Distance measures in stylometry

Joanna Byszuk, Maciej Eder, Jan Rybicki

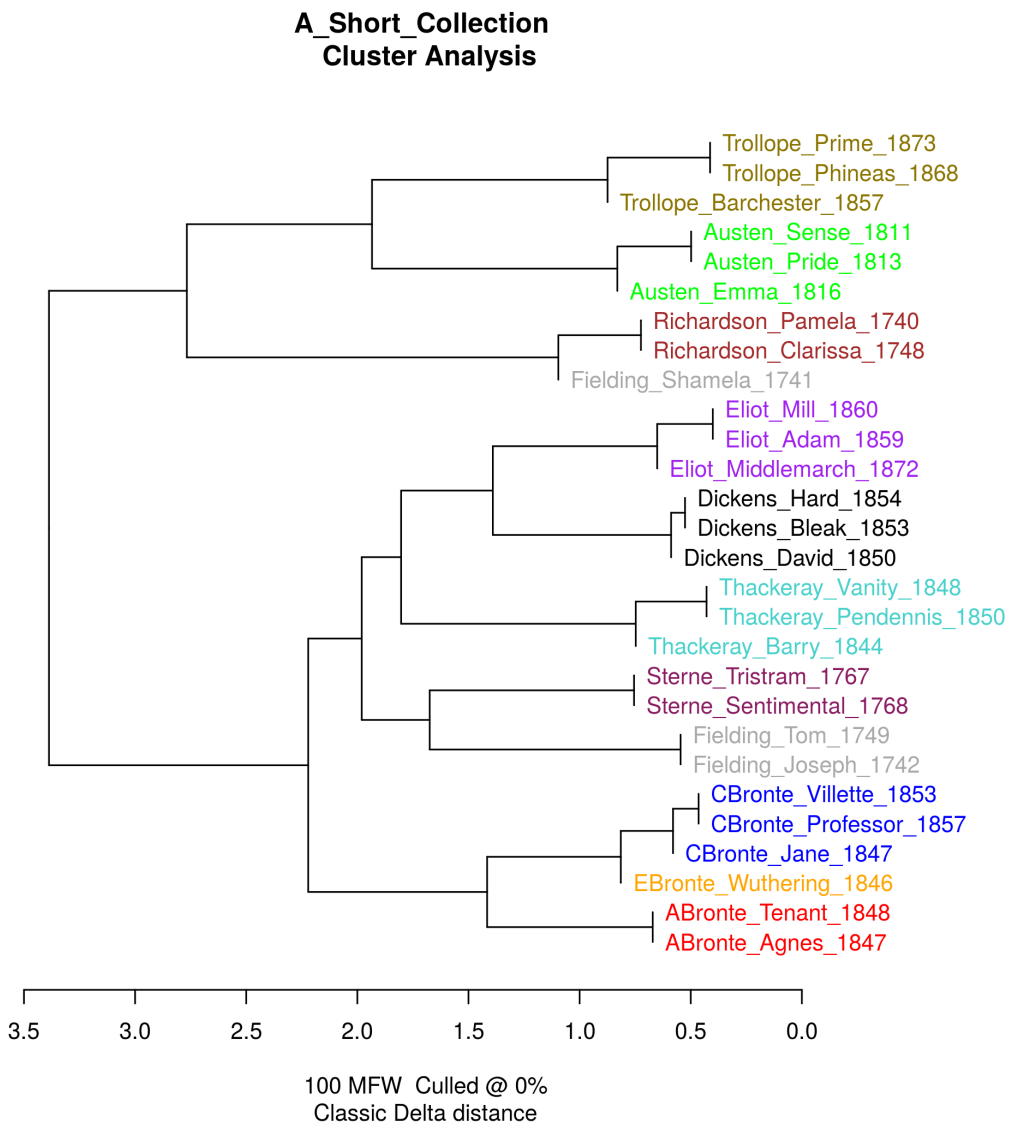
14.06.2018

Disclaimer

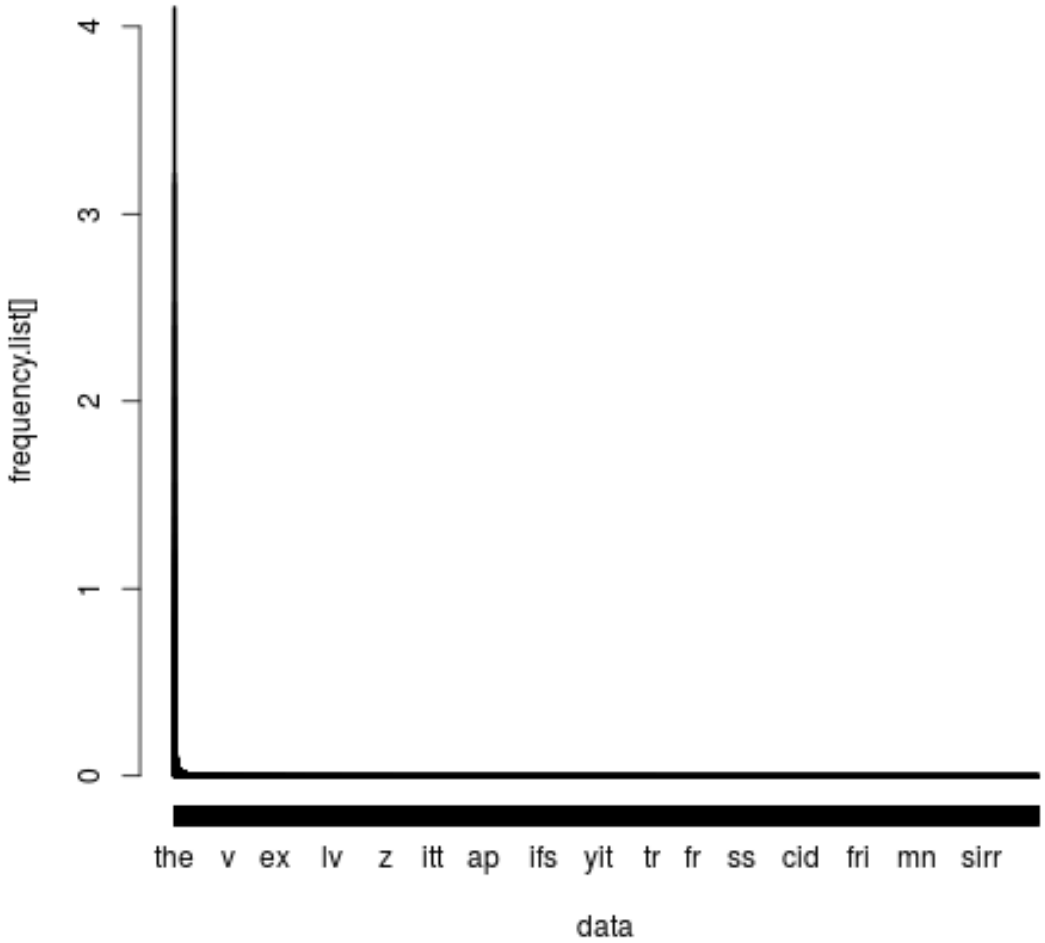
The presentation was adapted by Joanna Byszuk for the 'Stylometry with R' course at DHSI 2018 (co-taught by JB and JR) from the previous presentations on multidimensionality and distance measures created by Maciej Eder. Political system metaphors are of his invention.

All calculations/word lists were based on *A Short Collection of British Fiction* corpus as available via Computational Stylistics Group

Most frequent words - important?



Most frequent words - distribution



Zipf's law

Zipf's law states that given some corpus of natural language utterances, **the frequency of any word is inversely proportional to its rank in the frequency table**.

Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.: the rank-frequency distribution is an inverse relation.

Zipf's law - rank/length dependence

- First ten:
  - the, and, to, of, i, a, in, that, he, it
- 100-110:
  - made, miss, too, sir, shall, come, might, thought, himself, dear, make
- 10000-10010:
  - abel, accommodations, acquainting, acre, addicted, advertisement, area, assiduously, axe, balancing, bedad

Zipf's law - semantics

The more frequent a word, the more meanings it has.

Think of the words that: * are common in metaphors, e.g. *touch*, *drink* * can function as various parts of speech, e.g. *round* - verb, noun, adj, adv, prep

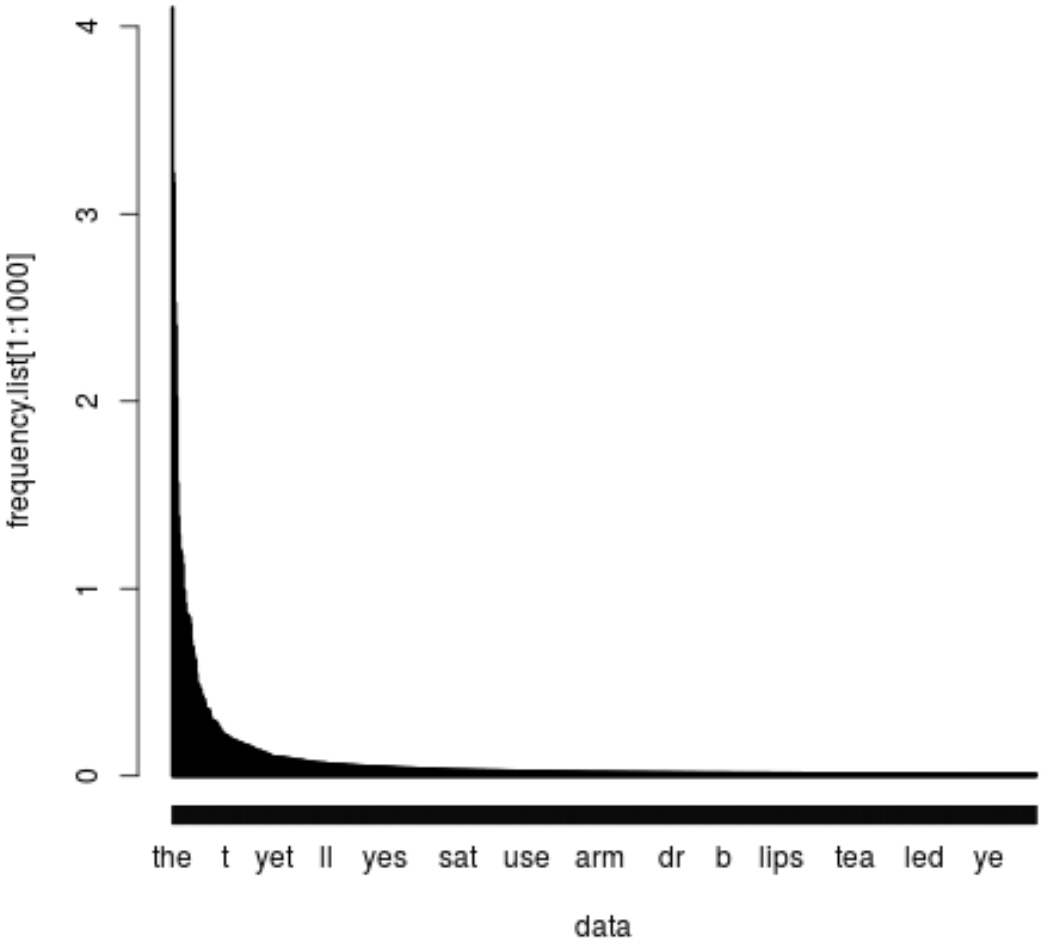
Zipf's law - Principle of Least Effort

Image source: learning libraries

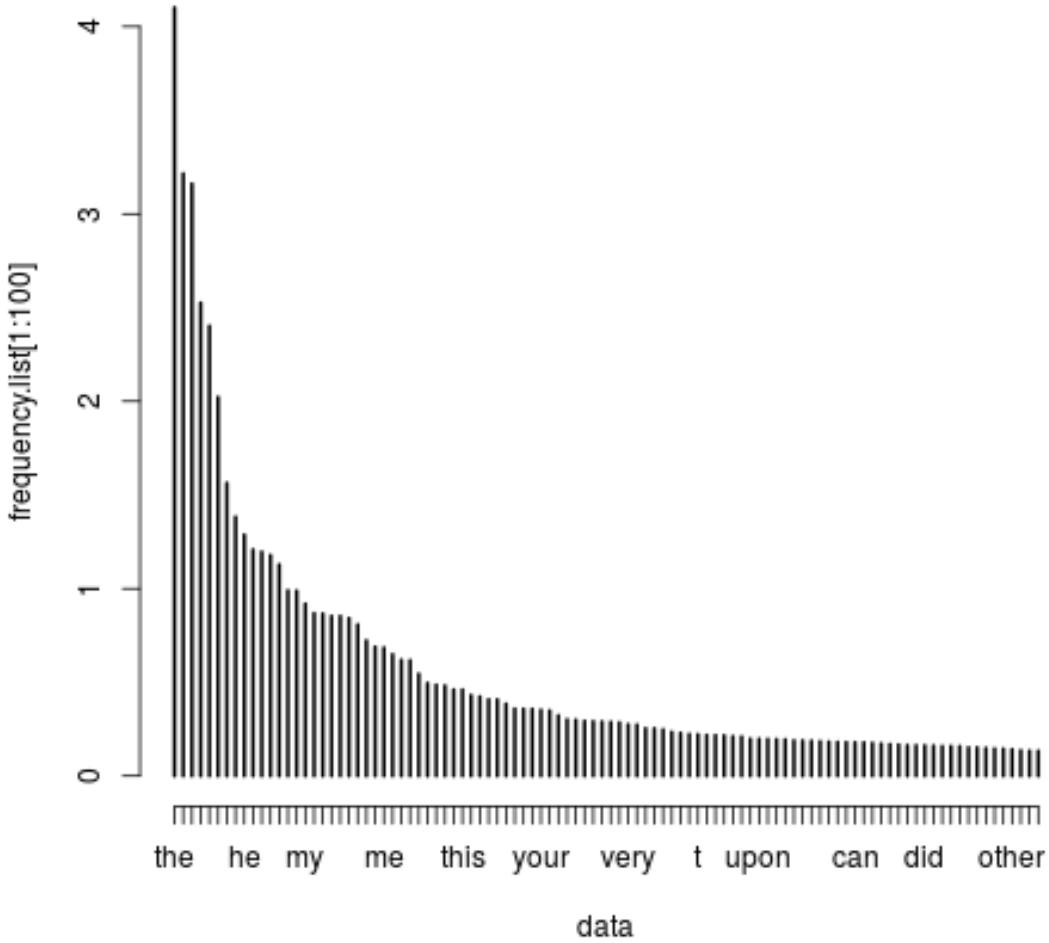
Zipf's law - Principle of Least Effort

- people naturally choose the path of least resistance, also when writing
- most authors use very uncommon words very rarely - this is sometimes measured with methods of *lexical richness*

