More technical aspects of authorship attribution

Presenting Evert et al.'s "Better Understanding Burrow's Δ"

J.Nerbonne

Characteristics of Δ

- Overall consensus since Mosteller & Wallace to focus on frequent words
- Why?
  - Less frequent words occur too sparsely to be reliable
  - More frequent words tend to be used automatically
- But how many most frequent words (MFW's)can be used?
  - 100?, 500?, 1,000?
  - … to be considered later

Basic components

- Given a document D, we count how often each word (feature) occurs in it. We might compare $f_{the}(D_1)$ to $f_{the}(D_3)$
  - But if D_1 is bigger than D_3, we expect higher raw frequencies in D_1
  - So $f_i(D_1)$ is always a relative frequency, e.g. occurrences/1000 wd
- Burrows (see interview with John Burrows) suggested comparing not raw frequencies but rather RELATIVE FREQUENCIES
- Burrows further asked whether more frequent words should count more heavily.
  - Should they?

Digression: standard deviations, SD's

- sd measures fluctuation in variable x or scores $x_1, x_2, \ldots, x_n$
- Given mean m_x, sd is roughly the root mean of the squared differences

$$sd(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2}$$

  - x=c(1,2,3,4,5)
  - sd(x)
    - [1] 1.581139

- Always positive

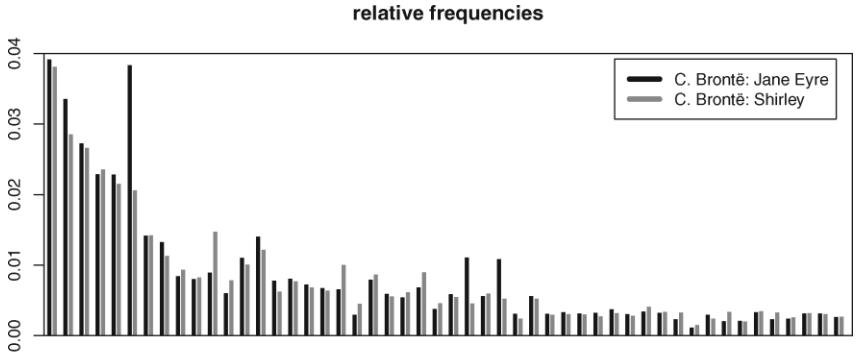
Frequency as a weight would distort

- We use lots of words automatically, prep., art., modals, …
  - Using relative frequency directly, *the* counts 100 × more than the 100th MWF, maybe *each,* and 1000 × more than …
- How to correct for this?
  - NORMALIZE frequencies
  - For each word, e.g. *the*, calculate its mean relative frequency in the collection of documents, m_{the}, and its standard deviation sd_{the}
  - Then use the NORMALIZED z-SCORE (wrt to document D):

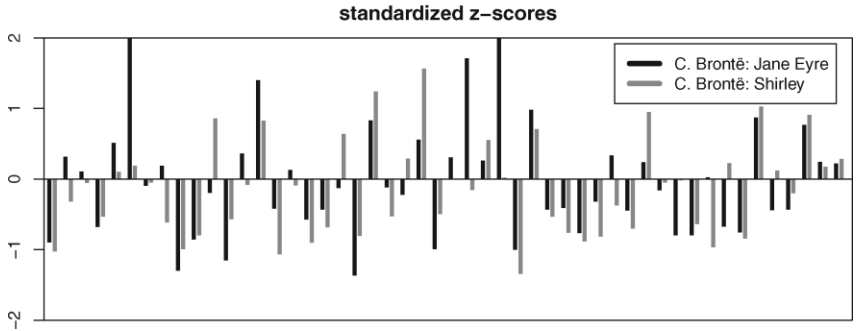
$$z_{the}(D) = \frac{f_{the}(D) - m_{the}}{sd_{the}}$$

Delta feature vectors. x-axis: 50 most frequent words, sorted by corpus frequency.

Relative frequencies



z-scores



Test later whether this gets rid of frequency effect!

