2. Descriptive Inference I (Flipped)

DS-GA 1015, Text as Data
Arthur Spirling

Feb 16, 2021

Housekeeping

Housekeeping

1 Lab straight after this lecture

Housekeeping

1. Lab straight after this lecture
2. Please do federal engagement survey

Housekeeping

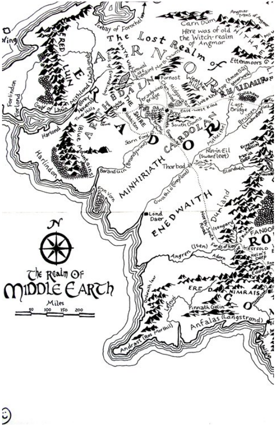
1. Lab straight after this lecture
2. Please do federal engagement survey
3. HW 1 out soon.

Where Are We?

Where Are We?

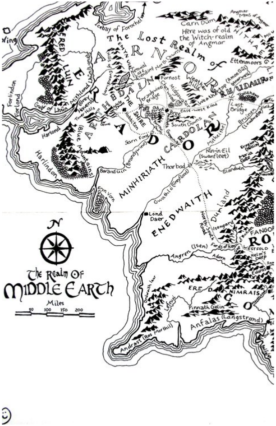
The Realm of
Middle Earth
Miles
50
100
150
200
Icebay of Forochel
The Lost Realm of
ARNOR
Carn Dûm
Here was of old
the Witch-realm
of Angmar
Ettenmoors
Fornost
North Downs
Weather Hills
Weathertop (Amon Sûl)
RHUDAUR
Trollshaws
Last Bridge
Bruinen (Loudwater)
Ered Luin
Forlindon
Forlond
Gulf of Lhûn
Mithlond (Grey Havens)
Tower Hills
Harlond
Harlindon
Blue Mountains
ERIADOR
Emyn Uial
Brandywine Bridge
East-West Road
Bree
South Downs
Hobbiton
THE SHIRE
Sarn Ford
CARDOLAN
Mitheithel
Nîn-in-Eil (Swanfleet)
Moria Gate
Glanduin
Tharbad
Baranduin (Brandywine)
MINHIRIATH
Eryn Vorn
Gwathló (Greyflood)
Lond Daer
ENEDWAITH
Dunland
North-South Road
Isengard
FANGORN
Fords of Isen
Angren (Isen)
Adorn
WESTFOLD
Helm's Deep
Edoras
ERED NIMRAIS
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).

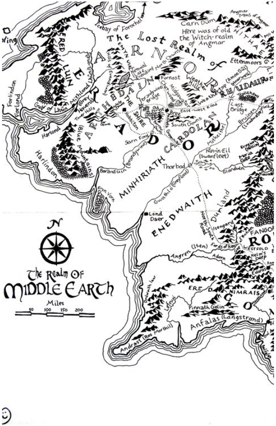
Our fundamental unit of text analysis is the document term matrix.

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.

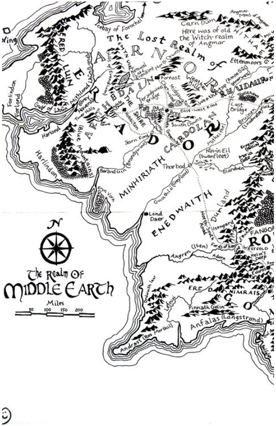
This is a set of stacked vectors, with each entry in each vector representing the ‘amount’ of a particular term.

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.

Relationships in Data

The vector space model generally results in a lossy compression.

Relationships in Data

The vector space model generally results in a lossy compression. What does this mean? Does it matter?

Relationships in Data

The vector space model generally results in a lossy compression. What does this mean? Does it matter?

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

Relationships in Data

The vector space model generally results in a lossy compression. What does this mean? Does it matter?

Heap's Law tells us the relationship between number of tokens and number of types. What does it look like? Why?

Relationships in Data

The vector space model generally results in a lossy compression. What does this mean? Does it matter?

