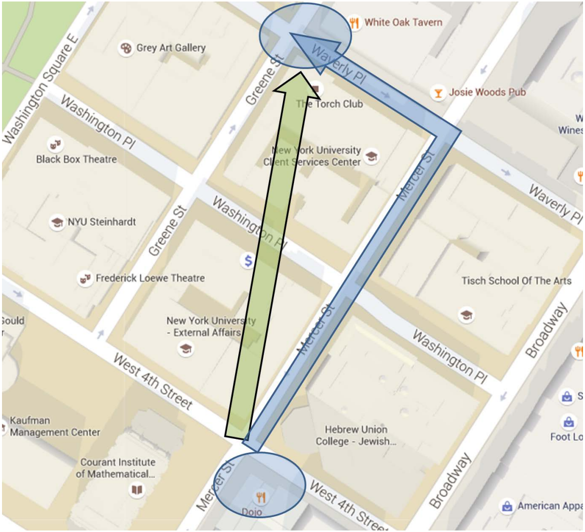

White Oak Tavern
Grey Art Gallery
Washington Square E
Waverly Pl
Greene St
The Torch Club
Josie Woods Pub
Washington Pl
Black Box Theatre
New York University Client Services Center
Wines
Mercer St
Waverly Pl
NYU Steinhardt
Greene St
Washington Pl
Frederick Loewe Theatre
Tisch School Of The Arts
ould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

Solution



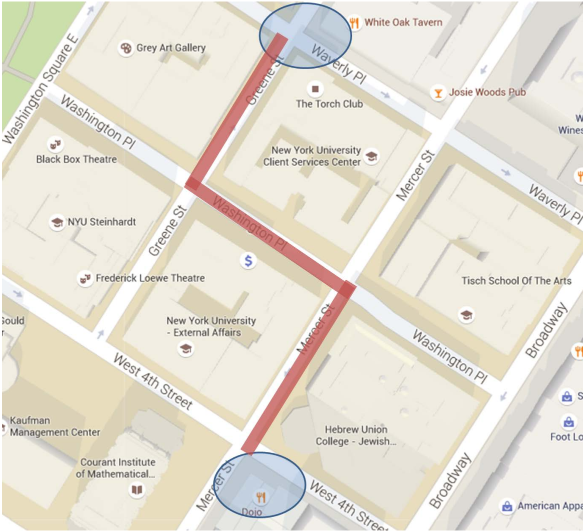
- Euclidean ($\sqrt{5}$)

Solution



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

Solution



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

Collocations, phrasemes and co-occurrence

Collocations, phrasemes and co-occurrence

So far,

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style.

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations:

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document,

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others.

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others. Mean something specific. Describe as strong collocation if the link between the words is fixed and restrictive.

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others. Mean something specific. Describe as strong collocation if the link between the words is fixed and restrictive.

e.g. '`do business`' (not '`make business`'),

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others. Mean something specific. Describe as strong collocation if the link between the words is fixed and restrictive.

e.g. '`do business`' (not '`make business`'), '`save money`' (not '`preserve money`').

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others. Mean something specific. Describe as strong collocation if the link between the words is fixed and restrictive.

e.g. '`do business`' (not '`make business`'), '`save money`' (not '`preserve money`'). Note that one of the terms is chosen freely (e.g. money),

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others. Mean something specific. Describe as strong collocation if the link between the words is fixed and restrictive.

e.g. '`do business`' (not '`make business`'), '`save money`' (not '`preserve money`'). Note that one of the terms is chosen freely (e.g. money), but the other is constrained by language.

Collocations, phrasemes and co-occurrence

So far, we've treated the features in a 'bag of words' style. In some cases, may also want to think seriously about collocations: words that co-occur in a document, adjacent (or close to adjacent) to one another.

defn group of two or more adjacent words that are seen together more often than we would 'expect' were words placed in document independent of the others. Mean something specific. Describe as strong collocation if the link between the words is fixed and restrictive.

e.g. '`do business`' (not '`make business`'), '`save money`' (not '`preserve money`'). Note that one of the terms is chosen freely (e.g. money), but the other is constrained by language.

Why study?

Why study?