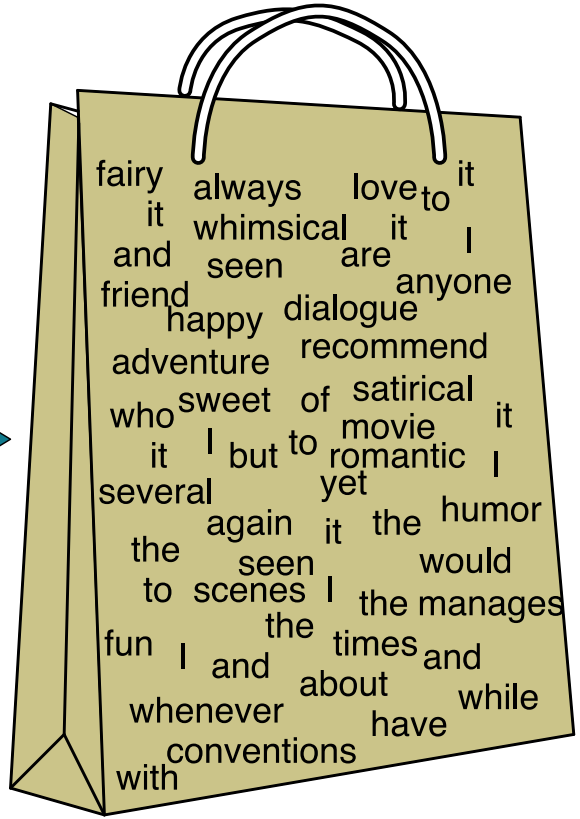
Evert et al. *Dig. Schol. Hum*. 2017

Collect all the normalized relative frequencies

- For each document being compared, collect the normalized relative frequencies for each of the n MFWs
  - Doc1 $D_1 =< z_{f_1\,(the)}, z_{f_1(be)}, \dots, z_{f_1(us)} >$
  - Doc2 $D_2 =< z_{f_2\,(the)}, z_{f_2(be)}, \dots, z_{f_2(us)} >$
- Documents compared on the basis of these n z-scores
  - So authorship attribution is a BAG-OF-WORDS model
- These are vectors in a 100-dimensional space (for 100 MFW)
- How to compare vectors?

The Bag of Words (BoW) Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Criticisms of BoW

- Words in text influence one another
  - Even in MFWs *across from, all but, every other, …*
  - Most methods assume stat. independence!
  - Argamom's (2008) ROTATED DELTA Δ_R removes independence assumption
    - But doesn't perform well

- Hard to interpret
  - If authors differ in their n MFWs, what does this mean wrt literary style?

Comparing documents

- Collect the normalized relative frequencies for the n MFWs
  - Doc1 $D_1 =< z_{f_1\,(the)}, z_{f_1(be)}, \dots, z_{f_1(us)} >$
  - Doc2 $D_2 =< z_{f_2\,(the)}, z_{f_2(be)}, \dots, z_{f_2(us)} >$
- Documents are compared on the basis of these n z-scores

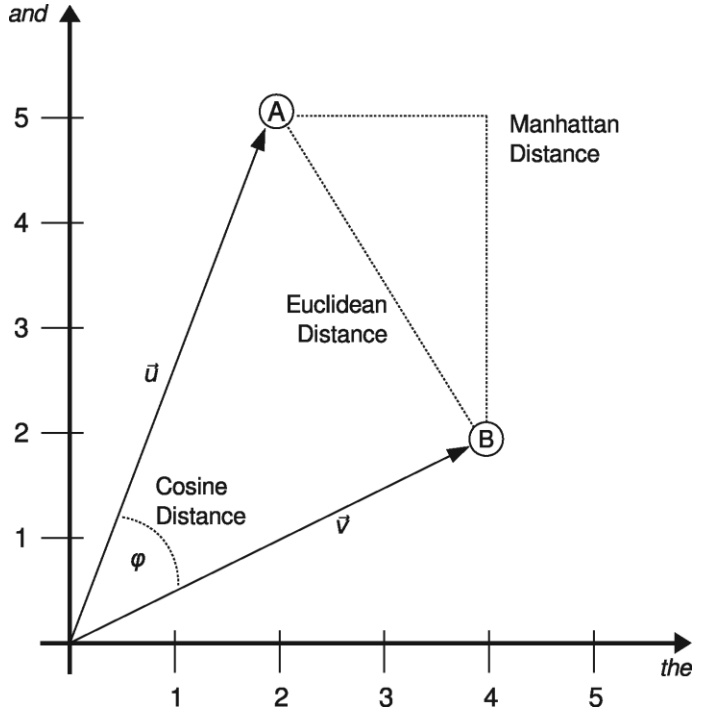
- These are vectors in a 100-dimensional space (for 100 MFW)
- Argamon (2008) asked then, how should we best compare these vectors?