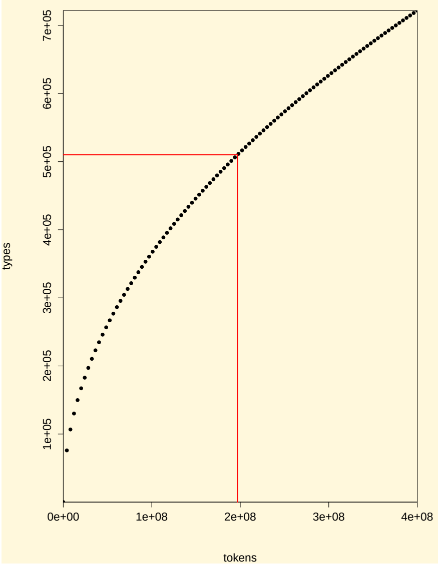
Heap's Law tells us the relationship between number of tokens and number of types. What does it look like? Why?

Zipf's Law tells us about the rank-frequency distribution. What does it say?

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

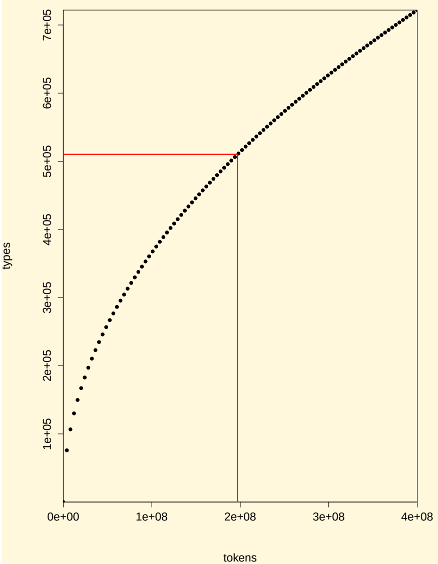
$k = 44, b = 0.49, T = 400,000$

RCV1.

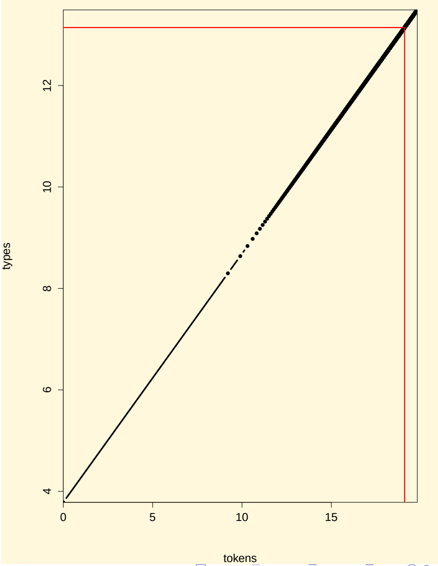


$k = 44, b = 0.49, T = 400,000$

RCV1.



RCV1, log-log.



Zipf's Law

Zipf's Law

corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

Zipf's Law

corpus frequency of ith most common term is $\boxed{\propto \frac{1}{i}}$

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

Zipf's Law

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

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

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

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

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

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,

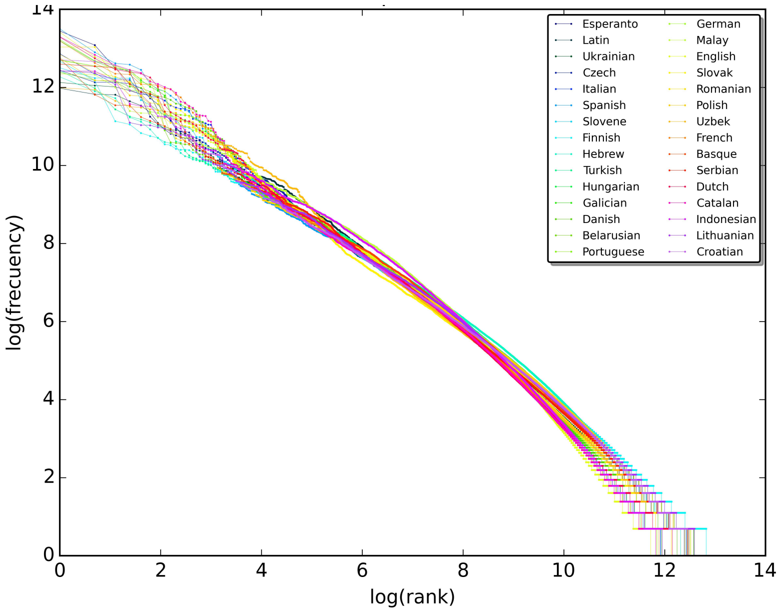
and third most common term is one third as common as most common,

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, $k = -1$.