Collocations are a type of phraseme: an idiomatic, ‘set phrase’ in a language.

Why study?

Collocations are a type of phraseme: an idiomatic, ‘set phrase’ in a language. Has *at least* one word or part that is constrained.

Why study?

Collocations are a type of phraseme: an idiomatic, ‘set phrase’ in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

Why study?

Collocations are a type of phraseme: an idiomatic, ‘set phrase’ in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

e.g. `‘He was pulling my leg’`

Why study?

Collocations are a type of phraseme: an idiomatic, 'set phrase' in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

e.g. `'He was pulling my leg'`, `'At the drop of a hat.'`

Why study?

Collocations are a type of phraseme: an idiomatic, ‘set phrase’ in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

e.g. ‘`He was pulling my leg`’, ‘`At the drop of a hat.`’

Phrasemes are interesting in that they tell us something about history or culture of language.

Why study?

Collocations are a type of phraseme: an idiomatic, 'set phrase' in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

e.g. '`He was pulling my leg`', '`At the drop of a hat.`'

Phrasemes are interesting in that they tell us something about history or culture of language.

Why do we talk of '`strong tea`' but '`powerful drugs`'?

Why study?

Collocations are a type of phraseme: an idiomatic, ‘set phrase’ in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

e.g. ‘`He was pulling my leg`’, ‘`At the drop of a hat.`’

Phrasemes are interesting in that they tell us something about history or culture of language.

Why do we talk of ‘`strong tea`’ but ‘`powerful drugs`’?

+ Very important when studying named entities, like people or cities

Why study?

Collocations are a type of phraseme: an idiomatic, 'set phrase' in a language. Has *at least* one word or part that is constrained. In the extreme, the meaning may not be implied by any of the lexical components.

e.g. '`He was pulling my leg`', '`At the drop of a hat.`'

Phrasemes are interesting in that they tell us something about history or culture of language.

Why do we talk of '`strong tea`' but '`powerful drugs`'?

+ Very important when studying named entities, like people or cities

e.g. '`Prime Minister`'

How to Find Them?

How to Find Them?

Frequency? Typically not enough to look for (say) common bigrams (NYT corpus):

How to Find Them?

Frequency? Typically not enough to look for (say) common bigrams (NYT corpus):

frequency	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

How to Find Them?

Frequency? Typically not enough to look for (say) common bigrams (NYT corpus):

frequency	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

→ most of these are uninteresting pairs of function words

How to Find Them?

Frequency? Typically not enough to look for (say) common bigrams (NYT corpus):

frequency	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

→ most of these are uninteresting pairs of function words (except ‘`New York`’)

How to Find Them?

Frequency? Typically not enough to look for (say) common bigrams (NYT corpus):

frequency	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

→ most of these are uninteresting pairs of function words (except ‘New York’)

Justeson and Katz (1995) improve performance considerably by applying parts-of-speech tagger,

How to Find Them?

Frequency? Typically not enough to look for (say) common bigrams (NYT corpus):

frequency	w_1	w_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

→ most of these are uninteresting pairs of function words (except ‘`New York`’)

Justeson and Katz (1995) improve performance considerably by applying parts-of-speech tagger, and only keeping those bigrams and trigrams that fulfill certain criteria. . .

Justeson and Katz (1995)

Justeson and Katz (1995)

N noun.

Justeson and Katz (1995)

N noun.

A adjective.

Justeson and Katz (1995)

N noun.

A adjective.

P preposition.

Justeson and Katz (1995)

N noun.

A adjective.

P preposition.

then following offers marked improvement:

Justeson and Katz (1995)

N noun.

A adjective.

P preposition.

then following offers marked improvement:

pattern	example
A N	Prime Minister
N N	surface area
A A N	little green men
A N N	real estate agent
N A N	home sweet home
N N N	term document matrix
N P N	Secretary of State

Reanalyzing NYT corpus: top ranked bi-grams

Reanalyzing NYT corpus: top ranked bi-grams

frequency	$w_1 w_2$	
11487	New York	A N
7261	United States	A N
5412	Los Angeles	N N
3301	last year	A N
3191	Saudi Arabia	N N
2699	last week	A N
2514	vice president	A N
2378	Persian Gulf	A N
2161	San Francisco	N N
2106	President Bush	N N
2001	Middle East	A N
1942	Saddam Hussein	N N
1867	Soviet Union	A N
1850	White House	A N
1633	United Nations	A N
1337	York City	N N
1328	oil prices	N N
1210	next year	A N
1074	chief executive	A N
1073	real estate	A N

Hypothesis Testing: Independence as Null Hypothesis

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together,

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together, consecutively,

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together, consecutively, if each word appears independently of all others.

