INTRINSIC PLAGIARISM DETECTION USING CHARACTER TRIGRAM DISTANCE SCORES

UNDER A NOVEL DOCUMENT REPRESENTATION
PLAGIARISM DETECTION

- External detection:
  - reference corpus = ALL source documents
  - ‘Closed’ world
- **Realistic?**
  - Growing potential reference collection (cf. web)
    - Computationally complex!
  - Not all sources digitally/publicly available
  - E.g. student hiring ghost writer for sections in master thesis: what if ghost writer himself did not plagiarize?
- Practically **relevant**
• Limited resources
• Only document itself...
• Seminal work: standard methodology

“The underlying approach to intrinsic plagiarism detection has not changed: a suspicious document $d$ is chunked, and [...] each chunk is compared with the whole of $d$. Then, chunks whose writing style differs significantly from the average writing style of the document are identified using outlier detection.” (PAN overview 2010)

• (Negative undertone?)
Segments, chunks, windows, …

Suspicious document

- Window size
  - Step size
  - \( W_1 \)
  - \( W_2 \)
  - \( W_3 \)
$D$ vs. $w_1, w_2, w_3, \ldots, w_n$

Entire suspicious document $D$

$\Delta(D, w_i)$

$W_1$  $W_2$  $W_3$  $W_4$
IMPLICIT ASSUMPTIONS?

1 – “It’s okay to compare a chunk to the document as a whole.”

2 – “The whole document is a reliable point of stylistic reference.”
COMMON PRACTICE?

Equal size

Different size
1 – “It’s okay to compare a chunk to the document as a whole.”

2 – “The whole document is a reliable point of stylistic reference.”
WORST-CASE SCENARIOS

Original text will be marked as plagiarized?

Which one is the original author?
QUESTIONABLE ASSUMPTIONS

1. “It’s ok to compare a chunk to the document as a whole”
2. “Whole document is reliable point of stylistic reference”

But is there an alternative?
WINDOW VS. WINDOW

• Instead of Document vs. Window...
• Window versus Window
  • No assumption of reliability of D as a whole
  • Comparing blocks of equal size
## SYMMETRICAL DISTANCE MATRIX

Cf. Distance tables for clustering

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>...</th>
<th>$w_{n-1}$</th>
<th>$w_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>0</td>
<td>$\Delta(w_1,w_2)$</td>
<td>...</td>
<td>$\Delta(w_1,w_{n-1})$</td>
<td>$\Delta(w_1,w_n)$</td>
</tr>
<tr>
<td>$w_2$</td>
<td>$\Delta(w_1,w_2)$</td>
<td>0</td>
<td>...</td>
<td>$\Delta(w_2,w_{n-1})$</td>
<td>$\Delta(w_2,w_n)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$w_{n-1}$</td>
<td>$\Delta(w_{n-1},w_1)$</td>
<td>$\Delta(w_{n-1},w_2)$</td>
<td>...</td>
<td>0</td>
<td>$\Delta(w_{n-1},w_n)$</td>
</tr>
<tr>
<td>$w_n$</td>
<td>$\Delta(w_n,w_1)$</td>
<td>$\Delta(w_n,w_2)$</td>
<td>...</td>
<td>$\Delta(w_n,w_{n-1})$</td>
<td>0</td>
</tr>
</tbody>
</table>
CLUSTERING OF PLAGIARISMS OF SAME SOURCE
DISTANCE MEASURE

- Stamatatos’s normalized distance
- Distance between two ‘text profiles’
- Profile = bag-of-character-trigrams

\[
\sum_{g \in P(w_x)} \frac{\left( \frac{2(f_{w_x}(g) - f_{w_y}(g))}{f_{w_x}(g) + f_{w_y}(g)} \right)^2}{4|P(w_x)|}
\]
SYMMETRIC ADAPTATION

- Originally: all trigrams from 1 document
- Asymmetrical: distance(A,B) != distance(B,A)
- Adaptation: restrict to $n=1000$ most frequent character trigrams from entire corpus
- Stylometric inspiration
- Computationally simple: symmetry!

\[
\begin{array}{cccccc}
  & w_1 & w_2 & \ldots & w_{n-1} & w_n \\
w_1 & 0 & \Delta(w_1,w_2) & \ldots & \Delta(w_1,w_{n-1}) & \Delta(w_1,w_n) \\
w_2 & & 0 & \ldots & \Delta(w_2,w_{n-1}) & \Delta(w_2,w_n) \\
\vdots & & & \ddots & \ldots & \ddots \\
w_{n-1} & & & & 0 & \Delta(w_{n-1},w_n) \\
w_n & & & & & 0 \\
\end{array}
\]
OUTLIERS?

