Authorship Verification...

Is this an instance of the class SAME-AUTHOR or DIFFERENT-AUTHOR?

---

What characterizes authors?

- Unusual word choice?
- Shorter sentences?
- More complex grammar?

Sidenote: Individual vs Joint

- Individual vs Joint features valid ways of classifying texts written by the same or different authors?

Sidenote: visual features

- ... incl. punctuation, letter case, line length... to detect layout differences in plays / poetry

All features and combinations of features

Features' performance

<table>
<thead>
<tr>
<th>Language</th>
<th>Helpful</th>
<th>Harmful</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>All Individual</td>
<td>All Joint</td>
</tr>
<tr>
<td>Greek</td>
<td>All Joint</td>
<td>Just Entropy</td>
</tr>
<tr>
<td>Spanish</td>
<td>All Joint</td>
<td>Just Ngram sim</td>
</tr>
<tr>
<td>English</td>
<td>Vis+ngram+tok</td>
<td>Just POS sim</td>
</tr>
</tbody>
</table>

Best feature combinations

- Combo 1: ngram overlap + visual features
- Combo 2: ngram overlap + token cosine similarity
- Combo 3: ngram overlap + joint features
- Combo 4: ngram overlap + POS sim

Results: combined scores on PAN 2015 data

<table>
<thead>
<tr>
<th>Language</th>
<th>Training</th>
<th>Test</th>
<th>Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (full set)</td>
<td>.55</td>
<td>.62</td>
<td>3/17</td>
</tr>
<tr>
<td>English (full set)</td>
<td>.56</td>
<td>.41</td>
<td>8/17</td>
</tr>
<tr>
<td>Greek (combo2)</td>
<td>.54</td>
<td>.60</td>
<td>4/15</td>
</tr>
<tr>
<td>Spanish (full set)</td>
<td>.90</td>
<td>.54</td>
<td>5/17</td>
</tr>
</tbody>
</table>

Average runtime: 1 minute

Discussion

- Why do complex features not outperform surface-level features?
- Are individual vs joint features valid ways of classifying texts written by the same of different authors?

Conclusion: Pros of the system

- Robust
  - ... works across genres
  - ... works across topics
  - ... works across languages

- Simple
  - ... easy to run
  - ... uses simple features
  - ... very fast runtime