Towards a better understanding of Burrows's Delta for literary authorship attribution

CLfL - June 4, 2015, Denver

Notation

- Text documents D in a collection $\mathcal{D}$ of size $n_{\mathcal{D}}$
- Each text D is represented by a profile of relative frequencies $f_i(D)$ of the n_w most frequent words $w_1, w_2, \dots\ w_{n_w}$
- The complete profile of D is given by the feature vector $\mathbf{f}(D) = (f_1(D), \dots, f_{n_w}(D))$
- Features are standardized using a z-transformation $z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i}$
- Dissimilarities between the scaled feature vectors are computed according to a distance metric

Delta Measures

- Burrows's Delta [1]: Manhattan Distance

$$\Delta_B(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_1 = \sum_{i=1}^{n_w} |z_i(D) - z_i(D')|$$

- Quadratic Delta [2]: squared Euclidean Distance

$$\Delta_Q(D, D') = \|\mathbf{z}(D) - \mathbf{z}(D')\|_2^2 = \sum_{i=1}^{n_w} (z_i(D) - z_i(D'))^2$$

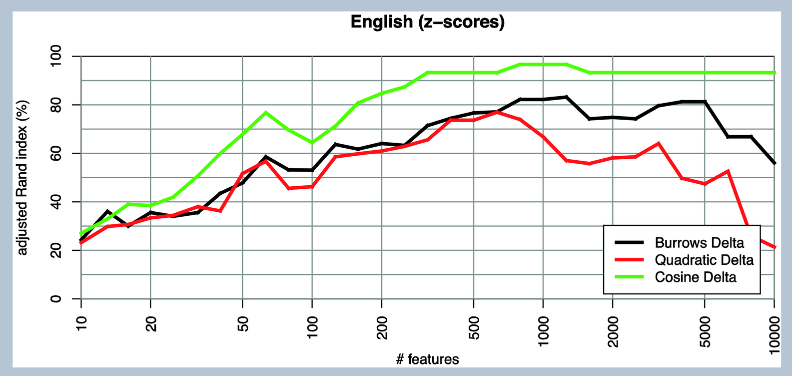
- Cosine Delta [3]: angle α between two feature vectors, computed from cosine similarity of $\mathbf{x} = \mathbf{z}(D)$ and $\mathbf{y} = \mathbf{z}(D')$

$$\Delta_\angle(D, D') = \alpha, \text{ with } \cos\alpha = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2}$$

Understanding the parameters of Delta

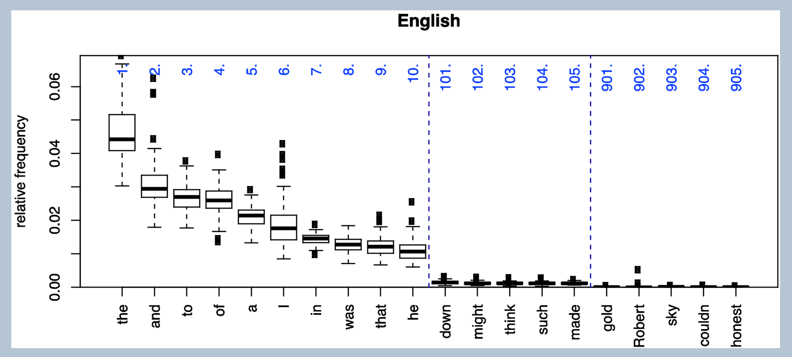
MFW

- Same clustering quality for Δ_B and Δ_Q for $n_w \leq 500$, but Δ_B proves to be more robust if n_w is increased, cf. [4]
- $\Delta_\angle$ outperforms the other variants, is robust, degrades more slowly and achieves impressive clustering quality
- Optimal n_w depends on many factors (language, text type, text length,…) and cannot be known a priori