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together, consecutively, if each word appears independently of all others.

→ if they appear independently,

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together, consecutively, if each word appears independently of all others.

→ if they appear independently, then the probability of a particular combination is just the product of the probability of seeing each one.

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together, consecutively, if each word appears independently of all others.

→ if they appear independently, then the probability of a particular combination is just the product of the probability of seeing each one.

so the null hypothesis is that $\Pr(w_1 w_2) = \Pr(w_1)\Pr(w_2)$ (this is naive, but useful)

Hypothesis Testing: Independence as Null Hypothesis

Ultimately, we are interested in how likely we are to see two words (w_1 and w_2) together, consecutively, if each word appears independently of all others.

→ if they appear independently, then the probability of a particular combination is just the product of the probability of seeing each one.

so the null hypothesis is that $\Pr(w_1 w_2) = \Pr(w_1)\Pr(w_2)$ (this is naive, but useful)

can construct 2 × 2 table, and consider expected vs observed frequency. . .

2×2 table

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

e.g. O_{11} is the number of times the collocation of interest, say ‘`New York`’,

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

e.g. O_{11} is the number of times the collocation of interest, say ‘`New York`’, occurred in the corpus.

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

e.g. O_{11} is the number of times the collocation of interest, say ‘`New York`’, occurred in the corpus.

but what is its expected frequency?

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

e.g. O_{11} is the number of times the collocation of interest, say ‘`New York`’, occurred in the corpus.

but what is its expected frequency? Under independence,

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

e.g. O_{11} is the number of times the collocation of interest, say ‘`New York`’, occurred in the corpus.

but what is its expected frequency? Under independence, it must be the proportion of all bigrams that start with ‘`New`’, multiplied by the proportion that end with ‘`York`’, multiplied by the number of all bigrams (N).

2×2 table

		Second Word		
		w_2	$\neg w_2$	total
First Word	w_1	O_{11}	O_{12}	$O_{11}+O_{12}$
	$\neg w_1$	O_{21}	O_{22}	$O_{21}+O_{22}$
	total	$O_{11}+O_{21}$	$O_{12}+O_{22}$	N

Here, O_{ij} is the observed frequency of that combination.

e.g. O_{11} is the number of times the collocation of interest, say ‘`New York`’, occurred in the corpus.

but what is its expected frequency? Under independence, it must be the proportion of all bigrams that start with ‘`New`’, multiplied by the proportion that end with ‘`York`’, multiplied by the number of all bigrams (N).

$$\rightarrow \frac{(O_{11}+O_{12})}{N} \times \frac{(O_{11}+O_{21})}{N} \times N = E_{11}$$

Brown Corpus: ‘New York’

Brown Corpus: ‘New York’

			Second Word	
		York	¬ York	total
First Word	New	303 New York	240 (e.g. ‘new day’)	543
	¬ New	6 (e.g. ‘from York’)	909219 (e.g. ‘red eye’)	909225
	total	309	909459	$N = 909768$

Brown Corpus: 'New York'

		Second Word		
		York	¬ York	total
First Word	New	303 New York	240 (e.g. 'new day')	543
	¬ New	6 (e.g. 'from York')	909219 (e.g. 'red eye')	909225
	total	309	909459	$N = 909768$

$$O_{11} = 303;\ E_{11} = \frac{(309)\times(543)}{909768} = 0.18$$

Brown Corpus: ‘New York’

		Second Word		
		York	¬ York	total
First Word	New	303 New York	240 (e.g. ‘new day’)	543
	¬ New	6 (e.g. ‘from York’)	909219 (e.g. ‘red eye’)	909225
	total	309	909459	$N = 909768$

$$O_{11} = 303;\ E_{11} = \frac{(309)\times(543)}{909768} = 0.18$$

hmm seems considerably more than we’d expect,

Brown Corpus: ‘New York’