Comparing vectors

- Manhattan distance – sum of component distances
  - $\Delta_B(D, D') = \sum_{i=1}^{n} |z_i(D) - z_i(D')|$
- Euclidean distance (geometry) – sq. root of squared distances
  - $\Delta_Q(D, D') = \sqrt[2]{\sum_{i=1}^{n} (z_i(D) - z_i(D'))^2}$
- Cosine – reflecting the angle between the vectors
  - $\Delta_<(D, D') = \frac{D \cdot D'}{|D| \cdot |D'|}$, i.e., the dot product between vectors D, D', normalized by the product of their lengths

Vector distances between example documents A and B illustrated in 2-dim.

- Manhattan ‘city block’ distance
  - Sum of distances in all dimensions
- Euclidean distance – length of connecting line
- Cosine – angle between vectors



Digital Scholarship Humanities (32) ,2017, Pages ii4–ii16, https://doi.org/10.1093/llc/fqx023

Systematic comparison

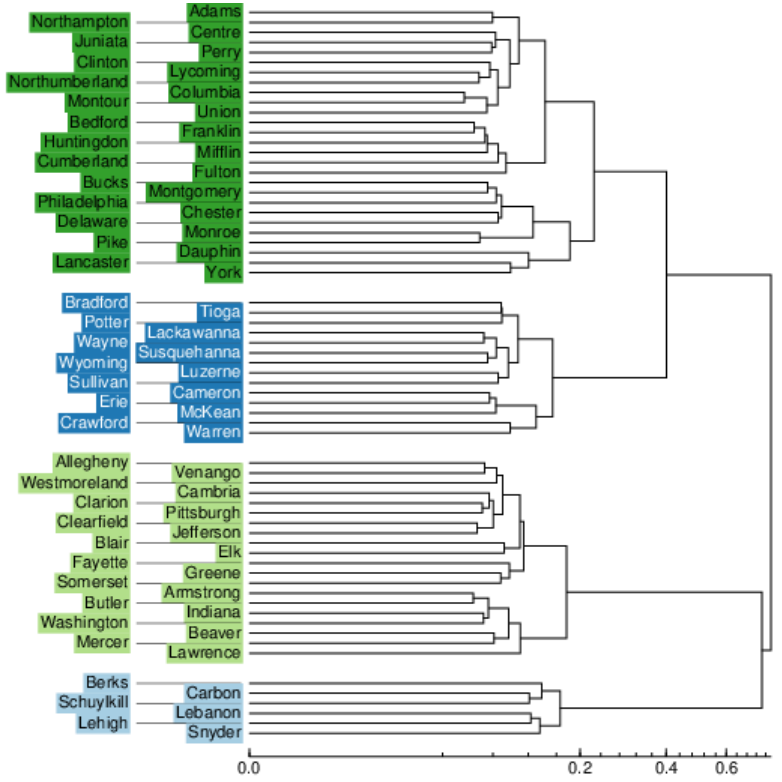
- Jannidis et al. (2015) tested different Δ's, numbers of MFWs
  - English, French & German novels
  - 75 novels from each language
  - 3 novels from each of 25 authors
- Evaluated via
  i. Differences in distances among (a) novels by same author vs. (b) novels by different authors
  ii. Clustering, i.e., how often authors' novels group together) using hierarchical clustering and partition around medoids