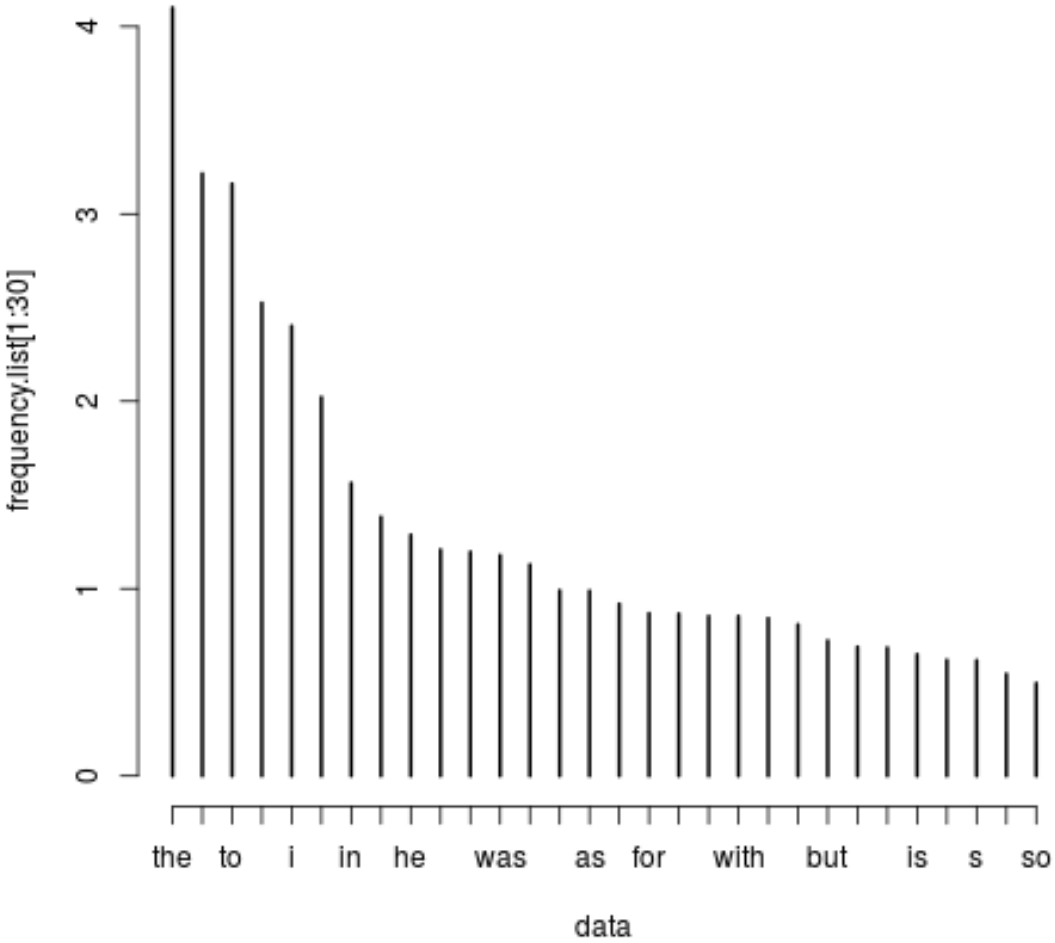
1000 most frequent words - distribution



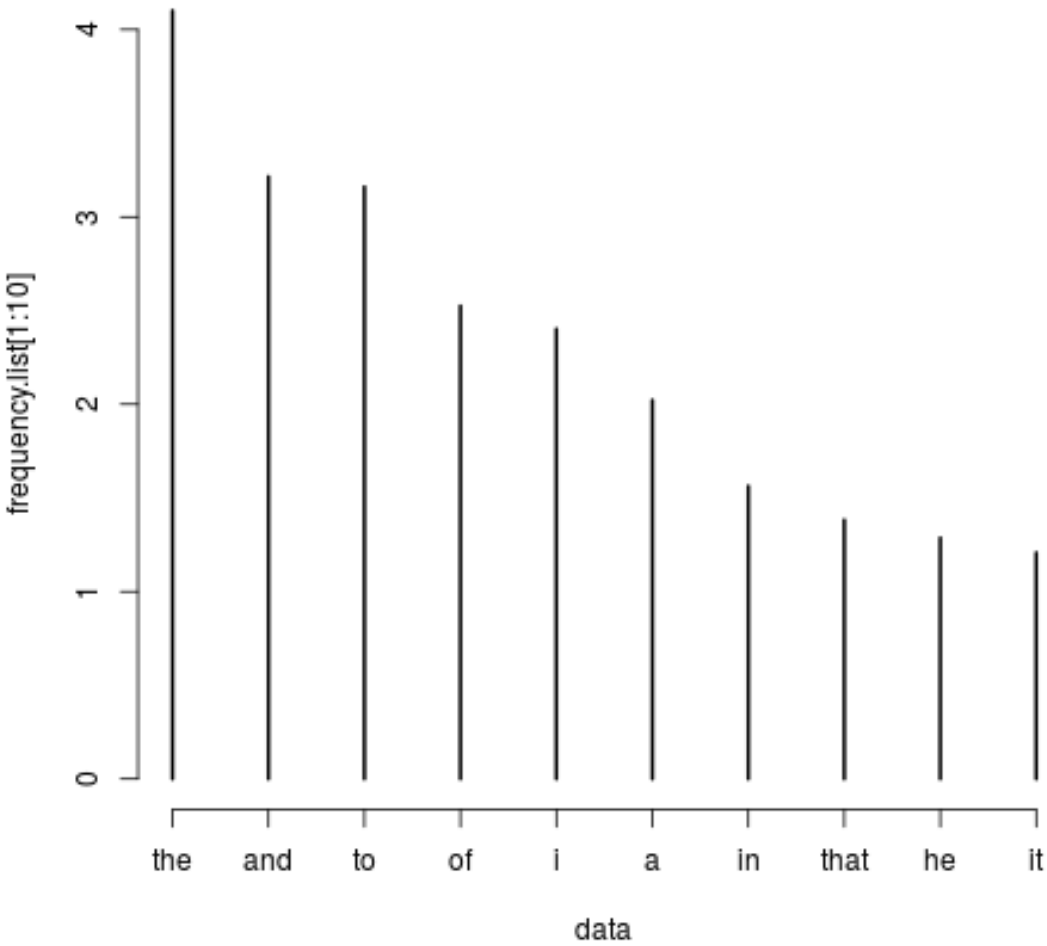
100 most frequent words - distribution



30 most frequent words - distribution



10 most frequent words - distribution



How many words to choose?

A difficult question that we face as we need to consider:
* which words are *really* important,
* where to cut the wordlist to obtain valid feature set,

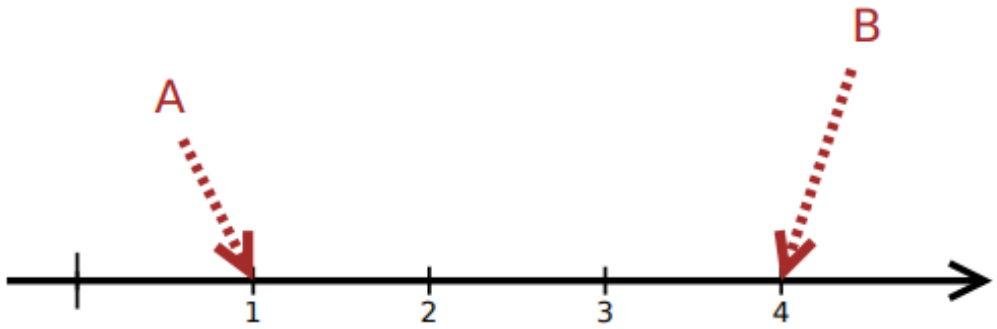
Are the words equally important?

The process of analysis of that consists of:
* selecting of words
* creating table of frequencies
* using distance measures to assess similarity

But what is the “distance” between words or texts?

Two texts, one dimension, (just one word as a difference)

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...

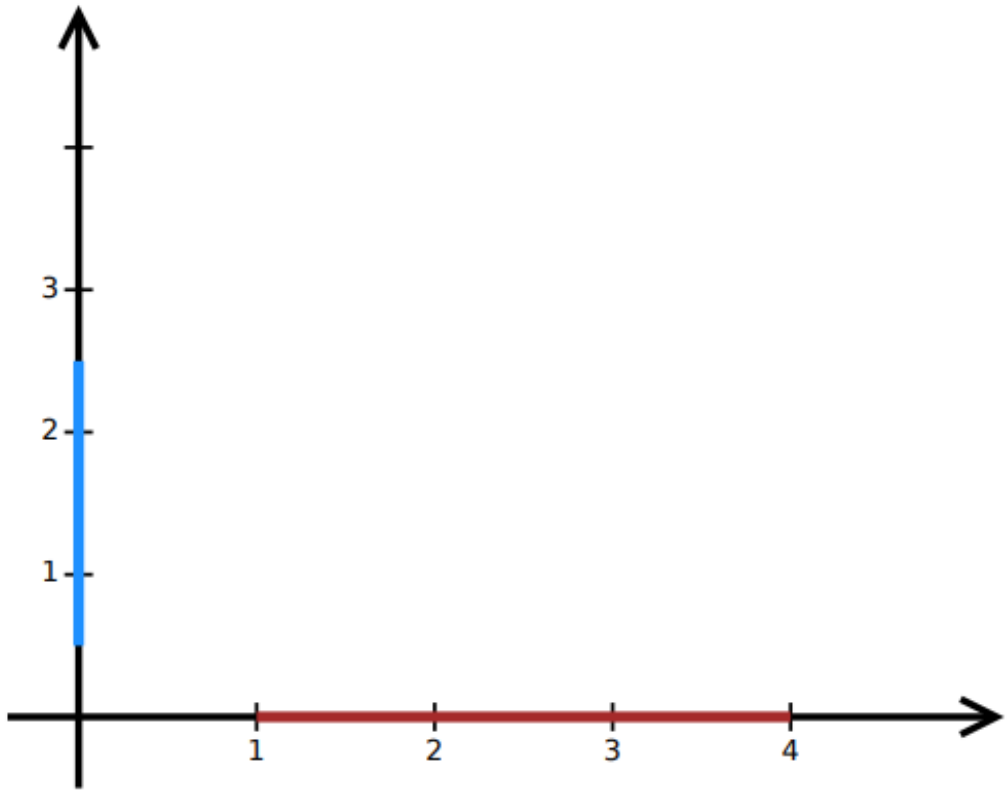


Two texts, one dimension, (just one word as a difference)

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...

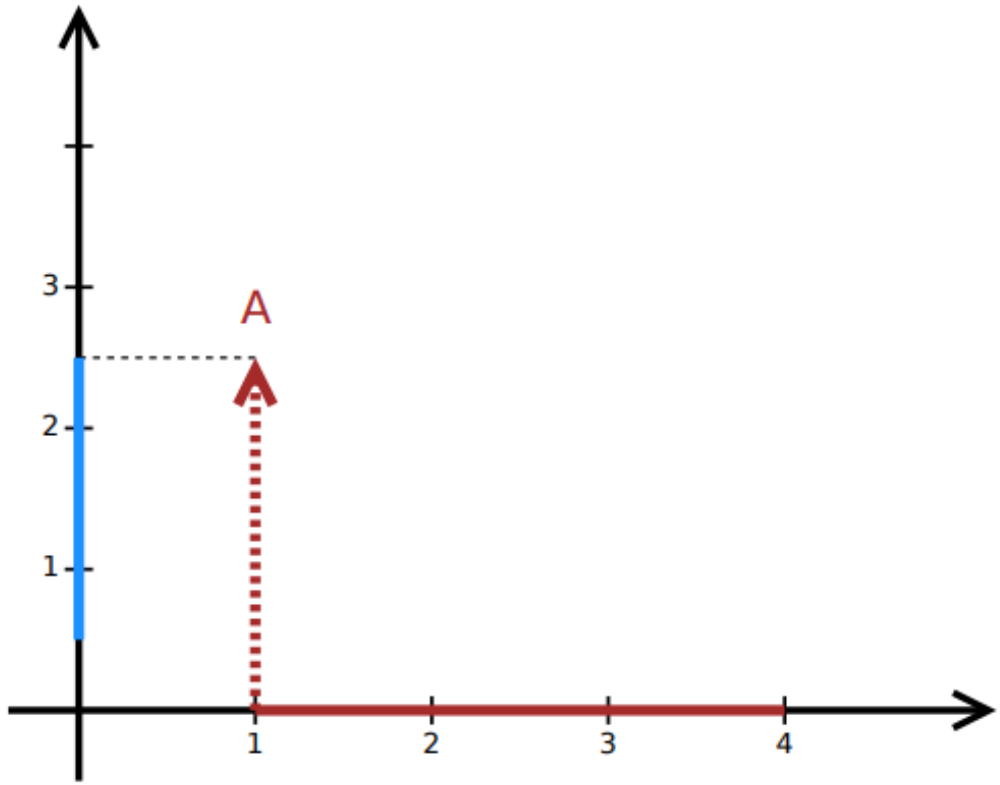
Two texts, two dimensions (difference based on 2 words)