Other Languages (Wikipedia)

Other Languages (Wikipedia)



Questions

Questions

1. Many (all?) languages follow Zipf's law: slope of log(rank) to log(frequency) is -1.

Questions

1 Many (all?) languages follow Zipf's law: slope of log(rank) to log(frequency) is -1. Babies babbling words does not follow it. Why? What does that imply?

Questions

1 Many (all?) languages follow Zipf's law: slope of log(rank) to log(frequency) is -1. Babies babbling words does not follow it. Why? What does that imply?

2 Astrophysicists study extraterrestrial signals looking for Zipf's law in the patterns. Why?

NAUTILUS | cosmos

Listening for Extraterrestrial Blah Blah

At the cosmic dinner party, intelligence is the loudest thing in the room.

By Laurance R. Doyle
Illustrations by Tianhua Mao

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?

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.

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

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.

$\boxed{c_{ij} \approx 0.90}$ Why are the numbers in general so ‘high’?

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.

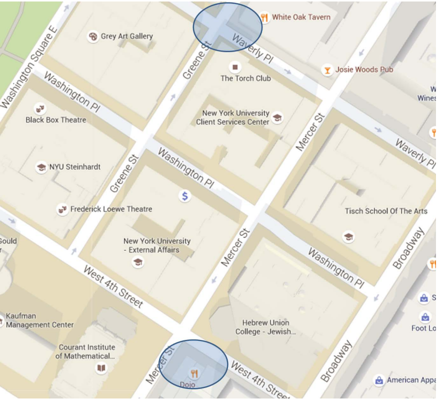
Canberra: weighted version of Manhattan distance. $\sum \frac{|\mathbf{y}_i - \mathbf{y}_j|}{|\mathbf{y}_i| + |\mathbf{y}_j|}$

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.

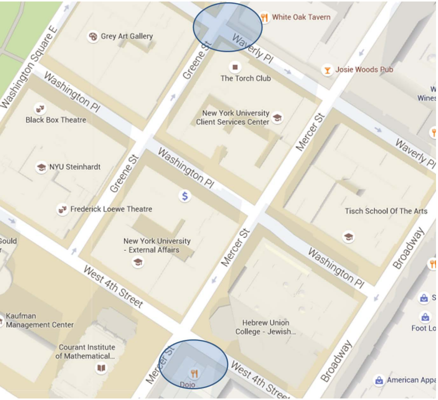
etc

Partner Exercise

Partner Exercise

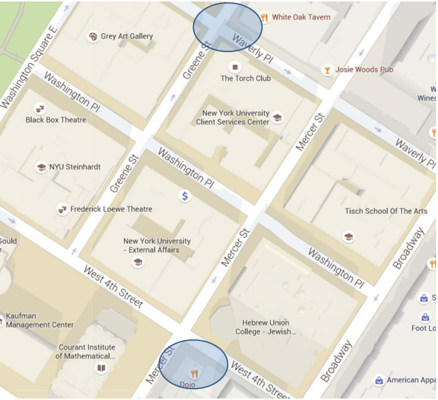


Partner Exercise



Look at the map. Suppose a block is one unit long and one unit wide.

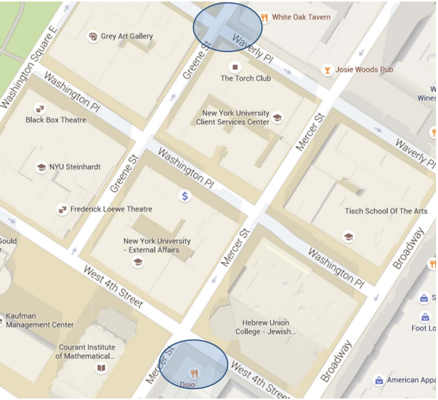
Partner Exercise



Look at the map. Suppose a block is one unit long and one unit wide.