Digression on clustering – finding groups

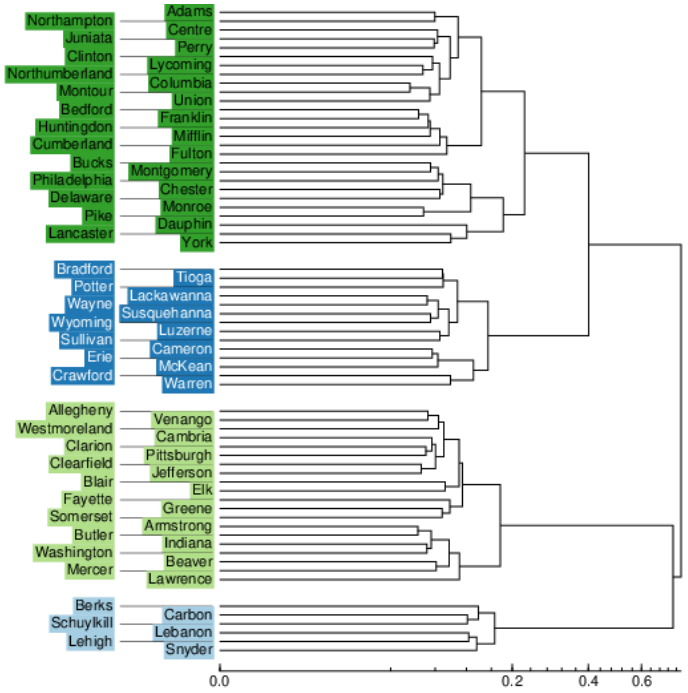
- Clustering seeks similar groups
  - Technique in exploratory statistics (not hypothesis testing, not inferential)
  - Used in various DH research lines
  - Similarity (in items and clusters) can be measured via Euclidean distance, Manhattan distance, …
- Flat vs. hierarchical clustering
  - Flat clustering partitions data in (k) groups – no subgroups
- Hard vs. soft clustering
  - Hard: Each item belongs exactly one (immediate) supergroup
  - Soft: Items may belong to more than one supergroup

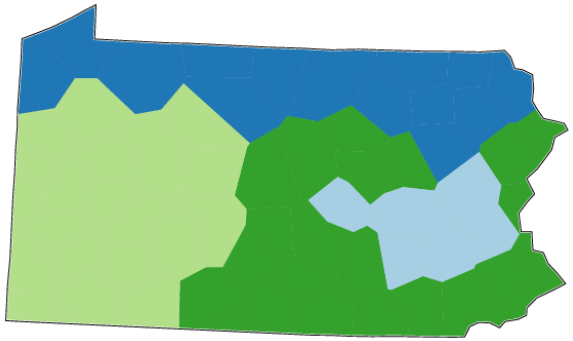
Clustering attractive for dialectology

- Whenever we suspect the existence of groups (regions)
  - Given traditional emphasis on dialect regions, clustering is a godsend!
- The results are presented in DENDROGRAMS (right)
- The branch length to the point of fusion indicates how different the fused varieties are
- One test of its correctness is the projection to geography