• Distance table (cf. clustering)
• Multivariate, higher-dimensional
• Mvoutlier (R, Filzmoser et al.)
• Principal Components Analysis
• Reduces dimensionality before detection
The smaller the windows, the better (but more expensive)

<table>
<thead>
<tr>
<th>ws</th>
<th>ss</th>
<th>plagdet recall</th>
<th>precision</th>
<th>granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>20,000</td>
<td>20,000</td>
<td>19.48</td>
<td>20.02</td>
<td>19.01</td>
</tr>
<tr>
<td>20,000</td>
<td>15,000</td>
<td>20.59</td>
<td>21.84</td>
<td>19.88</td>
</tr>
<tr>
<td>20,000</td>
<td>10,000</td>
<td>23.80</td>
<td>27.79</td>
<td>21.00</td>
</tr>
<tr>
<td>20,000</td>
<td>5,000</td>
<td>25.84</td>
<td>39.55</td>
<td>19.52</td>
</tr>
<tr>
<td>20,000</td>
<td>1,000</td>
<td>26.36</td>
<td>44.99</td>
<td>18.91</td>
</tr>
<tr>
<td>15,000</td>
<td>15,000</td>
<td>20.04</td>
<td>20.29</td>
<td>20.71</td>
</tr>
<tr>
<td>15,000</td>
<td>11,250</td>
<td>22.41</td>
<td>23.09</td>
<td>22.41</td>
</tr>
<tr>
<td>15,000</td>
<td>7,500</td>
<td>25.97</td>
<td>29.69</td>
<td>23.44</td>
</tr>
<tr>
<td>15,000</td>
<td>3750</td>
<td>26.79</td>
<td>40.17</td>
<td>20.63</td>
</tr>
<tr>
<td>15,000</td>
<td>750</td>
<td><strong>27.21</strong></td>
<td>45.09</td>
<td>19.89</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
<td>21.33</td>
<td>20.35</td>
<td>23.34</td>
</tr>
<tr>
<td>10,000</td>
<td>7,500</td>
<td>24.14</td>
<td>24.05</td>
<td>25.95</td>
</tr>
<tr>
<td>10,000</td>
<td>5,000</td>
<td>27.26</td>
<td>29.98</td>
<td>25.89</td>
</tr>
<tr>
<td>10,000</td>
<td>2,500</td>
<td><strong>27.53</strong></td>
<td>40.00</td>
<td>22.03</td>
</tr>
<tr>
<td>5,000</td>
<td>5,000</td>
<td>21.77</td>
<td>20.38</td>
<td>28.09</td>
</tr>
<tr>
<td>5,000</td>
<td>3,750</td>
<td>24.03</td>
<td>24.18</td>
<td>29.79</td>
</tr>
<tr>
<td>5,000</td>
<td>2,500</td>
<td><strong>27.52</strong></td>
<td>30.42</td>
<td>28.50</td>
</tr>
<tr>
<td>5,000</td>
<td>1,250</td>
<td><strong>27.49</strong></td>
<td>37.56</td>
<td>24.55</td>
</tr>
</tbody>
</table>
### OUTBOUND PARAMETER

<table>
<thead>
<tr>
<th>outbound</th>
<th>ws</th>
<th>ss</th>
<th>plagdet</th>
<th>recall</th>
<th>precision</th>
<th>granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>.20</td>
<td>20,000</td>
<td>20,000</td>
<td>19.92</td>
<td>21.17</td>
<td>18.84</td>
<td>1.00</td>
</tr>
<tr>
<td>.20</td>
<td>20,000</td>
<td>5,000</td>
<td>25.87</td>
<td>41.84</td>
<td>19.06</td>
<td>1.02</td>
</tr>
<tr>
<td>.30</td>
<td>20,000</td>
<td>5,000</td>
<td>25.66</td>
<td>36.60</td>
<td>20.09</td>
<td>1.01</td>
</tr>
<tr>
<td>.30</td>
<td>15,000</td>
<td>3,750</td>
<td>26.82</td>
<td>37.24</td>
<td>21.48</td>
<td>1.02</td>
</tr>
<tr>
<td>.35</td>
<td>15,000</td>
<td>3,750</td>
<td>25.68</td>
<td>30.01</td>
<td>22.91</td>
<td>1.02</td>
</tr>
<tr>
<td>.30</td>
<td>10,000</td>
<td>2,500</td>
<td>27.61</td>
<td>36.93</td>
<td>23.13</td>
<td>1.04</td>
</tr>
<tr>
<td>.20</td>
<td>10,000</td>
<td>2,500</td>
<td>27.29</td>
<td>42.25</td>
<td>21.17</td>
<td>1.04</td>
</tr>
</tbody>
</table>

- Controlled ratio of outliers detected
- Higher outbound pushed precision
- Lower outbound pushed recall (even more)
RESULTS

Training corpus (PAN 2010)
• Plagdet: 28.60
• Recall: 36.57
• Precision: 26.70
• Granularity: 1.11

Test corpus (PAN 2011-INTR)
• Plagdet: 16.79 (2nd place)
• Recall: 42.79 (!)
• Precision: 10.75 (?)
• Granularity: 1.03

Comparison
• ws = 5000, ss = 2500, n = 2500, outbound = .20
• Disappointing precision – dramatic drop
• Method does invariably great in recall
• Shorter documents in test?
REFERENCES