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

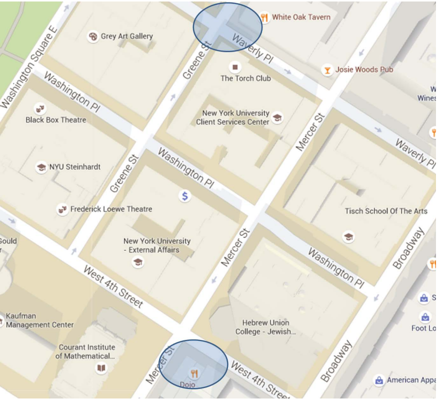
Partner Exercise



Look at the map. 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

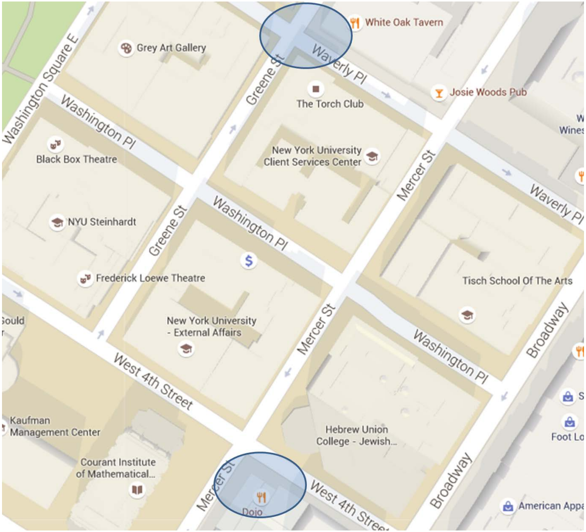


Look at the map. Suppose a block is one unit long and one unit wide.

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

Solution

Solution

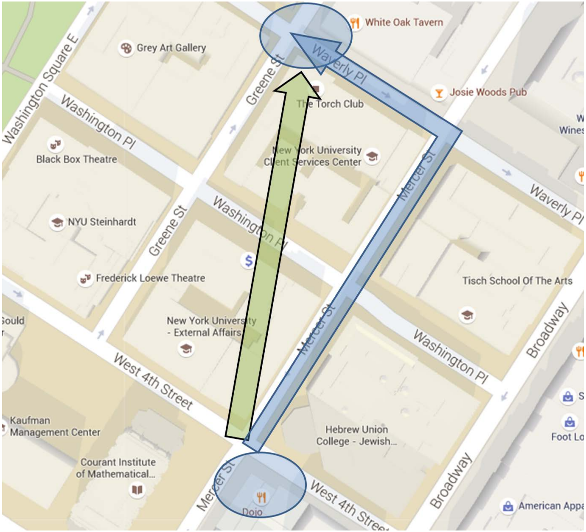


White Oak Tavern
Grey Art Gallery
Waverly Pl
Greene St
The Torch Club
Josie Woods Pub
Washington Square E
Washington Pl
Black Box Theatre
New York University
Client Services Center
Mercer St
Waverly Pl
NYU Steinhardt
Greene St
Washington Pl
Frederick Loewe Theatre
Tisch School Of The Arts
New York University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

Solution

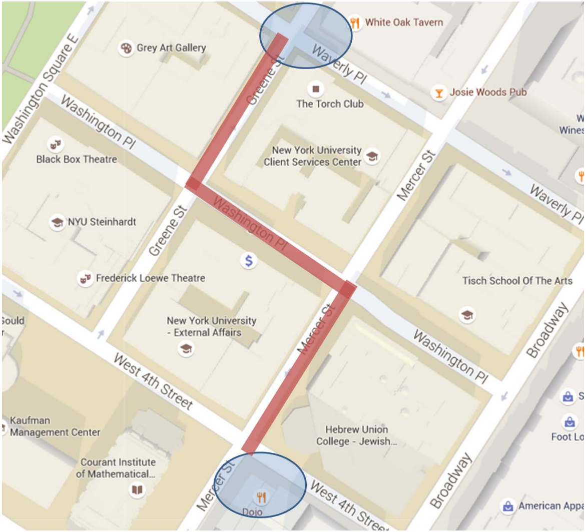
- Euclidean ($\sqrt{5}$)