First Word		Second Word: York	Second Word: ¬ York	total
	New	303 New York	240 (e.g. ‘new day’)	543
	¬ New	6 (e.g. ‘from York’)	909219 (e.g. ‘red eye’)	909225
	total	309	909459	$N = 909768$

$$O_{11} = 303;\ E_{11} = \frac{(309)\times(543)}{909768} = 0.18$$

hmm seems considerably more than we’d expect, by chance.

Brown Corpus: ‘New York’

			Second Word	
		York	¬ York	total
First Word	New	303 New York	240 (e.g. ‘new day’)	543
	¬ New	6 (e.g. ‘from York’)	909219 (e.g. ‘red eye’)	909225
	total	309	909459	$N = 909768$

$$O_{11} = 303;\ E_{11} = \frac{(309)\times(543)}{909768} = 0.18$$

hmm seems considerably more than we’d expect, by chance.

→ ‘york’ doesn’t occur often in the corpus,

Brown Corpus: ‘New York’

		Second Word		
		York	¬ York	total
First Word	New	303 New York	240 (e.g. ‘new day’)	543
	¬ New	6 (e.g. ‘from York’)	909219 (e.g. ‘red eye’)	909225
	total	309	909459	$N = 909768$

$$O_{11} = 303;\ E_{11} = \frac{(309)\times(543)}{909768} = 0.18$$

hmm seems considerably more than we’d expect, by chance.

→ ‘york’ doesn’t occur often in the corpus, but when it does, it’s almost always proceeded by ‘new’

Testing: χ^2

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is χ^2 distributed,

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ is χ^2 distributed, where i is the rows, j is the columns.

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is χ^2 distributed, where i is the rows, j is the columns.

and degrees of freedom is (number of rows minus 1) $\times$ (number of columns minus 1).

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is χ^2 distributed, where i is the rows, j is the columns.

and degrees of freedom is (number of rows minus 1) × (number of columns minus 1).

so for '`New York`',

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is χ^2 distributed, where i is the rows, j is the columns.

and degrees of freedom is (number of rows minus 1) × (number of columns minus 1).

so for '`New York`', $X^2 = 496020$ on 1 degree of freedom,

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is χ^2 distributed, where i is the rows, j is the columns.

and degrees of freedom is (number of rows minus 1) × (number of columns minus 1).

so for 'New York', $X^2 = 496020$ on 1 degree of freedom, $\rightarrow p < 0.001$

Testing: χ^2

The set up of the problem allows for a χ^2 approach.

i.e. $X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ is χ^2 distributed, where i is the rows, j is the columns.

and degrees of freedom is (number of rows minus 1) × (number of columns minus 1).

so for '`New York`', $X^2 = 496020$ on 1 degree of freedom, $\rightarrow p < 0.001$

⇒ reject the null hypothesis of independence: this word is a good choice as a collocation.

Likelihood Ratio

Likelihood Ratio

In general,

Likelihood Ratio

In general, the contingency tables are highly skewed

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large,

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for '`New York`')

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for `New York`)

but in such cases X^2 statistic, does not approximate the χ^2 distribution (on 1 degree of freedom) as well as G^2 (Dunning, 1993)

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for '`New York`')

but in such cases X^2 statistic, does not approximate the χ^2 distribution (on 1 degree of freedom) as well as G^2 (Dunning, 1993)

where

$$\boxed{G^2 = G = \;\; 2\sum_i \sum_j O_{ij} \ln \frac{O_{ij}}{E_{ij}} = \;\; 2\sum_{ij} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}}\right)}$$

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for '`New York`')

but in such cases X^2 statistic, does not approximate the χ^2 distribution (on 1 degree of freedom) as well as G^2 (Dunning, 1993)

where

$$\boxed{G^2 = G = \ 2\sum_i \sum_j O_{ij} \ln \frac{O_{ij}}{E_{ij}} = \ 2\sum_{ij} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}}\right)}$$

This formulation is equivalent to a log-likelihood ratio for a contingency table. . .

