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

Delta Measures
- Burrows's Delta [1]: Manhattan Distance $\Delta_\Delta(D, D') = |\mu(D) - \mu(D')|$
- Quadratic Delta [2]: squared Euclidean Distance $\Delta_\triangle(D, D') = \sum (x_i(D) - x_i(D'))^2$
- Cosine Delta [3]: angle $\alpha$ between two feature vectors, computed from cosine similarity of $x = x(D)$ and $y = x(D')$

MFW
- Same clustering quality for $\Delta_\Delta$ and $\Delta_\triangle$ for $n_w \leq 500$, but $\Delta_\Delta$ proves to be more robust if $n_w$ is increased, cf. [4]
- $\Delta_\triangle$ 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

Understanding the parameters of Delta

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
- Find the optimal number of features by pruning one feature at a time
- Classification and clustering with $\Delta_\triangle$ yields perfect results
- Classification accuracy of 0.97 on first test set indicates good generalization to unseen works from the same authors
- Classification and clustering with $\Delta_\Delta$ 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 somewhat author-dependent

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

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 $\Delta_\Delta$ 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

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 $\Delta_\Delta$ and $\Delta_\triangle$
- In $\Delta_\triangle$, 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 $\Delta_\Delta$ and $\Delta_\triangle$, might also improve other measures
- $\Delta_\triangle$ and $\Delta_\Delta$ 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

Table: The selected feature subset

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>nr. of features</td>
<td>246</td>
<td>381</td>
<td>234</td>
</tr>
<tr>
<td>SVC accuracy</td>
<td>0.99 (±0.04)</td>
<td>1.00 (±0.00)</td>
<td>1.00 (±0.00)</td>
</tr>
<tr>
<td>MaxEnt accuracy</td>
<td>1.00 (±0.00)</td>
<td>1.00 (±0.00)</td>
<td>1.00 (±0.00)</td>
</tr>
<tr>
<td>Cosine Delta ARI</td>
<td>0.966</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

References