Projecting clusters to a map



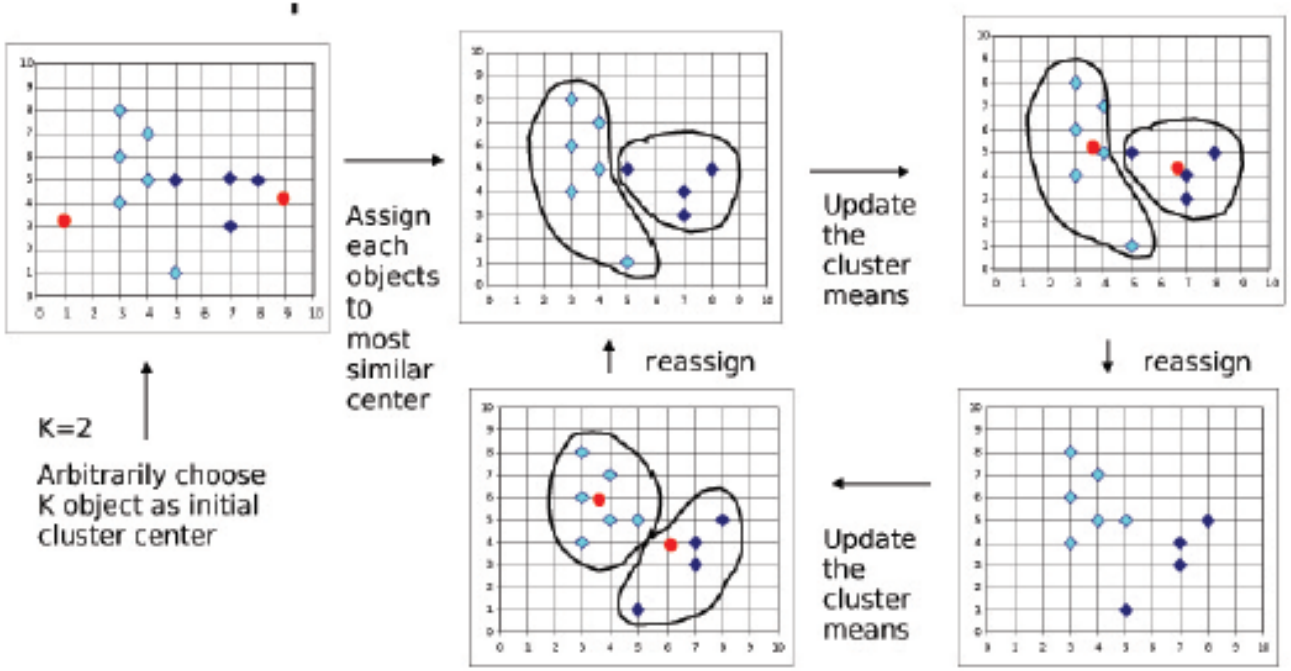


- The clusters don't have to be geographically coherent
- Usually a good sign if they are!
- From Gabmap: gabmap.nl

K-Means (or k–medoids) clustering

- Very popular in computational linguistics!
  - That's why we discuss it here
- Hard clustering algorithm
- Starts by partitioning the input points into k initial sets (randomly)
- Calculates the mean point, or centroid, of each set
  - Note the need for a distance measure
- Constructs a new partition by associating each point with the closest centroid
- Repeats last two steps until the objects no longer switch clusters

K-Means (Sketch by Jelena Prokić)



K-means clustering in dialectology

- Not often used
  - Why not?
- Dialectologists find hierarchical structure in the data
  - Feldkirch is southern Baden, which is Baden, which is Alemannic, which is continental west Germanic, which is ….
- It is seen occasionally in dialectology (when there's little interest in finer relations)
  - I don't recommend it for dialectology!
- Discussed here for its use elsewhere in DH, e.g., authorship attribution

The adjusted Rand Index evaluates classifications

- Example

imperfect grouping

	1(2)	2(35)	3(3)
$A(22)$	2(0.09)	17(0.43)	3(0.14)
$B(13)$	0	13(0.37)	0
$C(5)$	0	5(0.14)	0
Total	0.09	0.94	0.14

score $= 1.17/3$

perfect grouping

	1(2)	2(35)	3(3)
$A(22)$	0	22(1)	0
$B(13)$	13(1)	0	0
$C(5)$	0	0	5(1)
Total	1	1	1

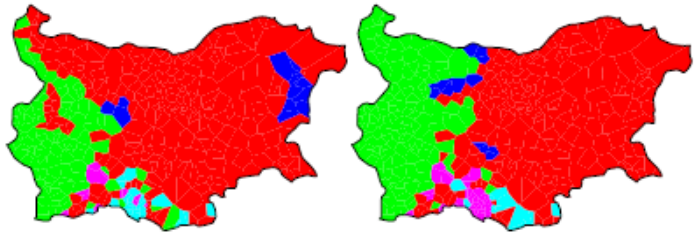
score $= 3/3$

- Evert et al. compare low-level clusters to novels by the same author.
- Less sensitive than comparing the vector distances, but also more task-specific and easier to interpret.