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for 'New York')

but in such cases X^2 statistic, does not approximate the χ^2 distribution (on 1 degree of freedom) as well as G^2 (Dunning, 1993)

where

$$\boxed{G^2 = G = \ 2\sum_i \sum_j O_{ij} \ln \frac{O_{ij}}{E_{ij}} = \ 2\sum_{ij} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}} \right)}$$

This formulation is equivalent to a log-likelihood ratio for a contingency table...

where the numerator is the maximum likelihood of the data under the null of independence (i.e. the ML consistent with H_0),

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for ‘`New York`’)

but in such cases X^2 statistic, does not approximate the χ^2 distribution (on 1 degree of freedom) as well as G^2 (Dunning, 1993)

where

$$\boxed{G^2 = G = \ 2\sum_i \sum_j O_{ij} \ln \frac{O_{ij}}{E_{ij}} = \ 2\sum_{ij} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}} \right)}$$

This formulation is equivalent to a log-likelihood ratio for a contingency table. . .

where the numerator is the maximum likelihood of the data under the null of independence (i.e. the ML consistent with H_0),

and the denominator is the unconstrained maximum likelihood,

Likelihood Ratio

In general, the contingency tables are highly skewed

i.e. N is large, O_{11} is small (as for ‘`New York`’)

but in such cases X^2 statistic, does not approximate the χ^2 distribution (on 1 degree of freedom) as well as G^2 (Dunning, 1993)

where

$$\boxed{G^2 = G = \ 2\sum_i \sum_j O_{ij} \ln \frac{O_{ij}}{E_{ij}} = \ 2\sum_{ij} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}}\right)}$$

This formulation is equivalent to a log-likelihood ratio for a contingency table. . .

where the numerator is the maximum likelihood of the data under the null of independence (i.e. the ML consistent with H_0),

and the denominator is the unconstrained maximum likelihood, where $\Pr(w_2|w_1)$ does not (have to) equal $\Pr(w_2|\neg w_1)$.

Other Ideas

Other Ideas

can perform *t*-test, in which

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$;

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

but assumes that probabilities are approximately normally distributed.

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

but assumes that probabilities are approximately normally distributed.

Pointwise Mutual Information:

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

but assumes that probabilities are approximately normally distributed.

Pointwise Mutual Information: $I(w_1 w_2) = \frac{\Pr(w_1 w_2)}{\Pr(w_1)\Pr(w_2)}$.

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

but assumes that probabilities are approximately normally distributed.

Pointwise Mutual Information: $I(w_1 w_2) = \frac{\Pr(w_1 w_2)}{\Pr(w_1)\Pr(w_2)}$.

but very rare bigrams tend to receive very high scores.

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

but assumes that probabilities are approximately normally distributed.

Pointwise Mutual Information: $I(w_1 w_2) = \frac{\Pr(w_1 w_2)}{\Pr(w_1)\Pr(w_2)}$.

but very rare bigrams tend to receive very high scores.

G^2 has been extended to trigrams.

Other Ideas

can perform *t*-test, in which

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where $\bar{x}$ is $\Pr(w_1 w_2) = \frac{O_{11}}{N}$; μ is $\Pr(w_1)\Pr(w_2)$ and $s^2 = p(1-p) \approx p = \bar{x}$.

but assumes that probabilities are approximately normally distributed.

Pointwise Mutual Information: $I(w_1 w_2) = \frac{\Pr(w_1 w_2)}{\Pr(w_1)\Pr(w_2)}$.

but very rare bigrams tend to receive very high scores.

G^2 has been extended to trigrams.

Partner Exercise

Partner Exercise

1. Ignoring parts of speech information, almost all bigrams in a corpus occur more often than chance would lead us to expect.

Partner Exercise

1 Ignoring parts of speech information, almost all bigrams in a corpus occur more often than chance would lead us to expect. Why?

Partner Exercise

1 Ignoring parts of speech information, almost all bigrams in a corpus occur more often than chance would lead us to expect. Why? Give an example, and explain why this matters when looking for collocations.