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



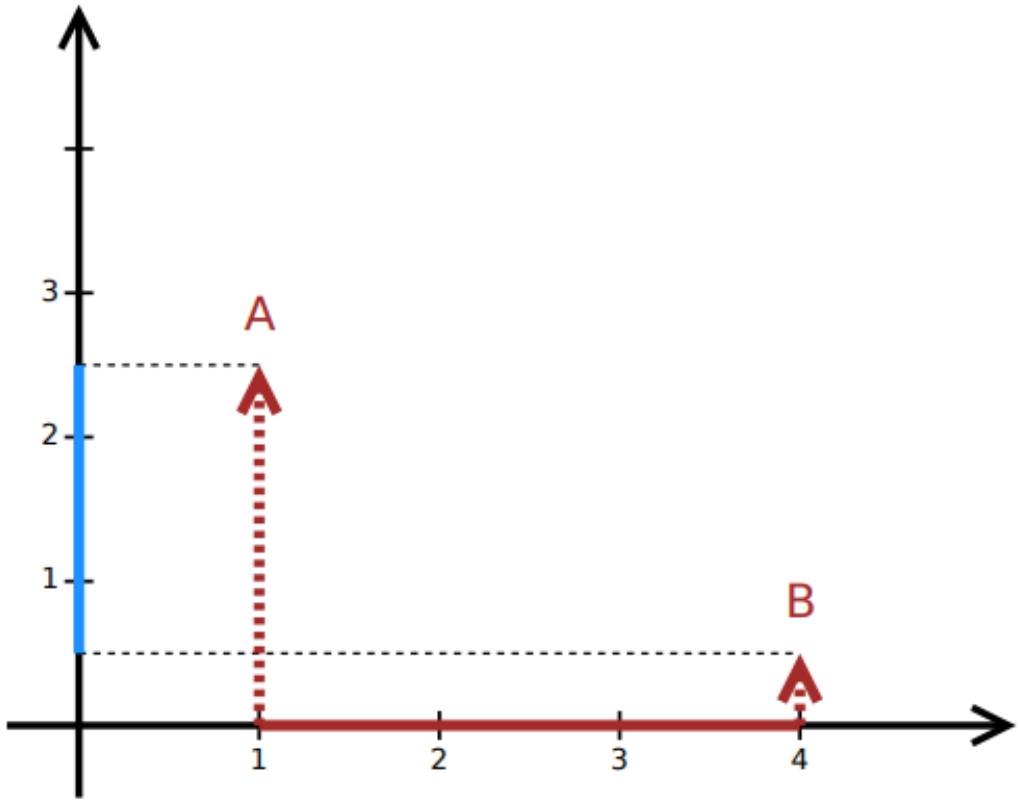
Constructing a two-dimensional space

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



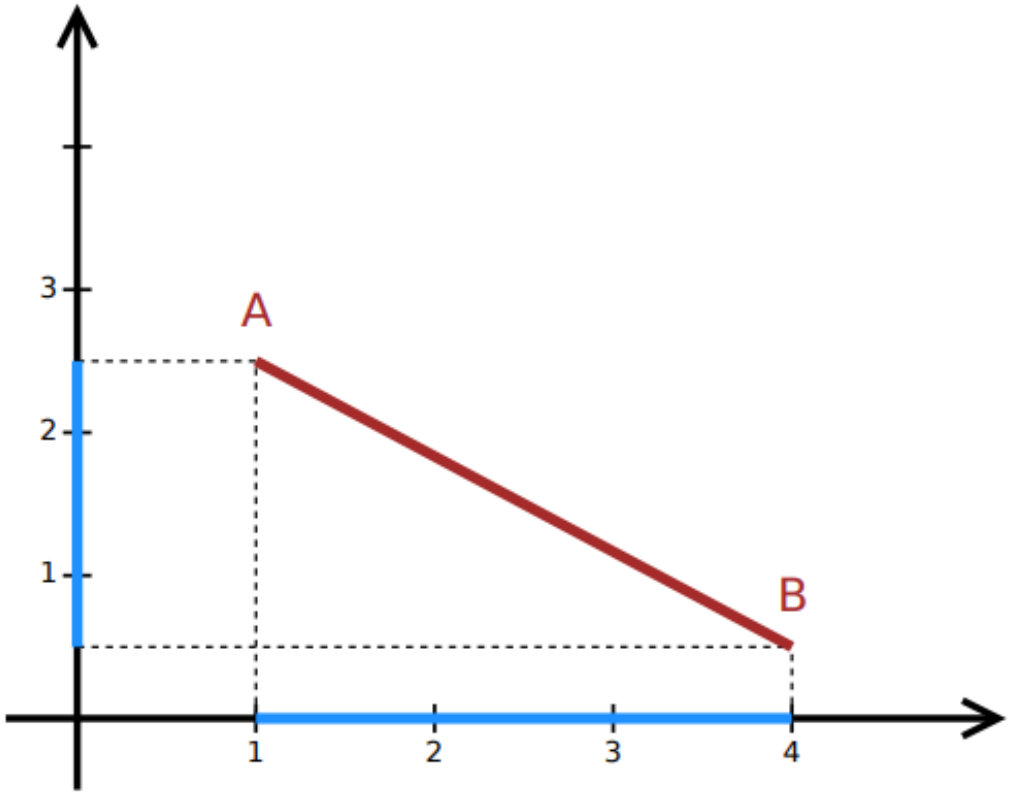
Constructing a two-dimensional space

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



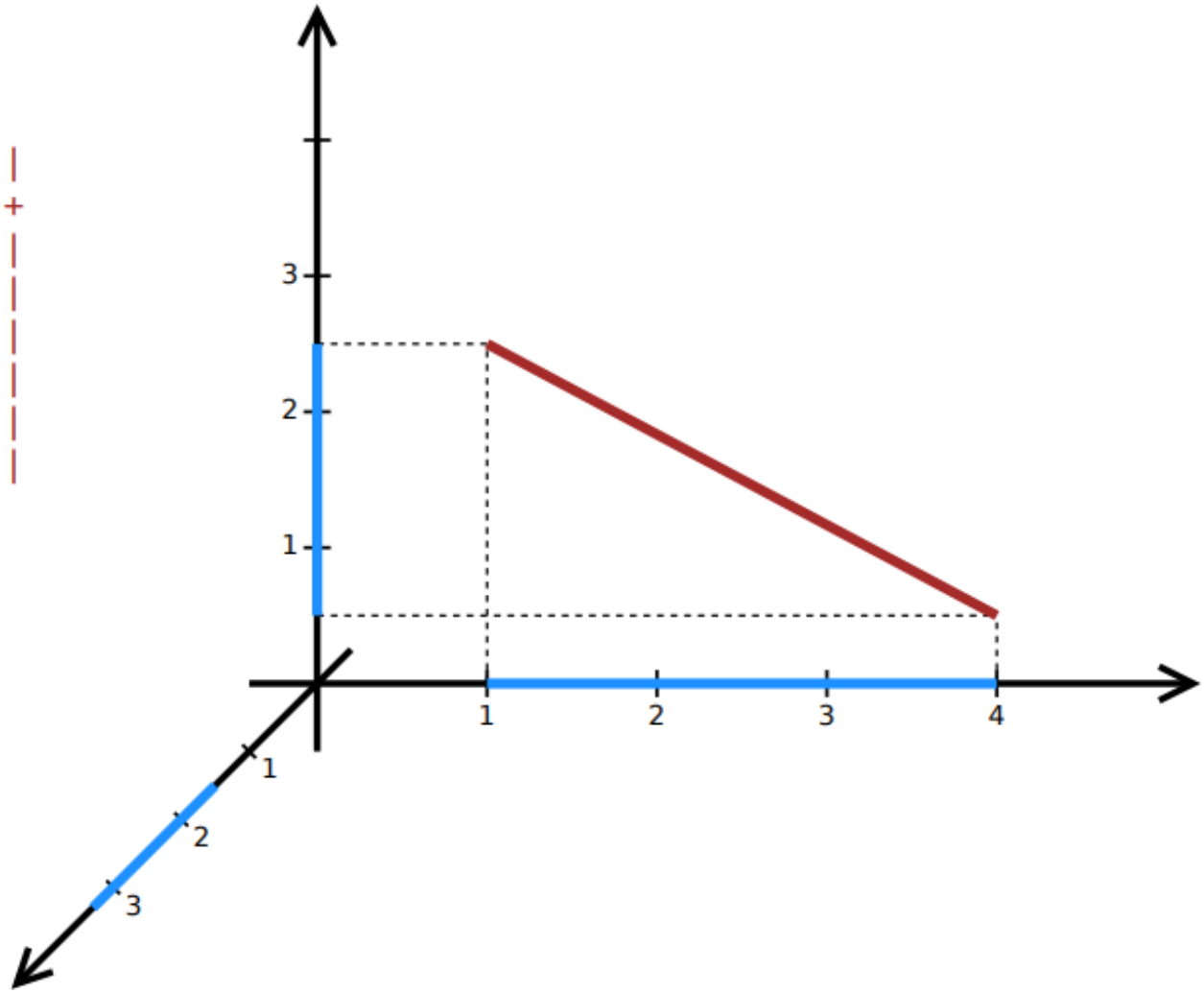
Two texts, two dimensions (difference based on two words)

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



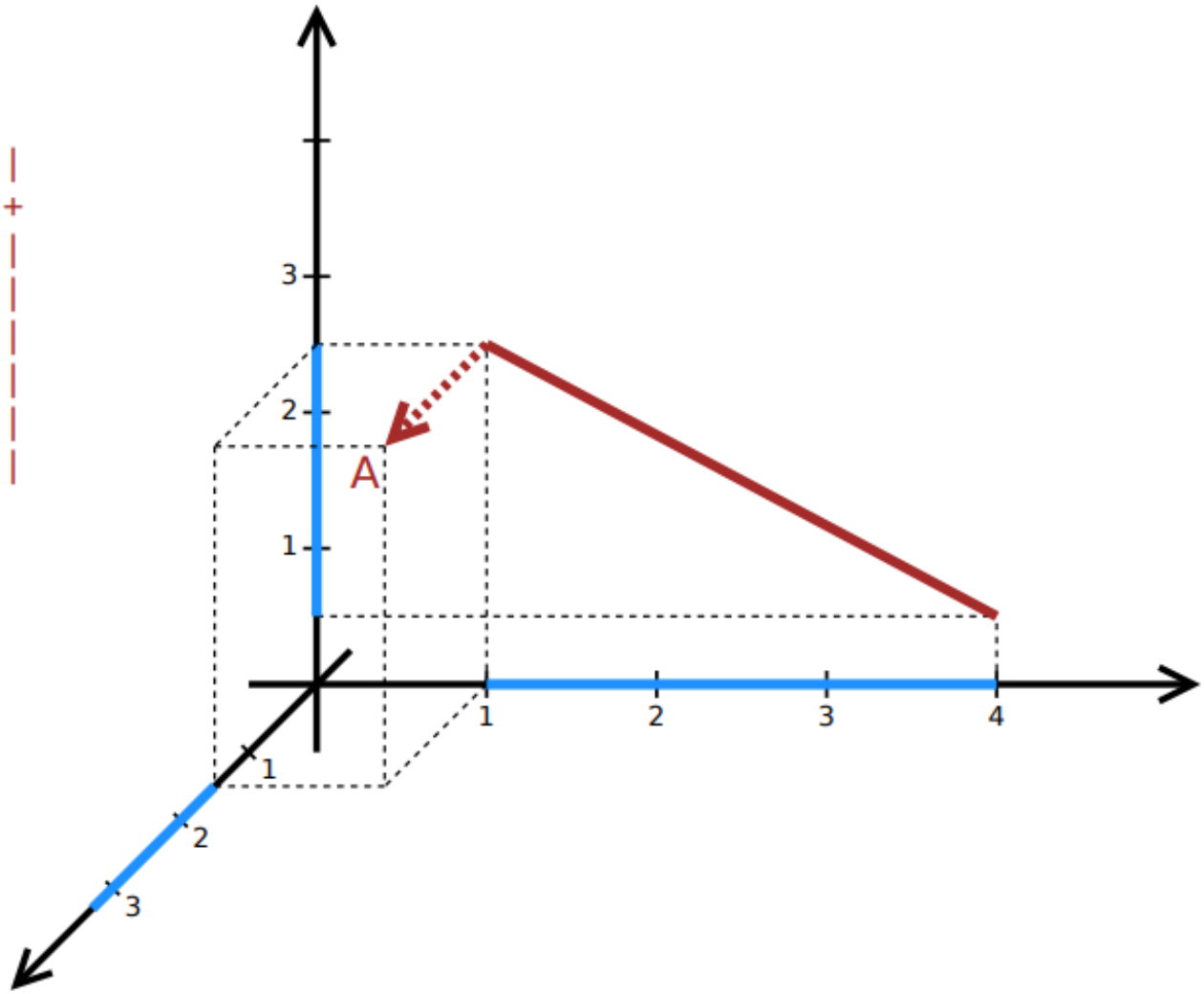
Two texts, three dimensions (difference based on three words)