- Heeringa, W., Nerbonne, J., & Kleiweg, P. (2002). Validating dialect comparison methods. In *Classification, Automation, and New Media,* 445-452. Springer: Berlin.

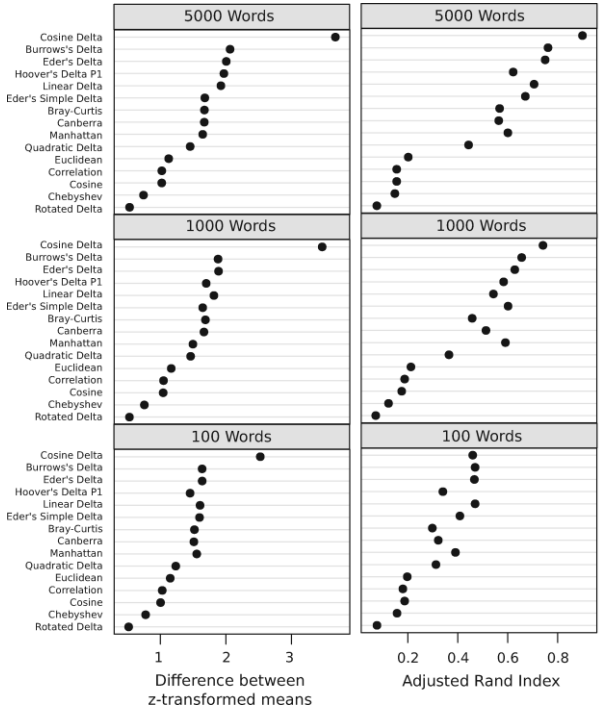
Quality of Clustering

- There's no perfect clustering algorithm
  - Kleinberg, Jon M. 2004. "An impossibility theorem for clustering" In: S. Thrun, S.Lawrence, & B. Schölkopf (eds.), *Advances in Neural Information Processing Systems* 16: Proc. NIPS 16. Cambridge, MA: MIT Press.
- Clustering has a serious stability problem
  - A process is STABLE if small changes in input change the results only a little
  - Two Bulgarian datasets (*r=0.97*)
  - Clustered, projected to map
- Cophenetic correlation
- Take care!



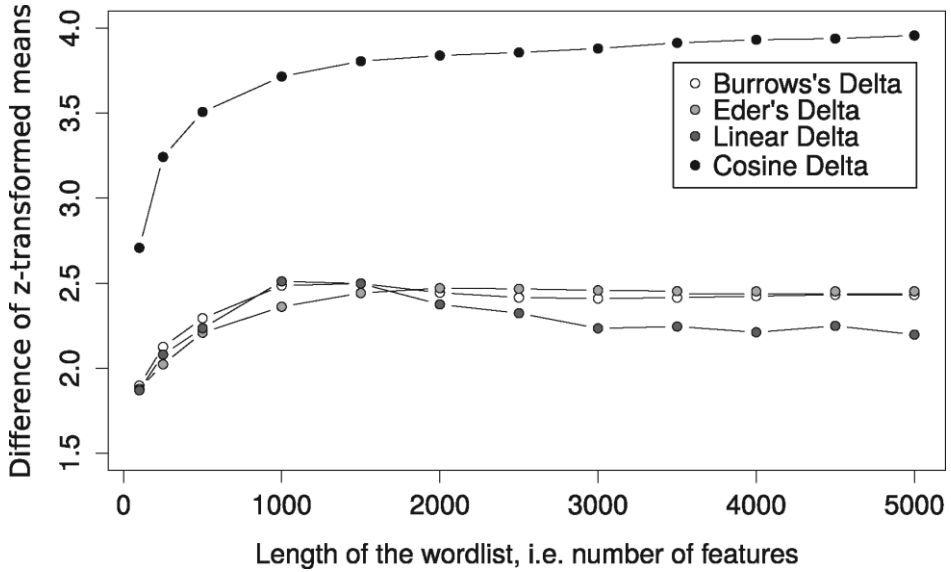
Performance of Δ's on English texts

- Based on z-score differences between books by the same author vs. books by different authors (left column)
- Based on classification success in clustering (right column) using adj. Rand index