Solution



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

Solution



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

Edit Distance: An Example

Edit Distance: An Example

original, s_1	`one million dollar limit for licensees in easternmost Florida`

Edit Distance: An Example

original, s_1	`one million dollar limit for licensees in easternmost Florida`
final, s_2	`one billion dollar limit for licensees in southern Florida`

Edit Distance: An Example

original, s_1	`one million dollar limit for licensees in easternmost Florida`
final, s_2	`one billion dollar limit for licensees in southern Florida`

Suppose we can do three things:

1. insert a character into a string

Edit Distance: An Example

original, s_1	`one million dollar limit for licensees in easternmost Florida`
final, s_2	`one billion dollar limit for licensees in southern Florida`

Suppose we can do three things:

1 insert a character into a string

2 delete a character from a string

Edit Distance: An Example

original, s_1	`one million dollar limit for licensees in easternmost Florida`
final, s_2	`one billion dollar limit for licensees in southern Florida`

Suppose we can do three things:

1. insert a character into a string
2. delete a character from a string
3. replace a character in a string by another character

Edit Distance: An Example

original, s_1	`one million dollar limit for licensees in easternmost Florida`
final, s_2	`one billion dollar limit for licensees in southern Florida`

Suppose we can do three things:

1. insert a character into a string
2. delete a character from a string
3. replace a character in a string by another character

The smallest number of operations taking us from s_1 to s_2 is the **Levenshtein distance** between those strings.

Levenshtein in Action

Levenshtein in Action

$s_1 = \texttt{easternmost}$

Levenshtein in Action

$s_1 = $ `easternmost`

$s_2 = $ `southern`

Levenshtein in Action

$s_1 = \texttt{easternmost}$

$s_2 = \texttt{southern}$

1 delete `m`, delete `o`, delete `s`, delete `t`.

Levenshtein in Action

$s_1 =$ easternmost

$s_2 =$ southern

1 delete m, delete o, delete s, delete t. $\longrightarrow$ eastern

Levenshtein in Action

$s_1 =$ `easternmost`

$s_2 =$ `southern`

1. delete `m`, delete `o`, delete `s`, delete `t`. $\longrightarrow$ `eastern`
2. insert `h`. $\longrightarrow$ `easthern`

Levenshtein in Action

$s_1 = \texttt{easternmost}$

$s_2 = \texttt{southern}$

1 delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`

2 insert `h`. $\rightarrow$ `easthern`

3 replace `e`, `a` and `s` with `s`, `o` and `u`.

Levenshtein in Action

$s_1 = \texttt{easternmost}$

$s_2 = \texttt{southern}$

1. delete m, delete o, delete s, delete t. $\rightarrow$ eastern
2. insert h. $\rightarrow$ easthern
3. replace e, a and s with s, o and u. $\rightarrow$ southern.

Levenshtein in Action

$s_1 = \texttt{easternmost}$

$s_2 = \texttt{southern}$

1. delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`
2. insert `h`. $\rightarrow$ `easthern`
3. replace `e`, `a` and `s` with `s`, `o` and `u`. $\rightarrow$ `southern`.

How many operations?

Levenshtein in Action

$s_1 =$ `easternmost`

$s_2 =$ `southern`

1. delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`
2. insert `h`. $\rightarrow$ `easthern`
3. replace `e`, `a` and `s` with `s`, `o` and `u`. $\rightarrow$ `southern`.

How many operations? 4

Levenshtein in Action

$s_1 = \texttt{easternmost}$

$s_2 = \texttt{southern}$

1. delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`
2. insert `h`. $\rightarrow$ `easthern`
3. replace `e`, `a` and `s` with `s`, `o` and `u`. $\rightarrow$ `southern`.

How many operations? $4 + 1$

Levenshtein in Action

$s_1 =$ `easternmost`

$s_2 =$ `southern`

1. delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`
2. insert `h`. $\rightarrow$ `easthern`
3. replace `e`, `a` and `s` with `s`, `o` and `u`. $\rightarrow$ `southern`.