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



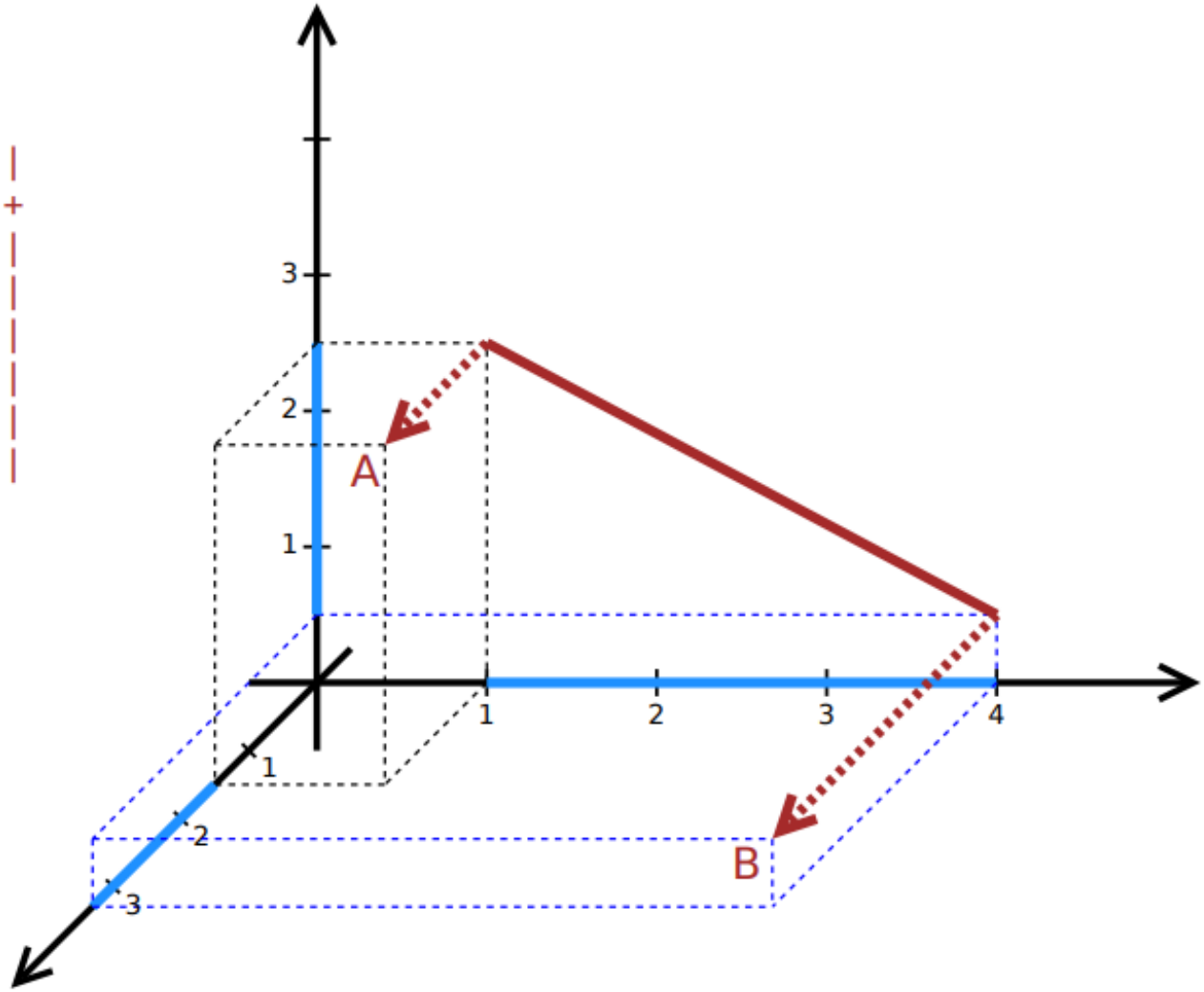
Constructing a three-dimensional space

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



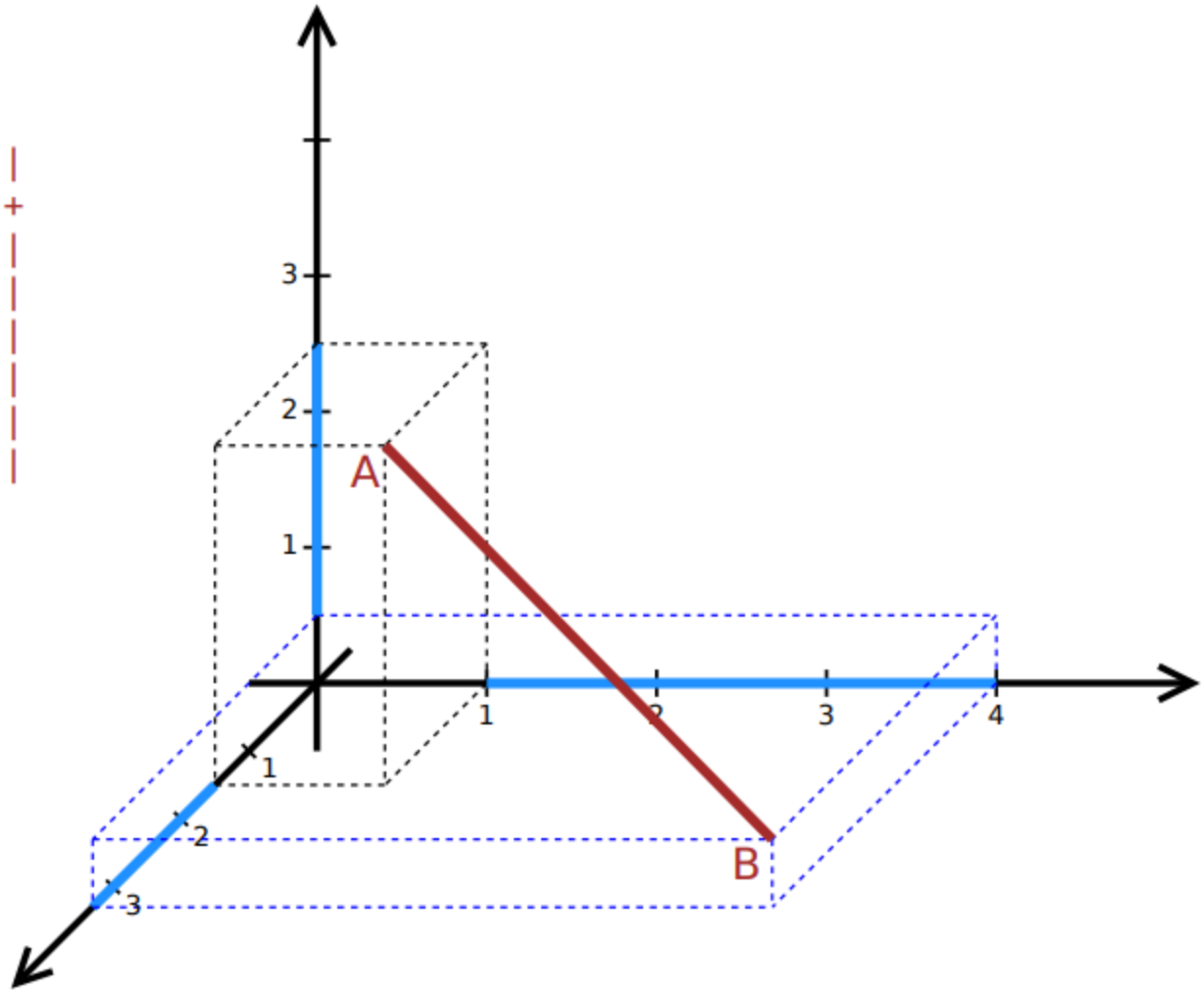
Constructing a three-dimensional space

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



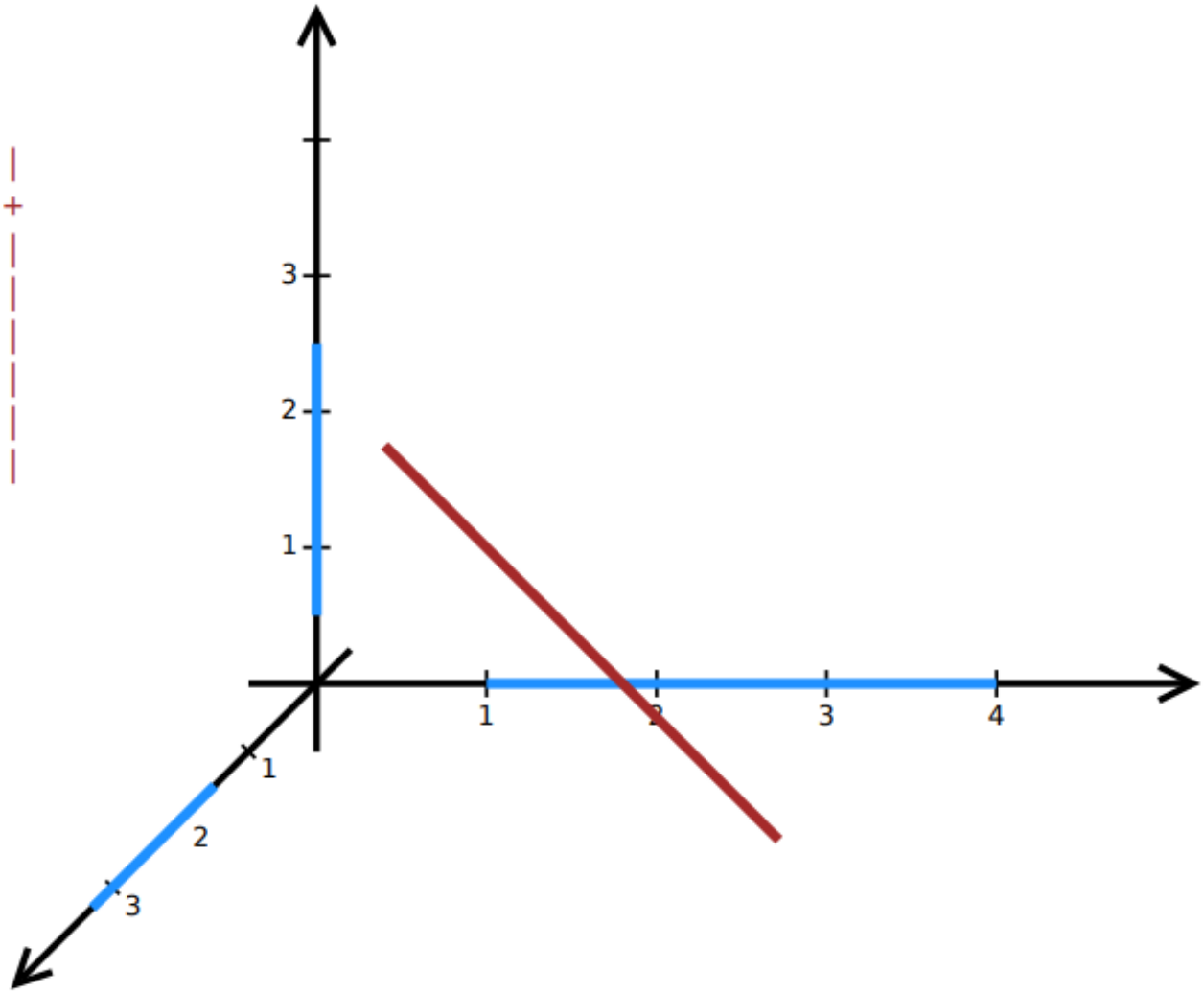
Two texts, three dimensions (difference based on three words)

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



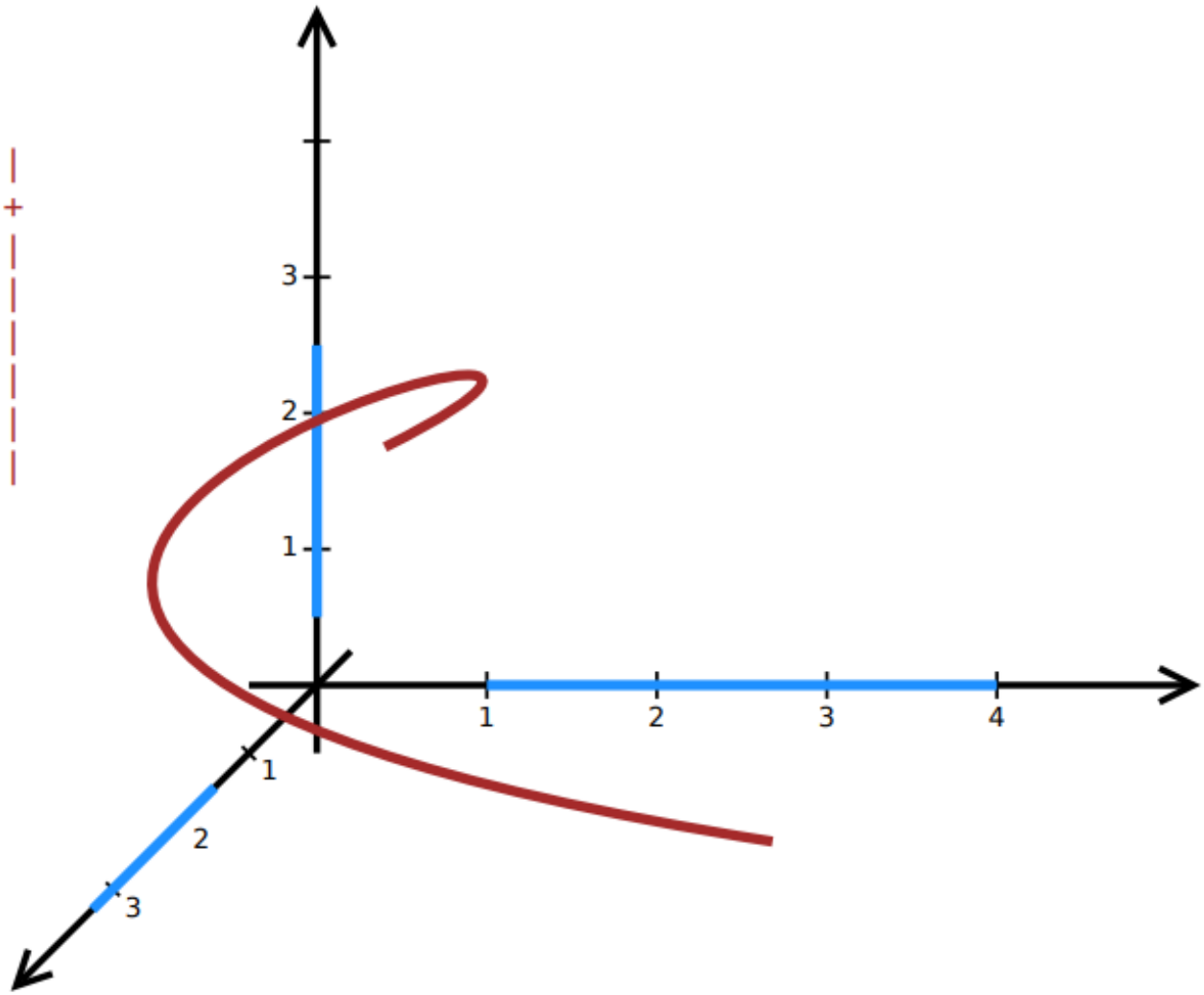
Shall we try the fourth dimension? What would it look like?