Digital Scholarship Humanities, 32, 2017, Pages ii4–ii16, https://doi.org/10.1093/llc/fqx023

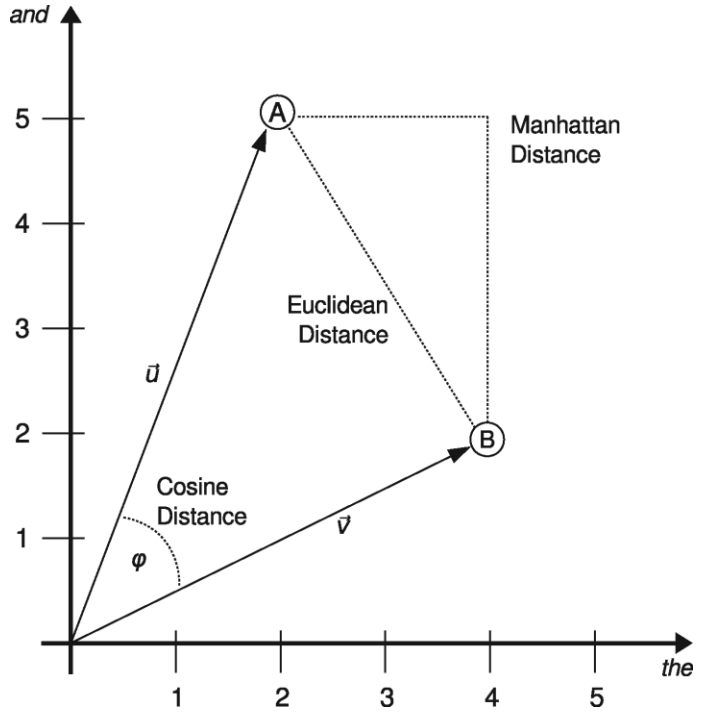
Difference between z-transformed means of ingroup and outgroup distances as a function of nMFW. Indicated for German texts.



Digital Scholarship Humanities, Volume 32, Issue suppl_2, December 2017, Pages ii4–ii16, https://doi.org/10.1093/llc/fqx023

Why is cosine delta $(\Delta_<)$ so much better?

- Recall scheme
- Consider what happens when vectors get larger
  - Euclidean & Manhattan distances also get larger
  - But cosine is constant!
- Normalization should reduce differences!



Other indications

- z-scores used to reduce influence of frequency (among MFWs)
  - Successful?
- Less freq. words contribute less to (Δ_B) , even less to (Δ_Q)
  - Why?
  - For most z-scores, $|z^2| \leq |z|$
- Normalize!

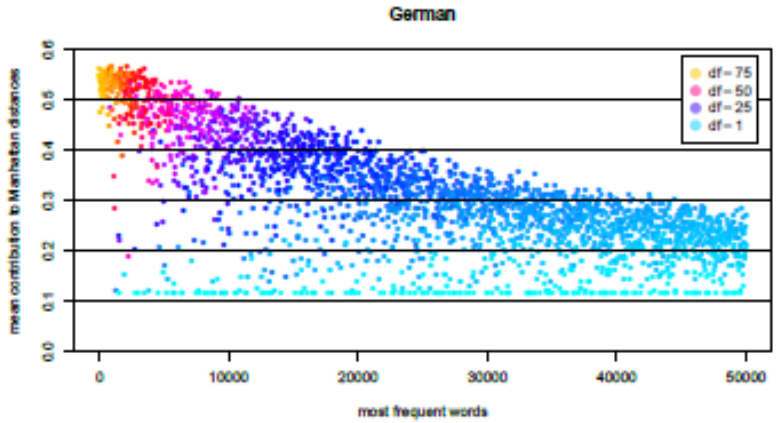


Figure 5: Average contribution of features to pairwise Δ_B distances; colour indicates document frequency (df)