How many operations? $4 + 1 + 3$

Levenshtein in Action

$s_1 =$ `easternmost`

$s_2 =$ `southern`

1. delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`
2. insert `h`. $\rightarrow$ `easthern`
3. replace `e`, `a` and `s` with `s`, `o` and `u`. $\rightarrow$ `southern`.

How many operations? $4 + 1 + 3 = 8$.

Levenshtein in Action

$s_1 = \texttt{easternmost}$

$s_2 = \texttt{southern}$

1. delete `m`, delete `o`, delete `s`, delete `t`. $\rightarrow$ `eastern`
2. insert `h`. $\rightarrow$ `easthern`
3. replace `e`, `a` and `s` with `s`, `o` and `u`. $\rightarrow$ `southern`.

How many operations? $4 + 1 + 3 = 8$. **Levenshtein distance** is 8.

Levenshtein Example

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

We want to suggest a correct spelling. Candidates might be

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

We want to suggest a correct spelling. Candidates might be

$c_1 =$ `hypocrisy`

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

We want to suggest a correct spelling. Candidates might be

$c_1 =$ `hypocrisy`

$c_2 =$ `hypocrite`

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ hipocrit

We want to suggest a correct spelling. Candidates might be

$c_1 =$ hypocrisy

$c_2 =$ hypocrite

Which is closest by Levenshtein distance?

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

We want to suggest a correct spelling. Candidates might be

$c_1 =$ `hypocrisy`

$c_2 =$ `hypocrite`

Which is closest by Levenshtein distance?

$s \to c_1$: replace `i`, replace `t`, add `y`.

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s = \texttt{hipocrit}$

We want to suggest a correct spelling. Candidates might be

$c_1 = \texttt{hypocrisy}$

$c_2 = \texttt{hypocrite}$

Which is closest by Levenshtein distance?

$s \to c_1$: replace `i`, replace `t`, add `y`. So, 3.

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

We want to suggest a correct spelling. Candidates might be

$c_1 =$ `hypocrisy`

$c_2 =$ `hypocrite`

Which is closest by Levenshtein distance?

$s \rightarrow c_1$: replace `i`, replace `t`, add `y`. So, 3.

$s \rightarrow c_2$: replace `i`, add `e`.

Levenshtein Example

We observe a misspelling, perhaps in an online forum or some other casual situation:

$s =$ `hipocrit`

We want to suggest a correct spelling. Candidates might be

$c_1 =$ `hypocrisy`

$c_2 =$ `hypocrite`

Which is closest by Levenshtein distance?

$s \rightarrow c_1$: replace `i`, replace `t`, add `y`. So, 3.

$s \rightarrow c_2$: replace `i`, add `e`. So, 2.

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’

		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.

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) $\times$ (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) $\times$ (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) $\times$ (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) $\times$ (number of columns minus 1).

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

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

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

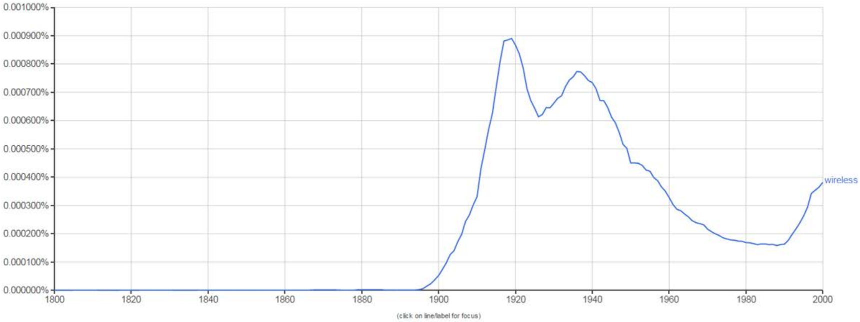
In Information Retrieval it is often extremely helpful to know how and where a particular token of interest appears, in terms of the words around it.

→ 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.

Use of 'Wireless' changes by context



Exercise

Exercise

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

Exercise

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

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