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



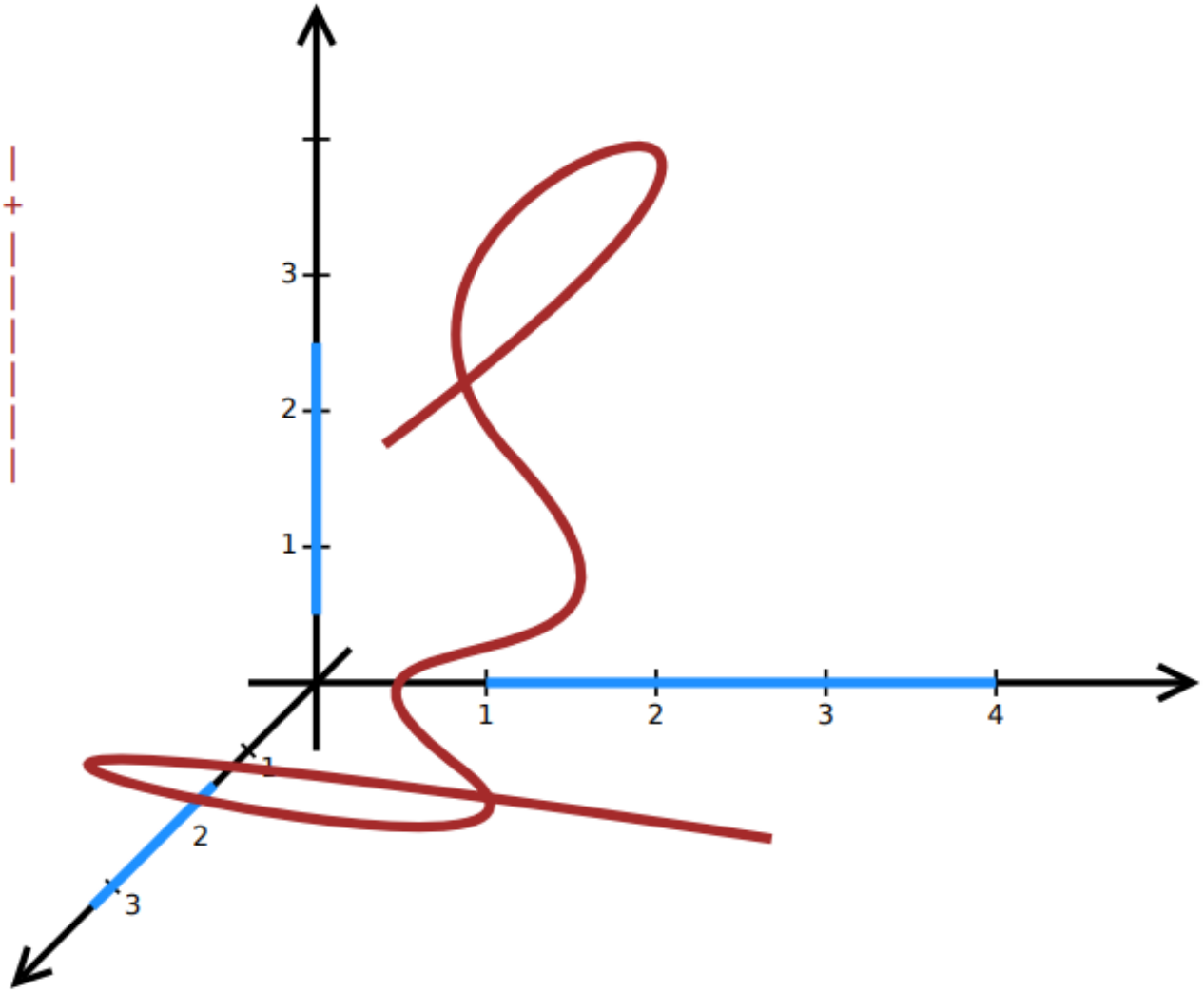
Two texts, MANY dimension, one difference based on many words

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...



One (hypothetical) difference in a multi-dimensional space?

	A	B
and	1	4
is	2.5	0.5
of	1.5	3.2
not	0.7	0.4
for	0.2	0.2
...	...	...

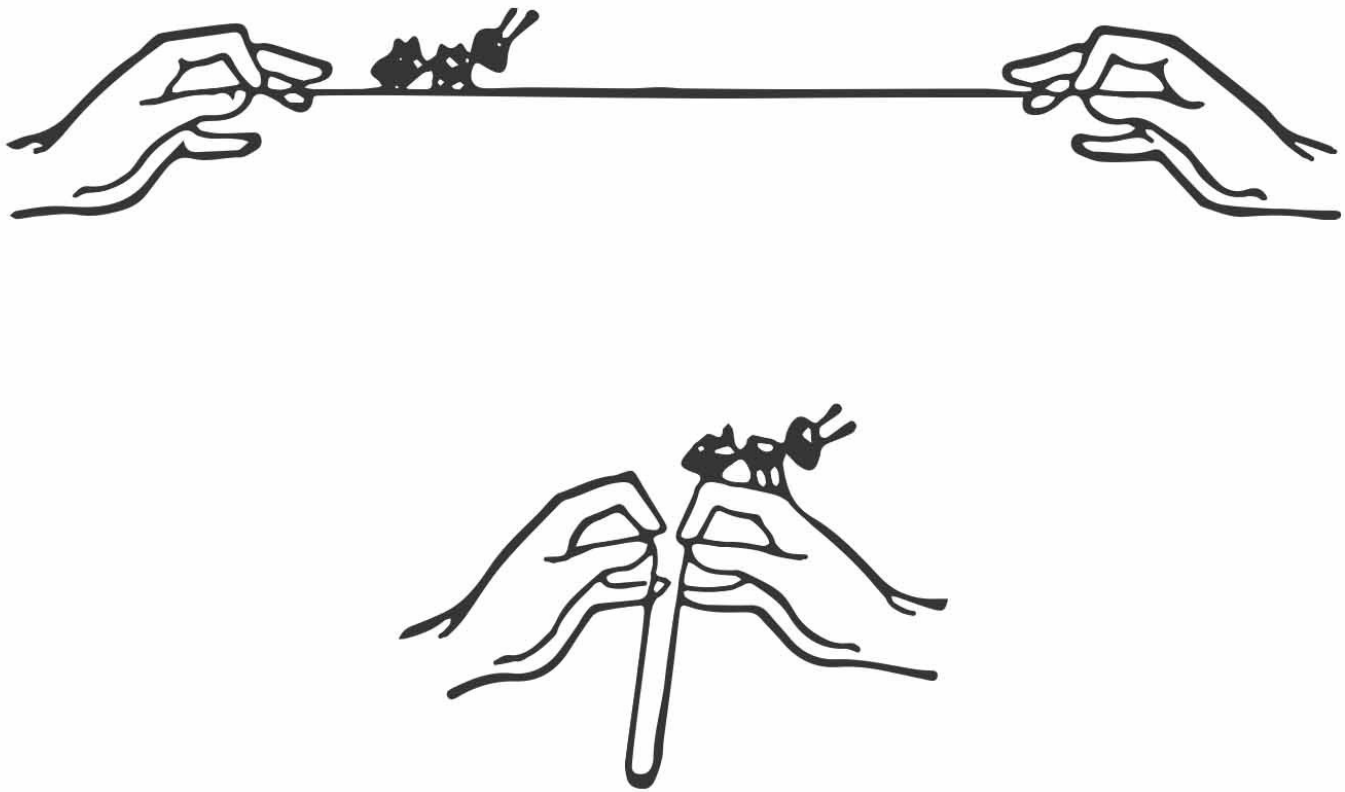


Dimension reduction

Why do we need it?
It's hard to visualise more dimensions than 3, and it's also hard to accurately compute and process multidimensional data. Each new unique word in a corpus adds a new dimension, so the number is rising to quite large values.

Now think of a spreadsheet of many columns and rows, and try to imagine it has so many further dimensions that it becomes a bit like in the last picture. Maybe it looks a bit like in Intestellar tesseract scene. . . ?

Or think of "A Wrinkle in Time"'s explanation of time travel!

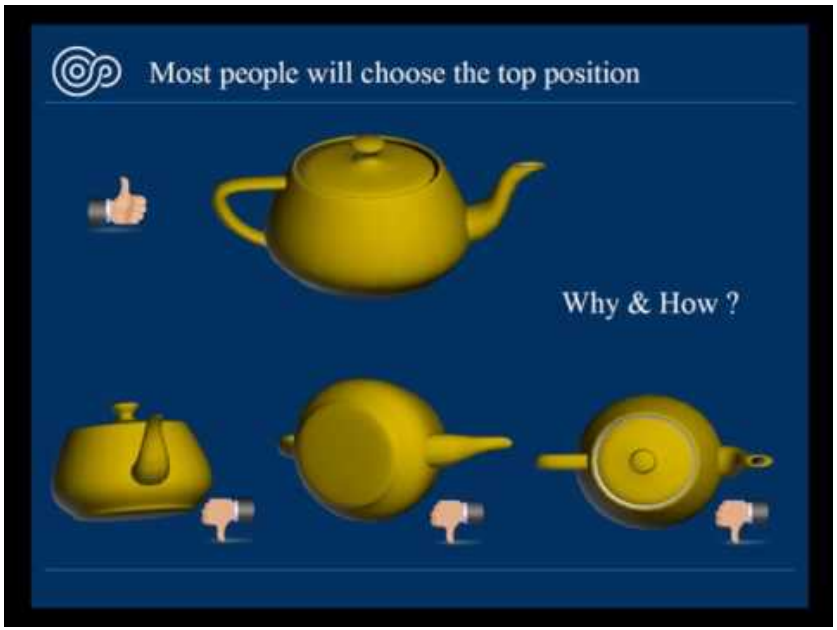


Two basic concepts related to dimension reduction:

- using a table (matrix) of distances (clustering, MDS)
- rotating a multidimensional space (PCA)

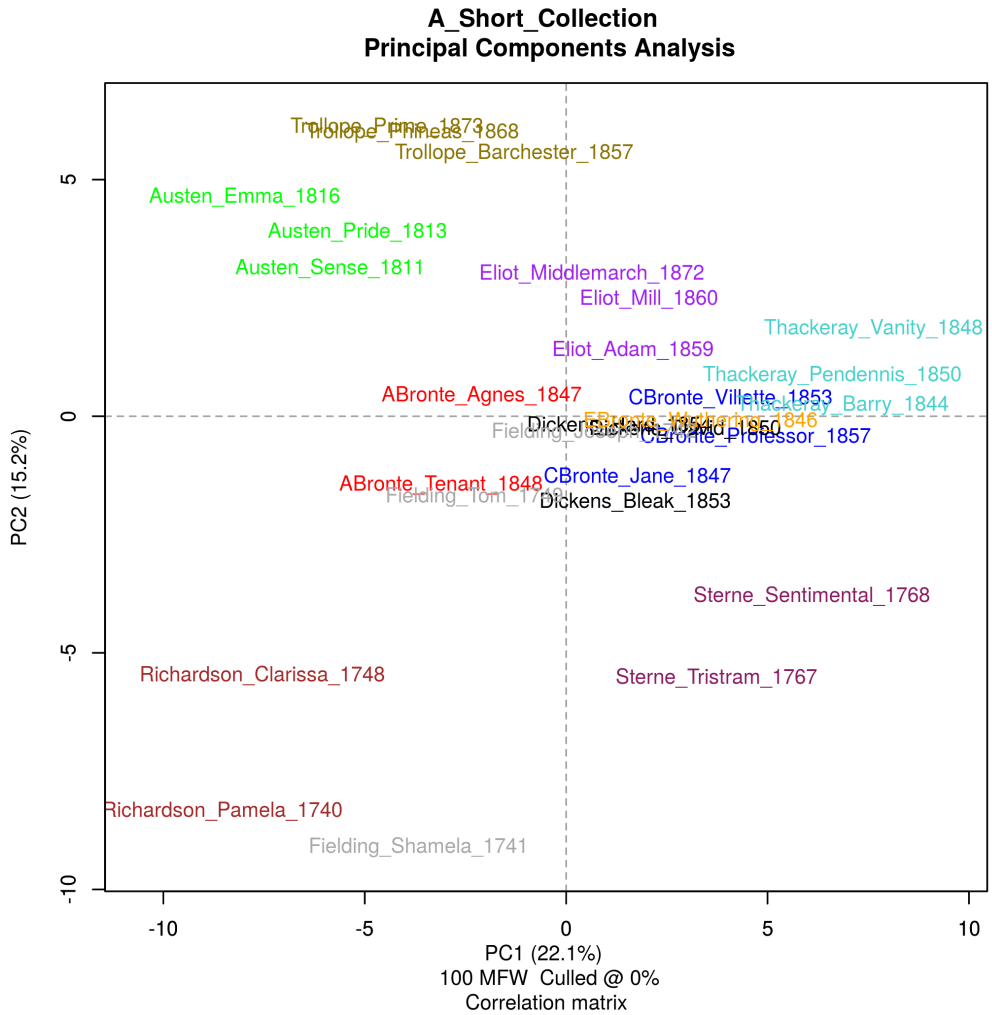
PCA = Principal Components Analysis

- Basic concept: looking at data to find coordinates that represent the biggest amount of information - finding the most optimal representation.
- E.g. drawing a kettle so as to give most details of what it looks like might look like this:



Information loss in PCA

- Components in PCA are sets of features which grant best information preservation, note however that they always have a value covering a relatively small percantage of all information.



Computing distances using table of frequencies

word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_Pr
the	3.68	3.54	3.24	3.56
and	4.00	4.01	3.04	2.94
to	3.46	3.34	3.23	3.40
of	2.34	2.23	2.67	2.96
i	3.22	3.63	1.99	1.70

Computing distances using table of frequencies

word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_PP
the	**3.68**	**3.54**	3.24	3.56
and	4.00	4.01	3.04	2.94
to	3.46	3.34	3.23	3.40
of	2.34	2.23	2.67	2.96
i	3.22	3.63	1.99	1.70

|a1 - b1|

Computing distances using table of frequencies

word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_PP
the	3.68	3.54	3.24	3.56
and	**4.00**	**4.01**	3.04	2.94
to	3.46	3.34	3.23	3.40
of	2.34	2.23	2.67	2.96
i	3.22	3.63	1.99	1.70

$$|\mathbf{a1} - \mathbf{b1}| + |\mathbf{a2} - \mathbf{b2}|$$

Computing distances using table of frequencies

word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_PP
the	3.68	3.54	3.24	3.56
and	4.00	4.01	3.04	2.94
to	**3.46**	**3.34**	3.23	3.40
of	2.34	2.23	2.67	2.96
i	3.22	3.63	1.99	1.70

$$|\mathbf{a1} - \mathbf{b1}| + |\mathbf{a2} - \mathbf{b2}| + |\mathbf{a3} - \mathbf{b3}|$$

Computing distances using table of frequencies

word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_PP
the	3.68	3.54	3.24	3.56
and	4.00	4.01	3.04	2.94
to	3.46	3.34	3.23	3.40
of	**2.34**	**2.23**	2.67	2.96
i	3.22	3.63	1.99	1.70

$$|\mathbf{a1} - \mathbf{b1}| + |\mathbf{a2} - \mathbf{b2}| + |\mathbf{a3} - \mathbf{b3}| + |\mathbf{a4} - \mathbf{b4}|$$

Computing distances using table of frequencies

word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_PP
the	3.68	3.54	3.24	3.56
and	4.00	4.01	3.04	2.94
to	3.46	3.34	3.23	3.40
of	2.34	2.23	2.67	2.96
i	**3.22**	**3.63**	1.99	1.70

$$|\mathbf{a1} - \mathbf{b1}| + |\mathbf{a2} - \mathbf{b2}| + |\mathbf{a3} - \mathbf{b3}| + |\mathbf{a4} - \mathbf{b4}| + |\mathbf{a5} - \mathbf{b5}|$$

Computing distances using table of frequencies

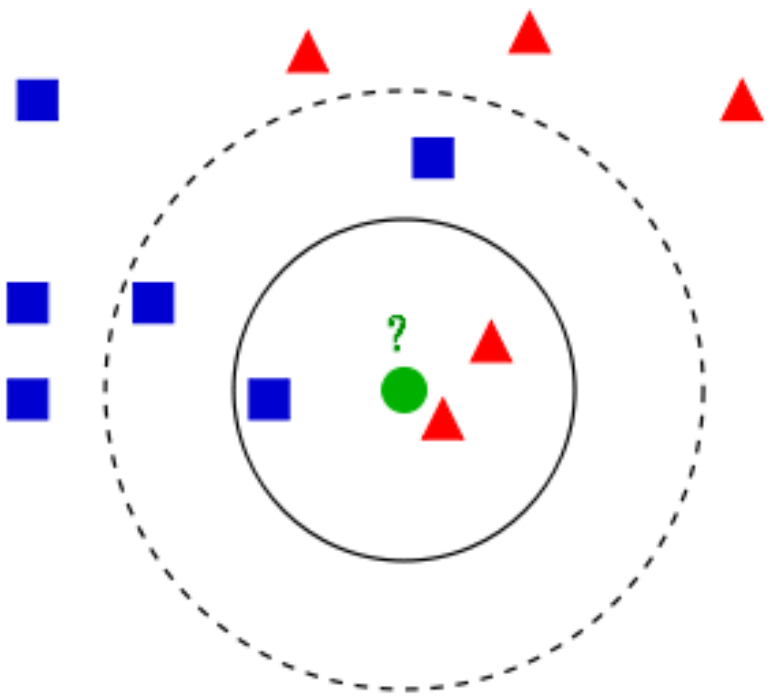
word	ABronte_Agn	ABronte_Ten	Austen_Em	Austen_PP
the	3.68	3.54	3.24	3.56
and	4.00	4.01	3.04	2.94
to	3.46	3.34	3.23	3.40
of	2.34	2.23	2.67	2.96
i	3.22	3.63	1.99	1.70

$$\sum_{i=1}^{10} |a_i - b_i|$$

Why do we need distance measures?