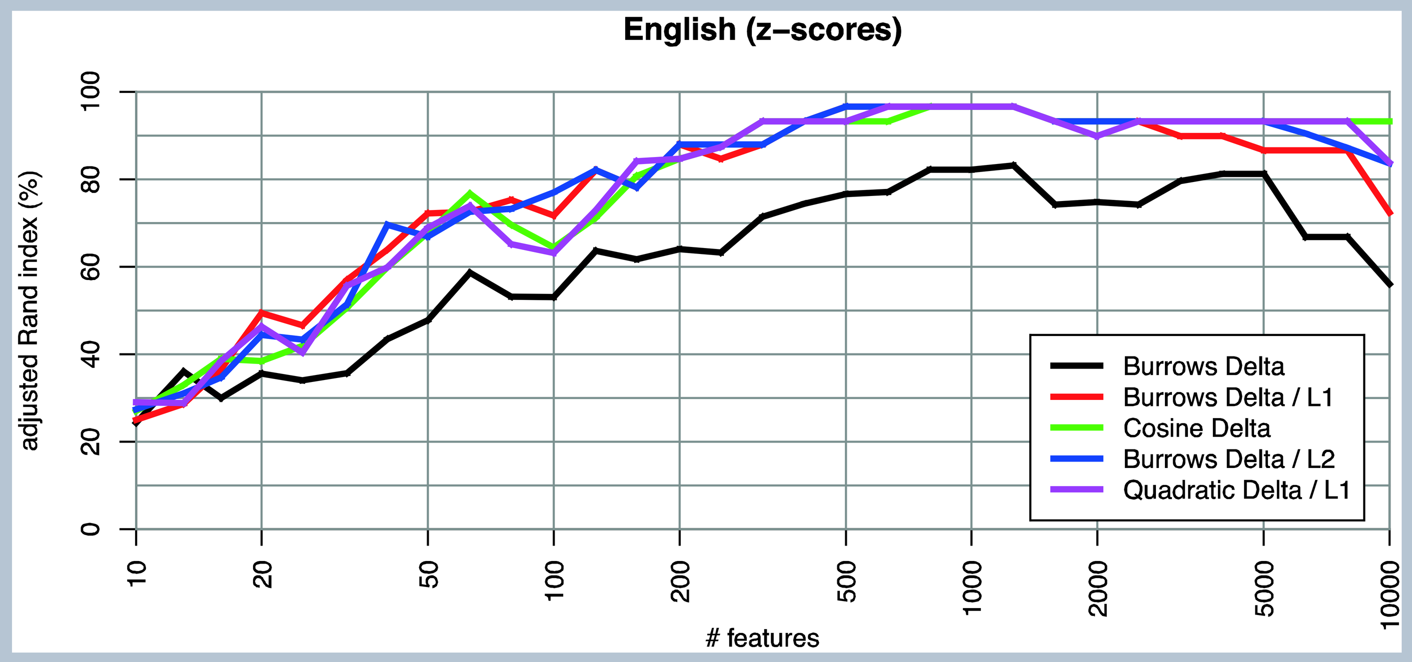
Feature scaling

- Without standardization, words with mfw ranks above 100 hardly make any contribution to the frequency profiles and hardly affect the delta scores
- Standardization gives all features equal weight in Δ_Q and $\Delta_\angle$
- In Δ_B, standardization gives less frequent words a moderately smaller weight; it also reduces the weight of words concentrated in a small number of texts. Experiments show that this results in better clustering quality than a scaling that gives equal weight to all features.



Vector normalization

- Normalization is the main difference between Δ_Q and $\Delta_\angle$, might also improve other measures
- Δ_Q and Δ_B are substantially improved by vector normalization, regardless of the type of normalization (L_1 vs. L_2)
- Authorial style reflected by positive and negative deviations of word frequencies from the average frequency across the collection
- Not to the same degree in all texts of one author, therefore differences in length (i.e. norm) of feature vectors
- Normalization makes the author's stylistic pattern stand out more clearly