Evert et al. NAACL-HLT Workshop 2015

Vector normalization

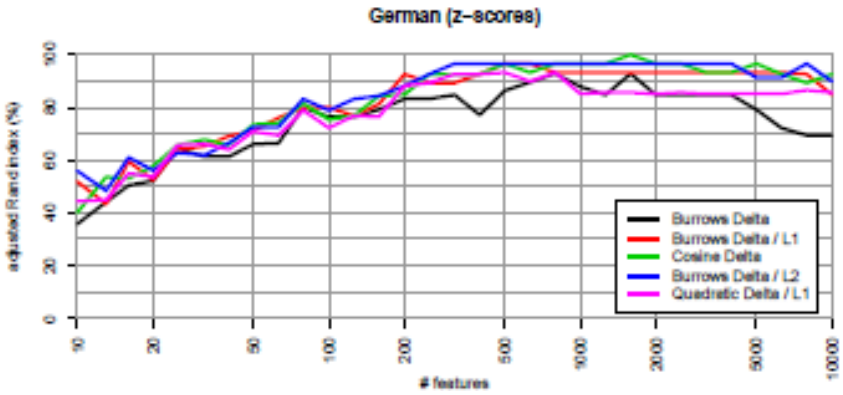
- When a vector is normalized, its length is set to one
  - Of course, how the length $\|\vec{v}\|$ is determined is important
  - Euclidean (L_2) or Manhattan (L_1)

- Let $\vec{v} = \langle 1,2,3 \rangle$, then $\|\vec{v}\|_2 = \sqrt[2]{1^2 + 2^2 + 3^2} = \sqrt[2]{14} \approx 3.75$
  - $L_2(\vec{v}) = \left\langle \frac{1}{3.75}, \frac{2}{3.75}, \frac{3}{3.75} \right\rangle = \langle 0.27, 0.53, 0.8 \rangle$
  - $\|L_2(\vec{v})\|_2 = \sqrt[2]{0.27^2 + 0.53^2 + 0.8^2} = \sqrt[2]{0.07 + 0.28 + 0.64} \approx 1.0$

- Manhattan distance normalized in the same way

Effects of normalization

- Hypothesis: normalization reduces effect of low frequency
- Tested on cluster quality
  - L_1, L_2 norms as above
- Confirming evidence!
- But what does this mean about authors' profiles?



Evert et al. NAACL-HLT, '15

Normalization & author profiles

- Why does normalization help (so much)?
  - Outlier elimination: Normalization improves detection because it suppresses the influence of outliers.
  - Key profile: Key to detection isn't the size of differences, but the pattern of overuse and underuse
- Evert et al. perform two more experiments to test these hypo's
  - "clamp" z-scores to $-1 \leq z \leq 1$ Scores ≤ -1 count -1, scores ≥ 1, as 1
    - This eliminates the effect of outliers!
  - Discretizing z-scores to three values, low-mid-high, eliminating magnitude

Outlier elimination or key profile

- Evidence is found for both hypotheses!
- Evert et al. experiment further with a Minkowski norm

$$\Delta_p(D, D') = \sqrt[p]{\sum_{i=1}^{n} |z_i(D) - z_i(D')|^p}$$, for various values of p

  - Generalization of Euclidean distance
- Possible presentation topic for the mathematically inclined

Current wisdom on Δ

- Focus on MFW, to guarantee sufficient sample size, and to focus on words used automatically (Mosteller & Wallace 1964)
- Use relative frequencies (wrt corpus frequency) to discount document size (Burrows 2002)
- Use normalized (z-score) frequencies to avoid (automatic) frequency weighting, again following Burrows
- Use cosine delta ($\Delta_<$) to measure vector distance OR first normalize vectors, using then $\Delta_<$, Δ_B (Manhattan), or Δ_Q(Euclidean) (Argamon 2008; Evert et al. 2015/17)

Next times

- Rolling stylometry
- Typical features (profiling)