You now know how to calculate distance between two words (or globally: two texts).

k-nearest neighbours (k-NN) classifier:



Why do we need distance measures?

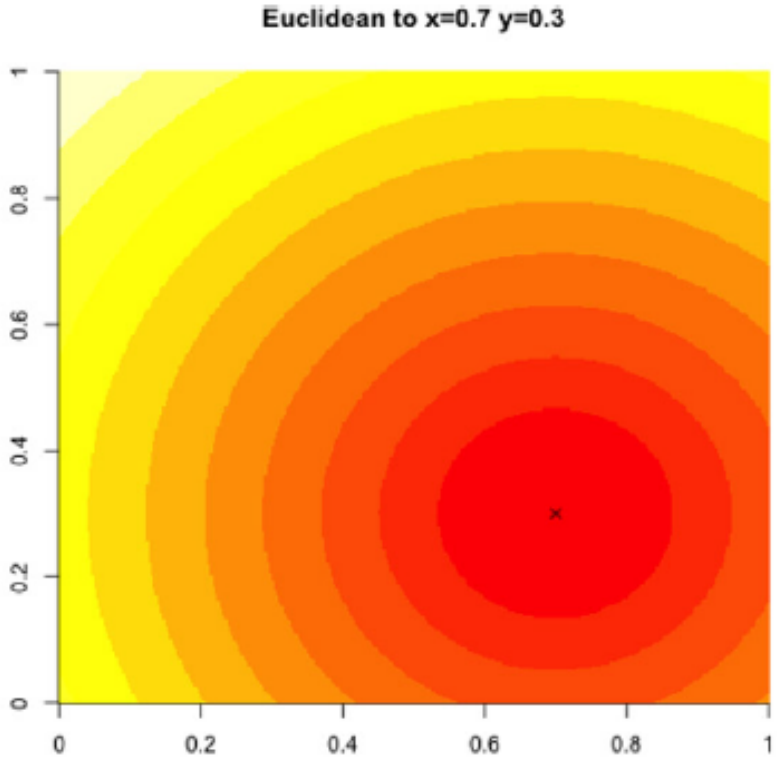
You now know how to calculate distance between two words (or globally: two texts).
The question remains:

* are we doing it in the way that works best for language?
* do we treat all the words the same if we know about Zipf's Law?

Stylometric distances as political systems

Distance measure	political system
Euclidean distance	tyranny
Manhattan distance	oligarchy
Classic delta	democracy
Eder's delta	feudal monarchy
Eder's simple	politeia (res publica)
Canberra	people's revolution

Euclidean distance

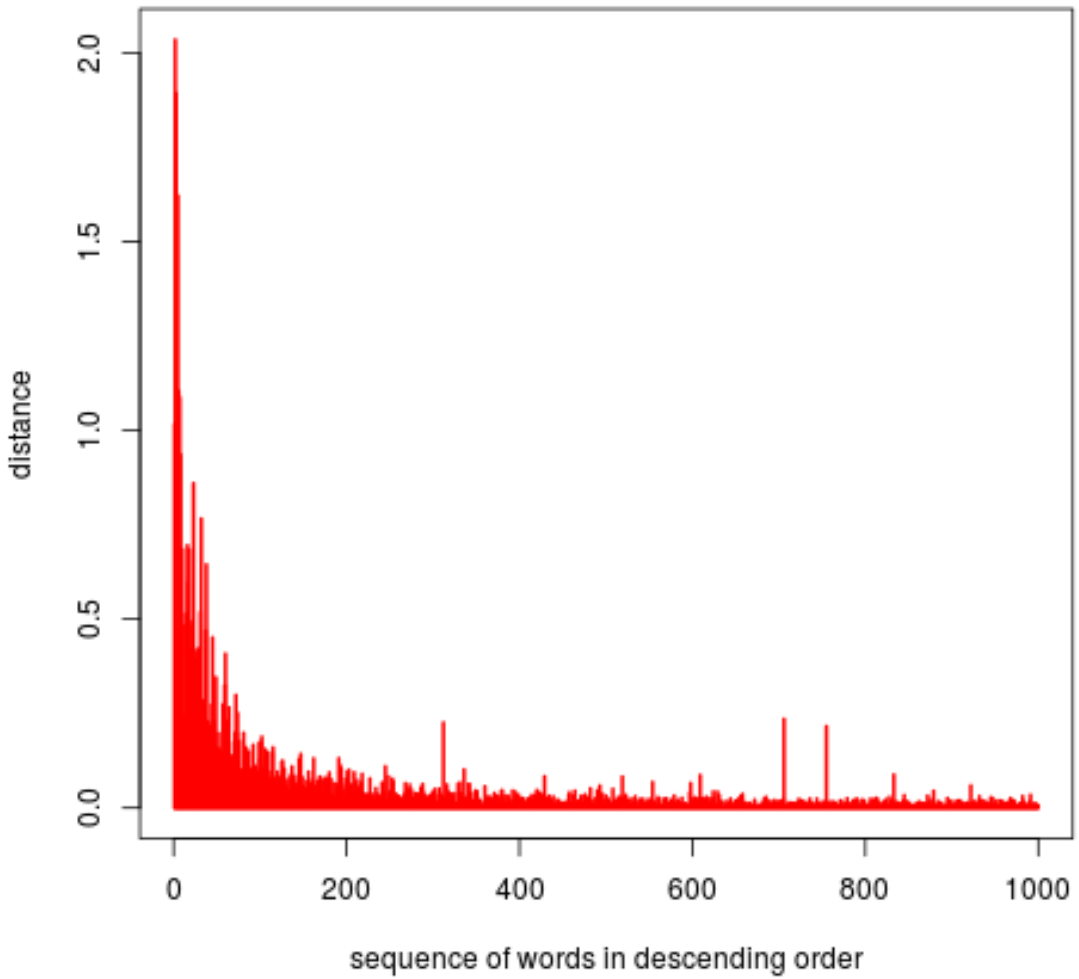


A classic distance measure:

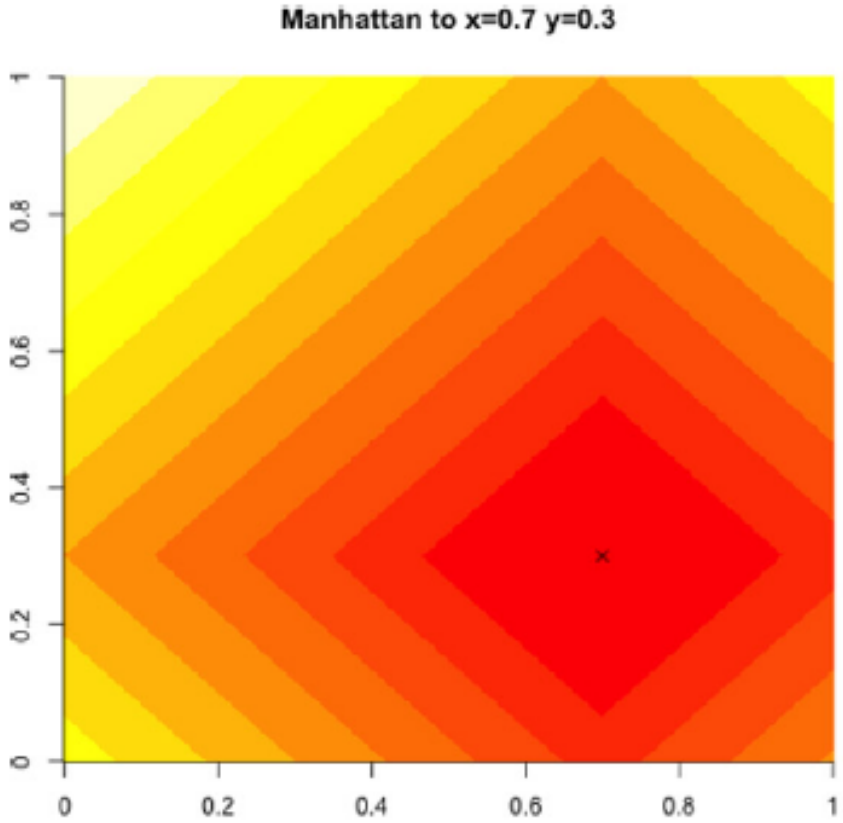
$$\delta_{AB} = \sum_{i=1}^{n} \sqrt{|f_i(A)^2 - f_i(B)^2|}$$

where:
n = the number of MFWs (most frequent words),
f_i = the frequency of a given word i,
A, B = text samples being compared

Euclidean distance: Tyranny of the most frequent word



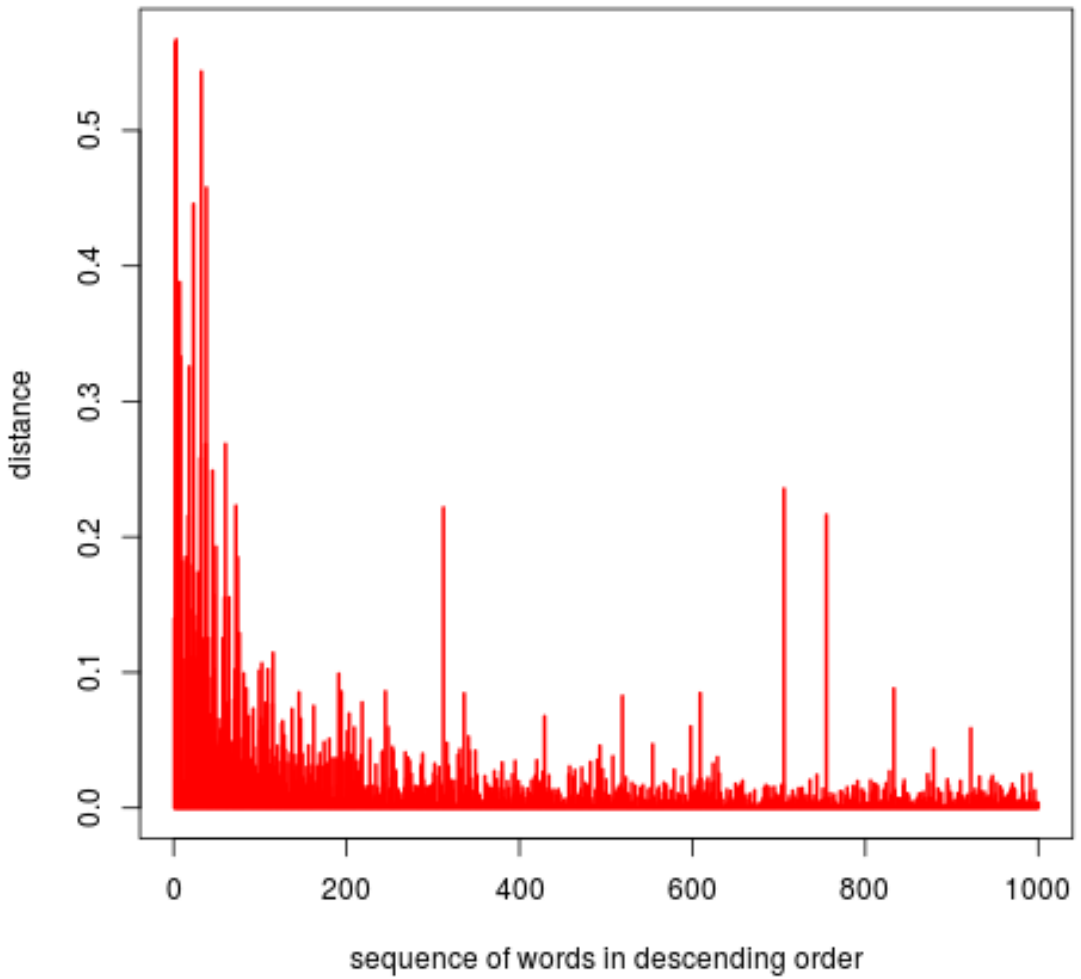
Manhattan distance



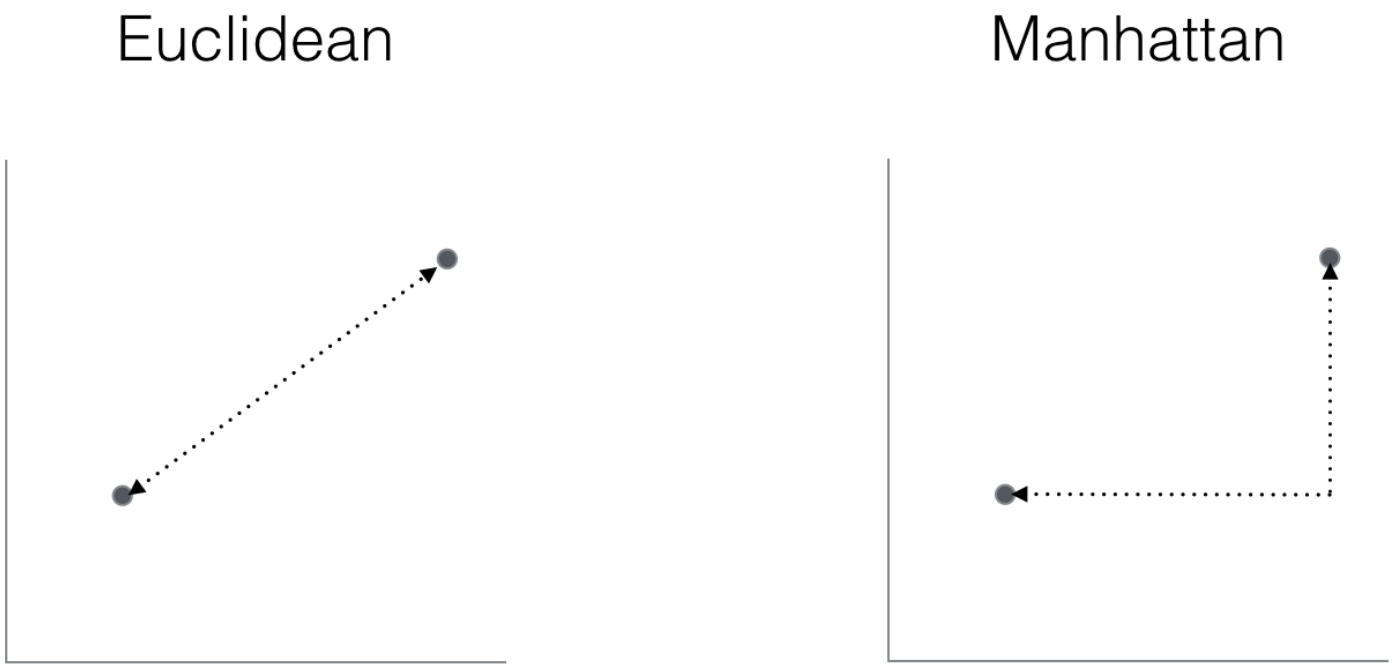
Another classic distance measure:

$$\delta_{AB} = \sum_{i=1}^{n} |f_i(A) - f_i(B)|$$

Manhattan distance: Oligarchy, or a small group of rulers



Euclidean vs Manhattan

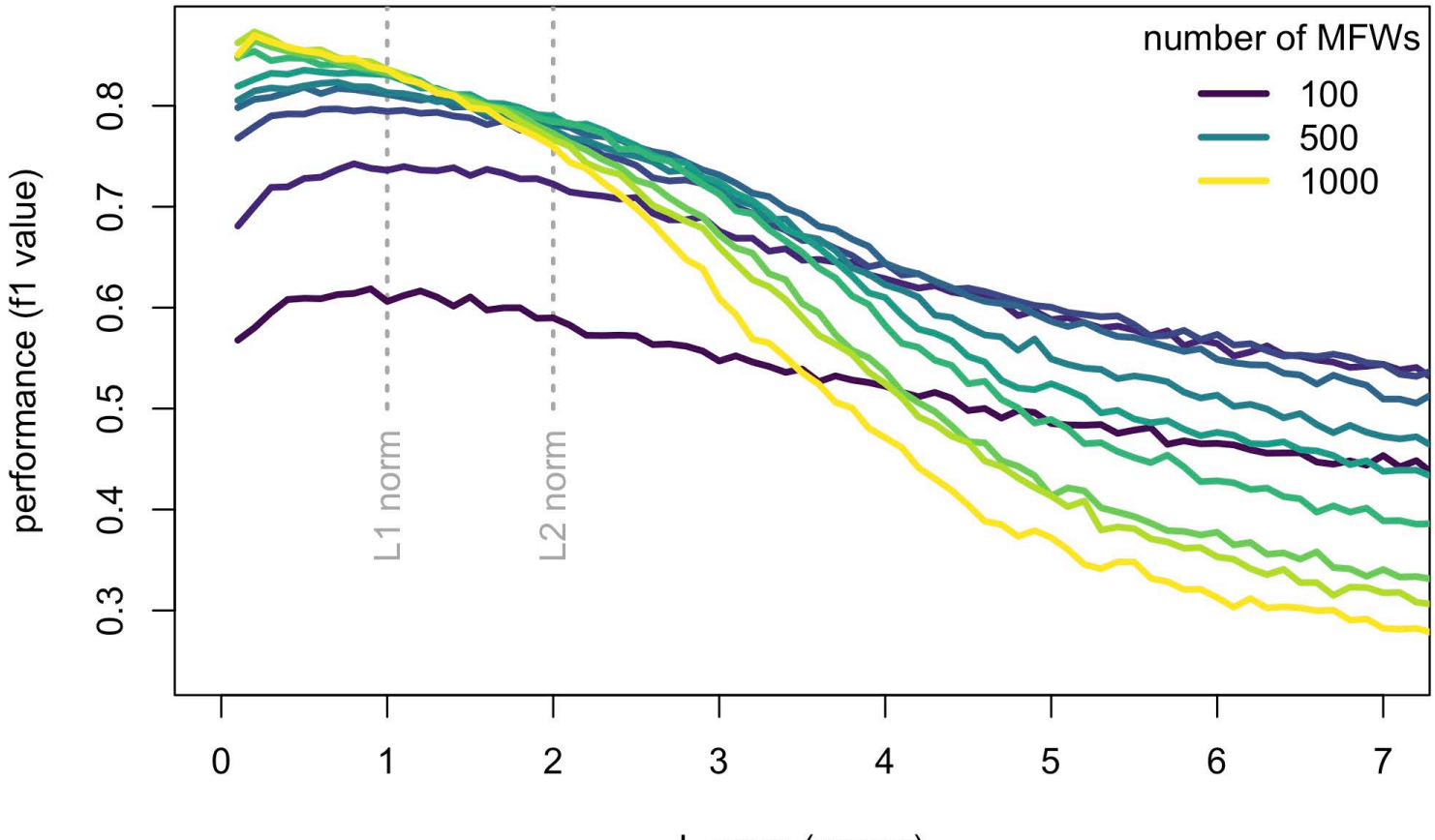


source: Kan Nishida - blog.exploratory.io

$$\delta_{AB} = \sum_{i=1}^{n} \sqrt{|f_i(A)^2 - f_i(B)^2|}$$

$$\delta_{AB} = \sum_{i=1}^{n} |A_i - B_i| \qquad \delta_{AB} = \sqrt[1]{\sum_{i=1}^{n} |A_i - B_i|^1}$$

$$\delta_{AB} = \sqrt[p]{\sum_{i=1}^{n} |A_i - B_i|^p}$$



$$\delta_{AB} = \sqrt[p]{\sum_{i=1}^{n} |A_i - B_i|^p}$$

Classic delta

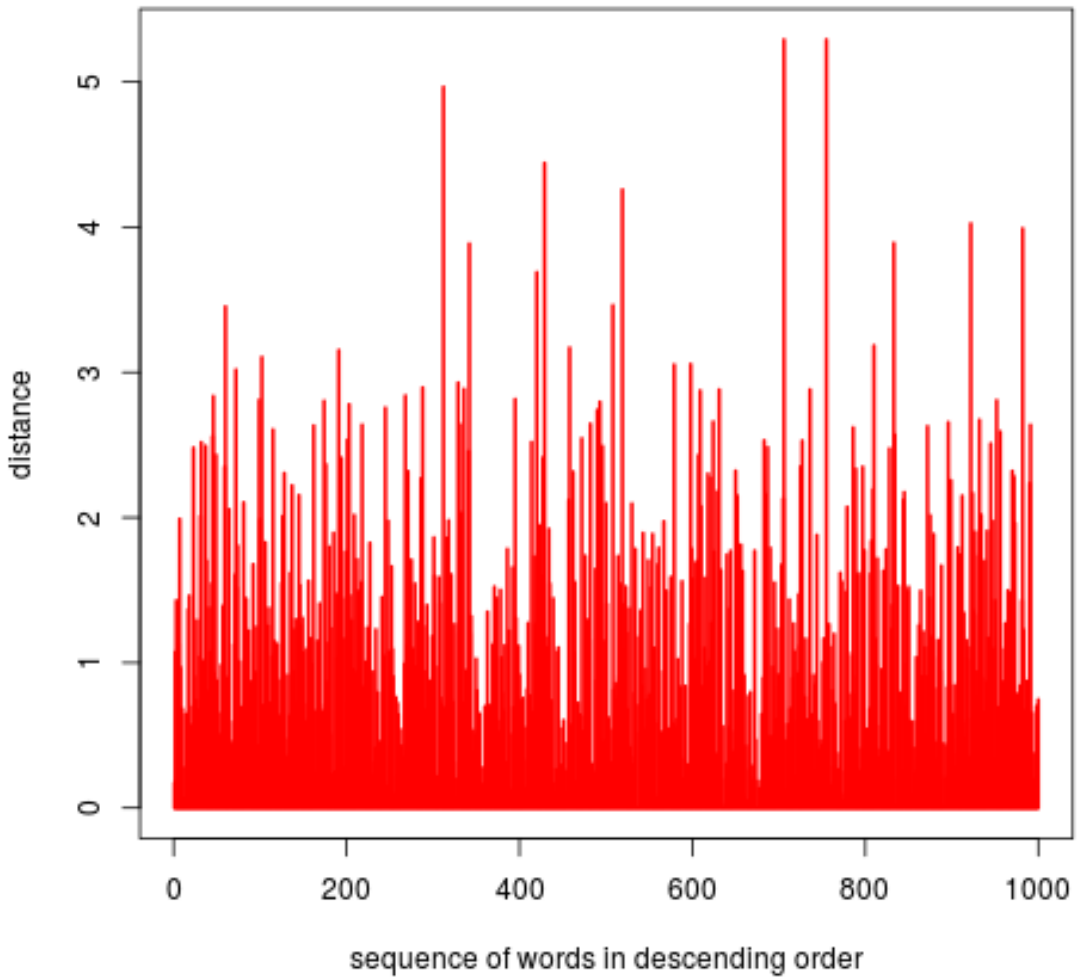
Delta distance as proposed by Burrows:

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i(A)-\mu_i}{\sigma_i} - \frac{f_i(B)-\mu_i}{\sigma_i}\right|$$

the same formula simplified (cf. Argamon, 2008):

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i(A)-f_i(B)}{\sigma_i}\right|$$

Classic delta: Democracy, or the same laws for all



Eder's modification to the delta distance

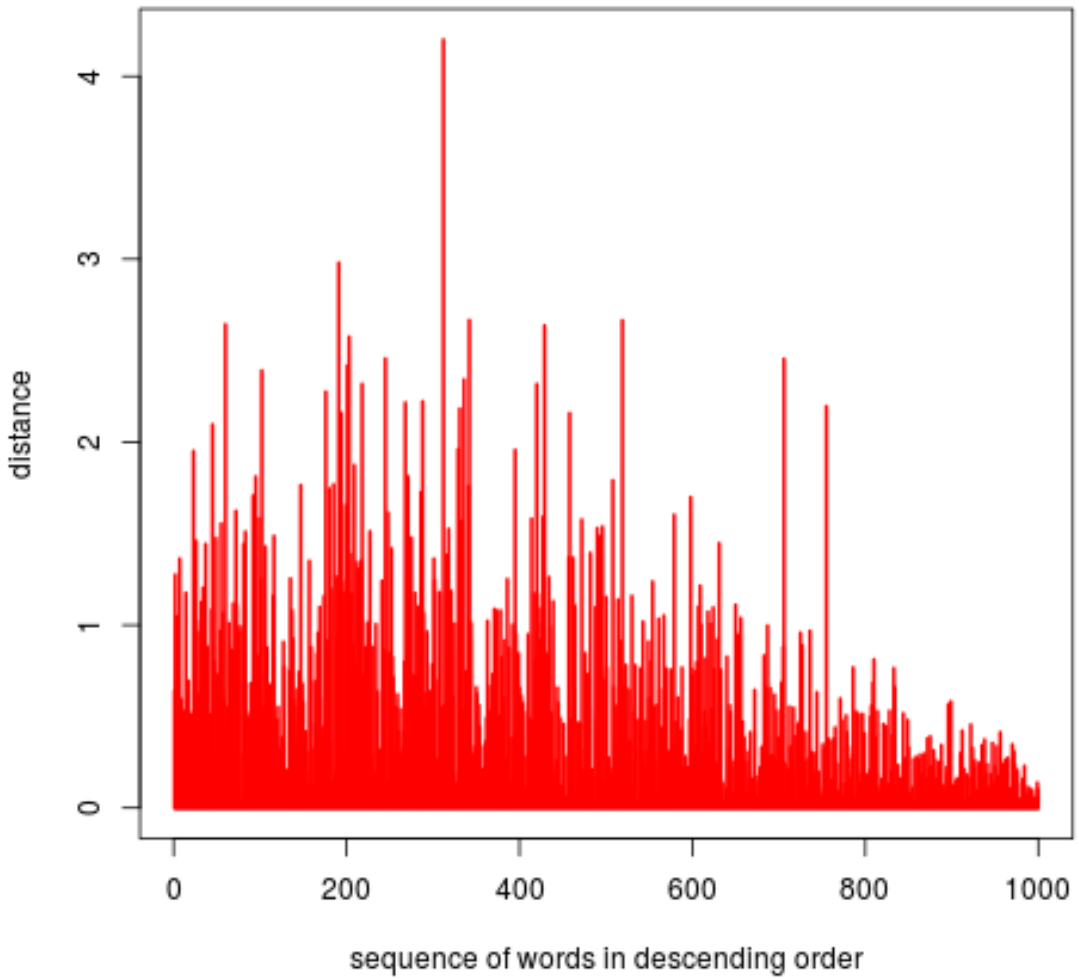
A method introducing weights to give slightly more influence to words that are more frequent:

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}(|\frac{f_i(A)-\mu_i}{\sigma_i} - \frac{f_i(B)-\mu_i}{\sigma_i}| \times \frac{n-n_i+1}{n})$$

the same formula simplified algebraically:

$$\Delta_{(AB)} = \frac{1}{n}\sum_{i=1}^{n}(|\frac{f_i(A)-f_i(B)}{\sigma_i}| \times \frac{n-n_i+1}{n})$$

