Intrinsic Plagiarism Detection
Using Character $n$-gram Profiles

Efstathios Stamatatos

University of the Aegean
Talk Layout

- Introduction
- The style change function
- Detecting plagiarism
- Evaluation
- Conclusions
Intrinsic Plagiarism Detection

• Ambitious and demanding task
• It can be used:
  – When no appropriate reference corpus is available
  – When the reference corpus is too large (web)
• Closely related to authorship verification
• Detection of irregularities of stylistic nature
  – However, not all stylistic irregularities are caused by plagiarism
Representing Writing Style

- Lexical features
- Character features
- Syntactic features
- Semantic features
- Application-specific features
Character $n$-grams

- Can be easily measured in any text
- Language-independent
- Domain-independent
- Require no text-preprocessing
- Very effective in authorship attribution
- Robust to noise
  - Obfuscation in plagiarism can be considered as noise insertion
The Proposed Approach

• The variation of document style is represented by the style change function
  – Using a sliding window over the text-length
• Writing style is represented by character $n$-gram profiles
  – The set of different character $n$-grams encountered in the text and their normalized frequencies
• A set of heuristic rules:
  – Decide whether or not the document is plagiarism-free
  – Detect the plagiarized section boundaries
  – Detect irrelevant stylistic inconsistencies
Representing Stylistic Changes

- Sliding window (length, step)
- Document
- Profile of the text window
- Profile of the whole document
- Distance estimation

- High value means stylistic anomaly
- Low value means stylistic consistency
Distance Estimation

• The sliding window text is shorter (or much shorter) than the whole document

• An accurate and robust function for imbalanced profiles is proposed by (Stamatatos, 2007):

\[
d_1(A, B) = \sum_{g \in P(A)} \left( \frac{2(f_A(g) - f_B(g))}{f_A(g) + f_B(g)} \right)^2
\]

• This is not a symmetric function
  – dissimilarity rather than distance measure
Style Change Function

- \( d_1 \) is normalized over the profile length:

\[
nd_1(A, B) = \frac{\sum_{g \in P(A)} \left(\frac{2(f_A(g) - f_B(g))}{f_A(g) + f_B(g)}\right)^2}{4|P(A)|}
\]

- Then, the style change function \( sc \) of a document \( D \) is:

\[
sc(i, D) = nd_1(w_i, D), \ i=1\ldots|w|
\]

- \(|w|\) depends on the text-length:
  - \( x \): text-length
  - \( l \): sliding window length
  - \( s \): sliding window step

\[
|w| = \left[ 1 + \frac{x - l}{s} \right]^{\frac{1}{n}}
\]
An Example

IPAT-DC
document #5
A Plagiarism-free Example

IPAT-DC
document #17
Detecting Plagiarism on the Document Level

• This is crucial to keep precision high
• Two options:
  – Pre-processing
  – Post-processing
• Plagiarism-free criterion: $S < t_1$
  where
  $S$: the standard deviation of the style change function
  $t_1$: a predefined threshold (0.02)
• Deficiencies:
  – Very short documents tend to have low sc values
  – Very long documents may contain stylistically inconsistent sections (high variance of sc)
A False Negative Example

IPAT-DC
Document #34
Identifying Plagiarized Passages

- It is assumed that at least half of the text is not plagiarized
  - The average sc value would correspond to the style of the alleged author
- In general, it is not known the amount of plagiarized text
  - All sc values greater than \( M + S \) are removed
  - \( M' \) and \( S' \) are then calculated
- Plagiarized passage criterion: \( sc(i', D) > M' + a*S' \)
  - \( a \) determines the sensitivity of the method (set to 2.0)
Detecting Irrelevant Style Changes

- Not all stylistic changes are caused by plagiarism
  - Text formatting affects style
  - Genre affects style
  - ...
- To reduce the formatting factor:
  - All text is transformed to lowercase
  - Every character $n$-gram that contains no letter characters (a-z) is removed from the profile
  - The sliding window parameters operate on letter characters
    - each window has the same number of letter characters (window length $l$) but different number of total characters (real window length $l'$)
Detecting Irrelevant Style Changes

• To reduce the multiple genre factor:
  – Special Section Criterion: $l' < t_2$
    where
  – $l'$: the real window length
  – $t_2$: a predefined threshold (1,500)
  – It combines with the plagiarized passage criterion

• Weaknesses
  – One can insert multiple non letter characters to obfuscate a plagiarized section
  – All special sections (table-of-contents, index) are considered plagiarism-free
An Example

IPAT-DC
Document #46
## Summary of Parameter Settings

<table>
<thead>
<tr>
<th>Description</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character $n$-gram length</td>
<td>$n$</td>
<td>3</td>
</tr>
<tr>
<td>Sliding window length</td>
<td>$l$</td>
<td>1,000</td>
</tr>
<tr>
<td>Sliding window step</td>
<td>$s$</td>
<td>200</td>
</tr>
<tr>
<td>Threshold of plagiarism-free criterion</td>
<td>$t_1$</td>
<td>0.02</td>
</tr>
<tr>
<td>Real window length threshold</td>
<td>$t_2$</td>
<td>1,500</td>
</tr>
<tr>
<td>Sensitivity of plagiarism detection</td>
<td>$a$</td>
<td>2</td>
</tr>
</tbody>
</table>

- Empirically derived, not optimized
### Evaluation on the Document Level

<table>
<thead>
<tr>
<th>Guess</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plagiarism-free</td>
<td>1102</td>
</tr>
<tr>
<td>Plagiarized</td>
<td>545 (22%)</td>
</tr>
<tr>
<td>Plagiarism-free</td>
<td>443</td>
</tr>
<tr>
<td>Plagiarized</td>
<td>1001 (78%)</td>
</tr>
</tbody>
</table>

- Results on IPAT-DC
False Negatives

- The majority of false negatives are relatively short documents (<30K chars)
- The shorter a document, the more likely to false negative
Evaluation on the Passage Level

<table>
<thead>
<tr>
<th>Corpus</th>
<th>IPAT-DC</th>
<th>IPAT-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.4552</td>
<td>0.4607</td>
</tr>
<tr>
<td>Precision</td>
<td>0.2183</td>
<td>0.2321</td>
</tr>
<tr>
<td>F-score</td>
<td>0.2876</td>
<td>0.3086</td>
</tr>
<tr>
<td>Granularity</td>
<td>1.22</td>
<td>1.25</td>
</tr>
<tr>
<td>Overall score</td>
<td>0.2358</td>
<td>0.2462</td>
</tr>
</tbody>
</table>

- Performance remains stable for both corpora
Recall and Precision vs. Text-length

- Recall is affected by decreasing text-length
  - A result of false negative distribution
Conclusions

• A fully-automated approach
  – Easy to follow (no text preprocessing)
  – Able to detect plagiarism-free documents
  – Able to detect plagiarized passage boundaries

• Nearly half of plagiarized passages are detected while precision remains low
  – An increased $a$ value can improve precision (and harm recall)

• Window length determines the shortest plagiarized passage that can be detected
Future Work

• Definition of more sophisticated criteria
• Parameter settings can be optimized by machine learning algorithms
• Different schemes to acquire style change function
  – Comparison of text window with the window complement
  – Comparison of text window with all the other text windows