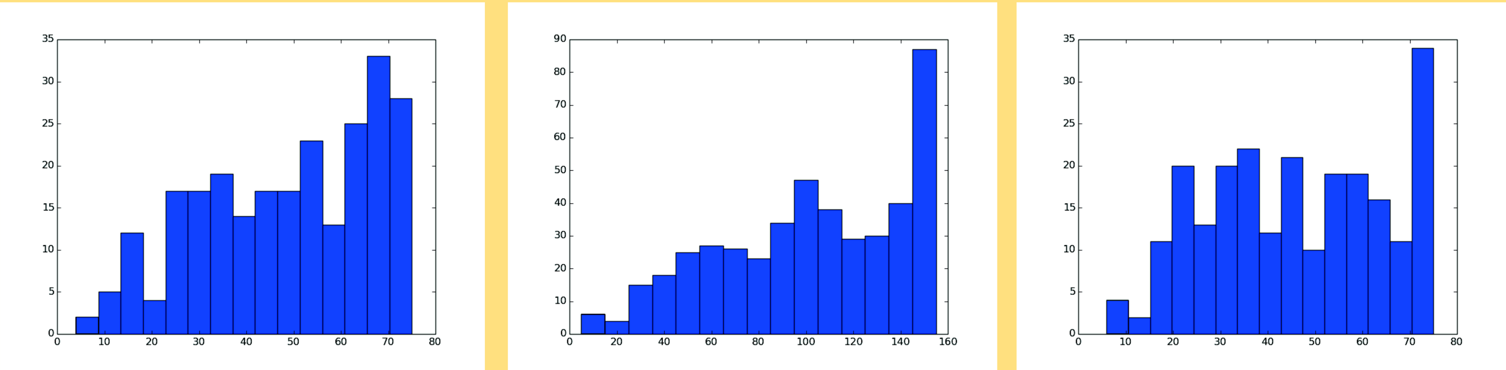
Recursive feature elimination

- Greedy algorithm which relies on a ranking of features and on each step selects only the top features, removing the remaining ones
- Reduction to 50000 features in steps of 10000, to 5000 in steps of 1000 and finally to 500 in steps of 100 features
- Find the optimal number of features by pruning one feature at a time with stratified threefold cross-validation after each step
- Both classification and clustering with $\Delta_\angle$ with optimal feature subset yield perfect results

	English	French	German
nr. of features	246	381	234
SVC accuracy	0.99 (±0.04)	1.00 (±0.00)	1.00 (±0.00)
MaxEnt accuracy	1.00 (±0.00)	1.00 (±0.00)	1.00 (±0.00)
Cosine Delta ARI	0.966	1.000	1.000

The selected feature subset

- Some features highly specific, occurring only in a fraction of texts, but most selected features have a rather high document frequency
- Not limited to function words
- Roman numerals in French and English collection characteristic of novels with unusually many chapters
- Artifacts in German collection due to historic orthographic variants



(a) English (b) French (c) German

Possible overfitting?

- Two additional unseen evaluation data sets, the second mainly consisting of additional authors
- Classification accuracy of 0.97 on first test set indicates good generalization to unseen works from the same authors
- Classification and clustering with $\Delta_\angle$ on the set with new authors and no singletons also yield good results
- Higher ARI for selected features than for 2000 mfw indicates that features are not overfitted and generalize well to unknown authors
- Difference in accuracy between the first and second test set indicates that features are somewhat author-dependent

	unscaled full fs	rescaled full fs	selected fs
SVC accuracy	0.91 (±0.03)	0.57 (±0.13)	0.84 (±0.14)
MaxEnt accuracy	0.95 (±0.03)	0.95 (±0.03)	0.90 (±0.08)
Cosine Delta ARI	0.835	0.835	0.871

References

[1] John Burrows. 2002. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267-287.

[2] Shlomo Argamon. 2008. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, 23(2): 131, 147, June.

[3] Peter W. H. Smith and W. Aldridge. 2011. Improving Authorship Attribution: Optimizing Burrows's Delta Method. *Journal of Quantitative Linguistics*, 18(1):63-88, February.

[4] Fotis Jannidis, Steffen Pielström, Christof Schöch and Thorsten Vitt. 2015. Improving Burrows's Delta – An empirical evaluation of text distance measures. In *Digital Humanities Conference 2015*, Sydney.

FAU Erlangen-Nürnberg

Stefan Evert, Thomas Proisl; {stefan.evert, thomas.proisl}@fau.de

University of Würzburg

Fotis Jannidis, Steffen Pielström, Christof Schöch, Thorsten Vitt

fotis.jannidis@uni-wuerzburg.de

poster design: Isabella Reger, isabella.reger@uni-wuerzburg.de