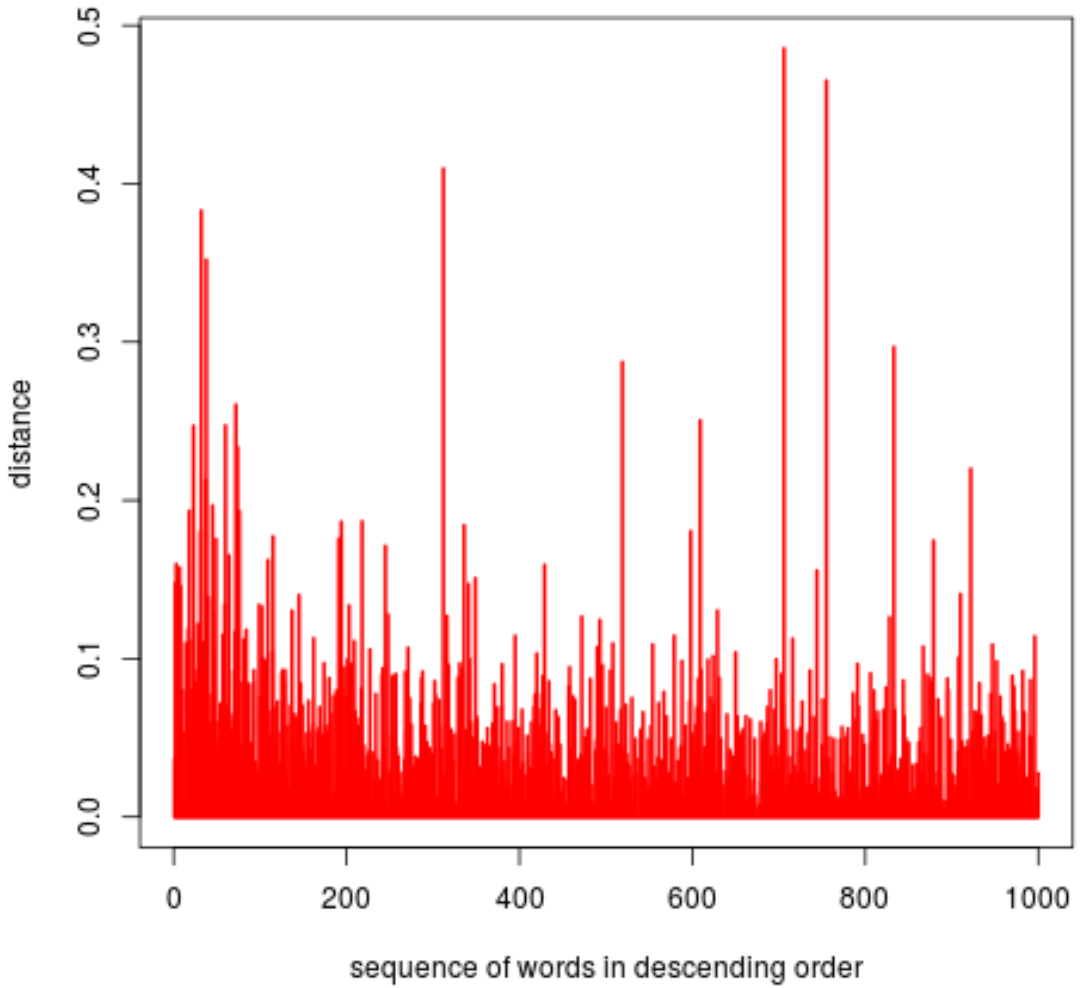
Eder's delta: Feudal monarchy, or the hierarchy of honors



Eder's simple (i.e. an anti-Zipf normalization)

$$\delta_{AB} = \sum_{i=1}^{n} |\sqrt{f_i(A)} - \sqrt{f_i(B)}|$$

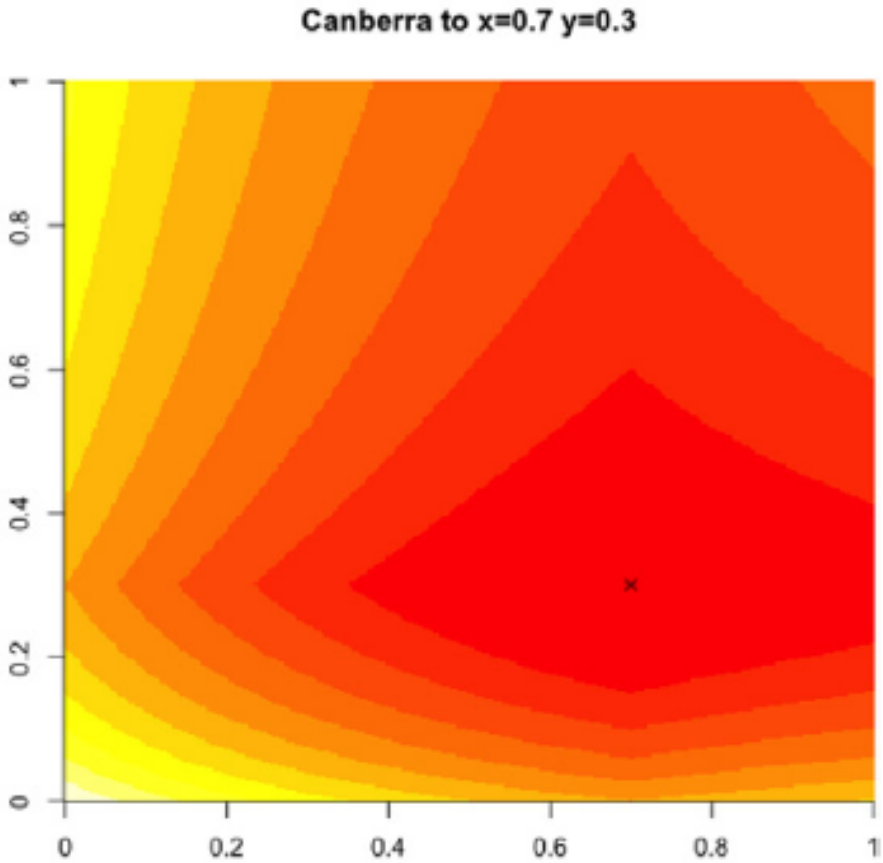
Eder's simple (monarchia mixta? politeia? res publica?)



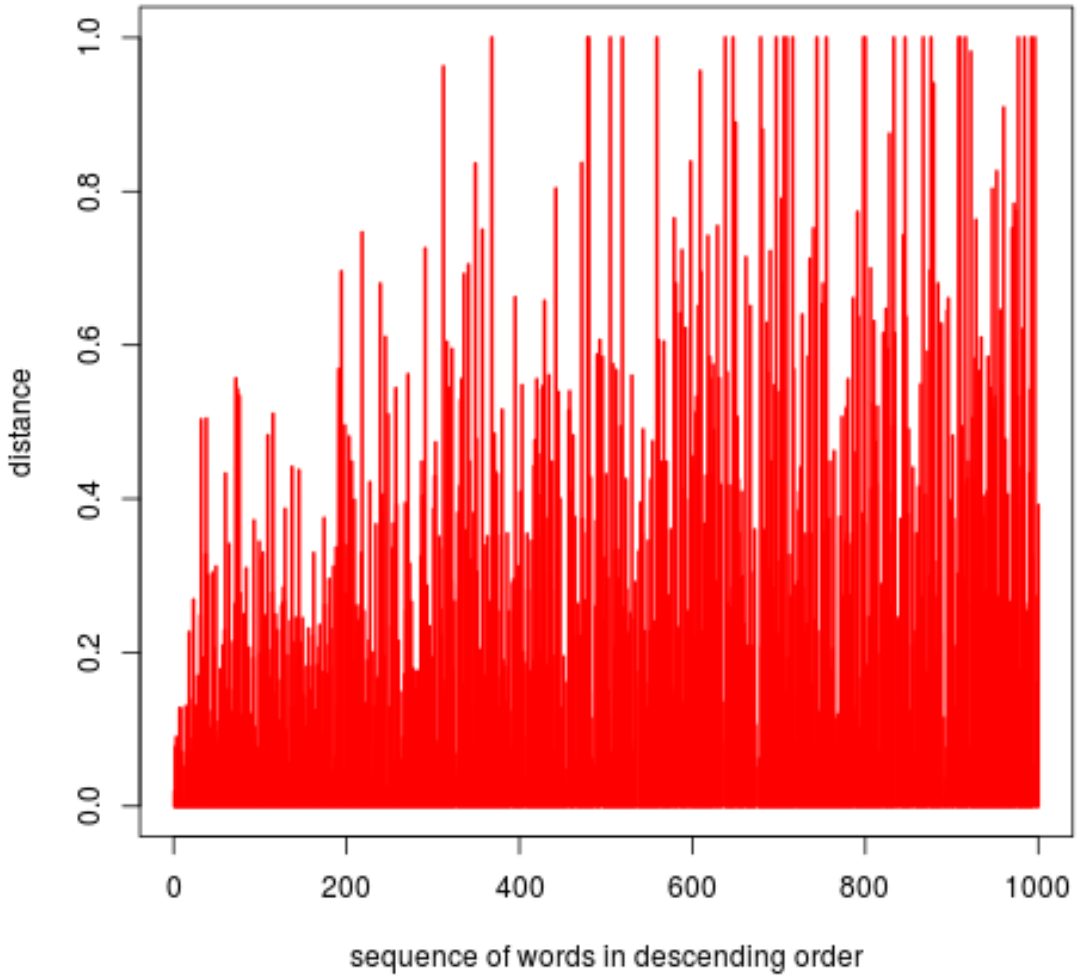
Canberra

Another classic method:

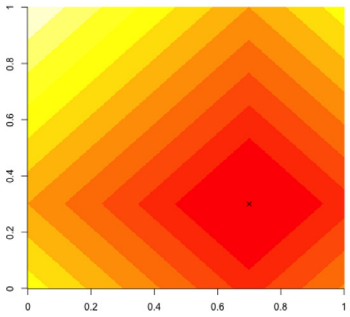
$$\delta_{AB} = \sum_{i=1}^{n} \frac{|A_i| - |B_i|}{|A_i + B_i|}$$



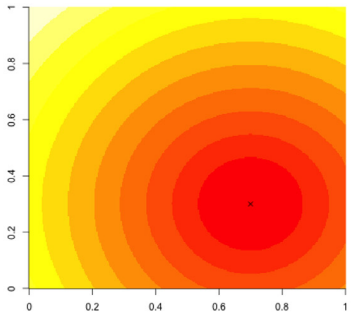
Canberra: People's revolution?



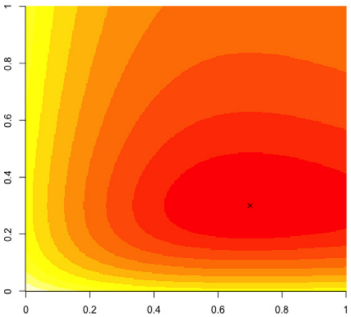
Manhattan to x=0.7 y=0.3



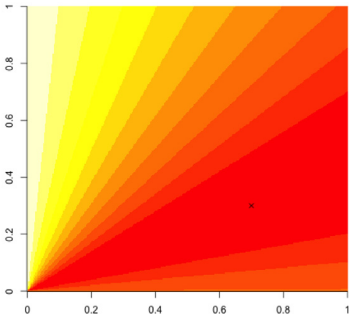
Euclidean to x=0.7 y=0.3



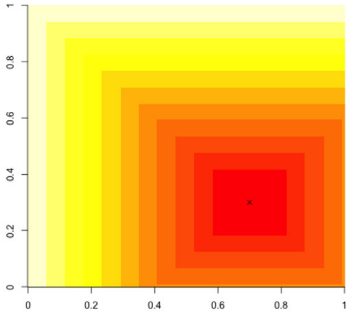
L2-Clark to x=0.7 y=0.3



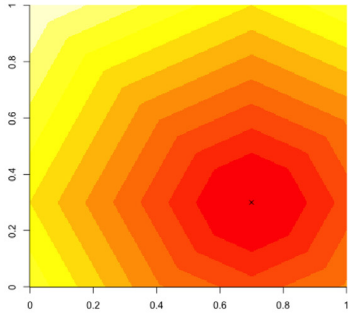
Cosine to x=0.7 y=0.3



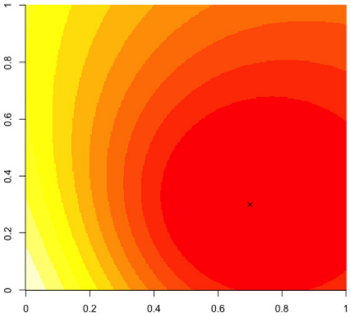
Chebyshev to x=0.7 y=0.3



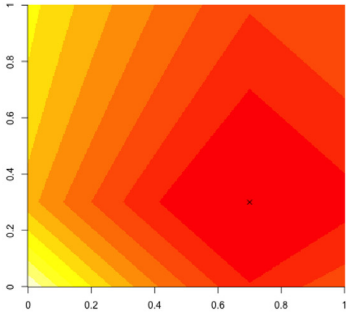
Average to x=0.7 y=0.3



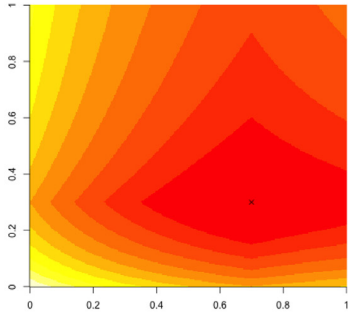
Jaccard to x=0.7 y=0.3



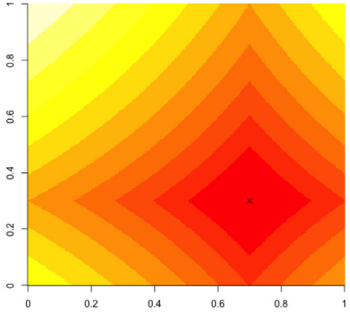
Sorensen to x=0.7 y=0.3



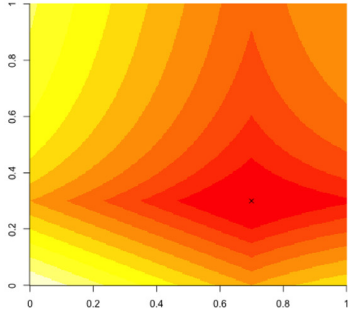
Canberra to x=0.7 y=0.3



Lorentzian to x=0.7 y=0.3



WaveHedges to x=0.7 y=0.3



Other methods implemented in stylo include

- Entropy
- Cosine Distance
- (Wurzburg) Cosine Delta
- Min-Max

  As of summer 2018, stylometric community is excited especially about Cosine Delta (Jannidis et al., 2015), as multiple tests have shown it to be most reliable and well adjusted to text analysis. You can read about it here.

We will hopefully soon include these methods in plotted comparisions. You can easily find their mathematic formulas online.

I want to know more!