Partner Exercise

1. Ignoring parts of speech information, almost all bigrams in a corpus occur more often than chance would lead us to expect. Why? Give an example, and explain why this matters when looking for collocations.

2. How would you implement the Justeson & Katz method in practice?

Partner Exercise

1 Ignoring parts of speech information, almost all bigrams in a corpus occur more often than chance would lead us to expect. Why? Give an example, and explain why this matters when looking for collocations.

2 How would you implement the Justeson & Katz method in practice?

3 The *G*-test is a type of likelihood ratio test. Assuming we are working without logs, what are the bounds on the calculated ratio statistic? Why?

Partner Exercise

1. Ignoring parts of speech information, almost all bigrams in a corpus occur more often than chance would lead us to expect. Why? Give an example, and explain why this matters when looking for collocations.

2. How would you implement the Justeson & Katz method in practice?

3. The G-test is a type of likelihood ratio test. Assuming we are working without logs, what are the bounds on the calculated ratio statistic? Why? (hint: remember that the null model is in the numerator)

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.

→ 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'

→ 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.

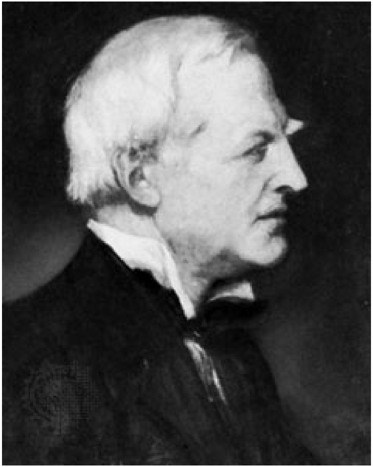
Detail: s267738.txt

Detail: s267738.txt

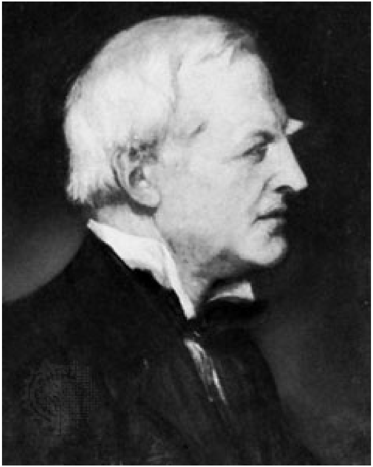
preword	word	postword
swept away in some further	democratic	change. And it is for
throne. When you get a	democratic	basis for your institutions, you
differences between ourselves and other	democratic	legislatures? Where is the democratic
democratic legislatures? Where is the	democratic	legislature which 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

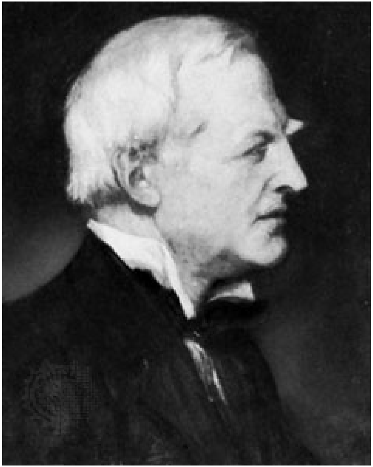


The Original Speaker and Speech



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?

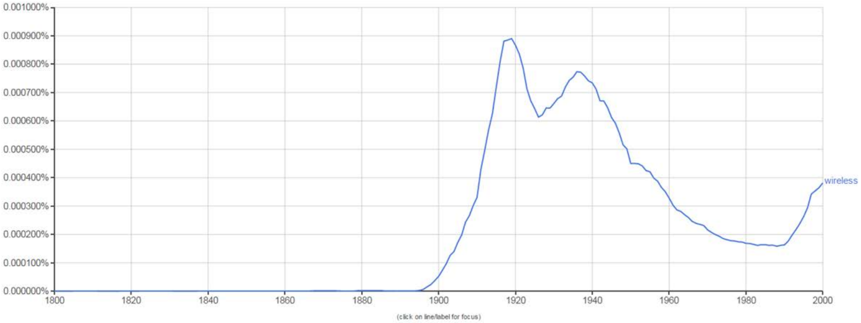
Partner Exercise

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’


wireless
(click on line/label for focus)