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

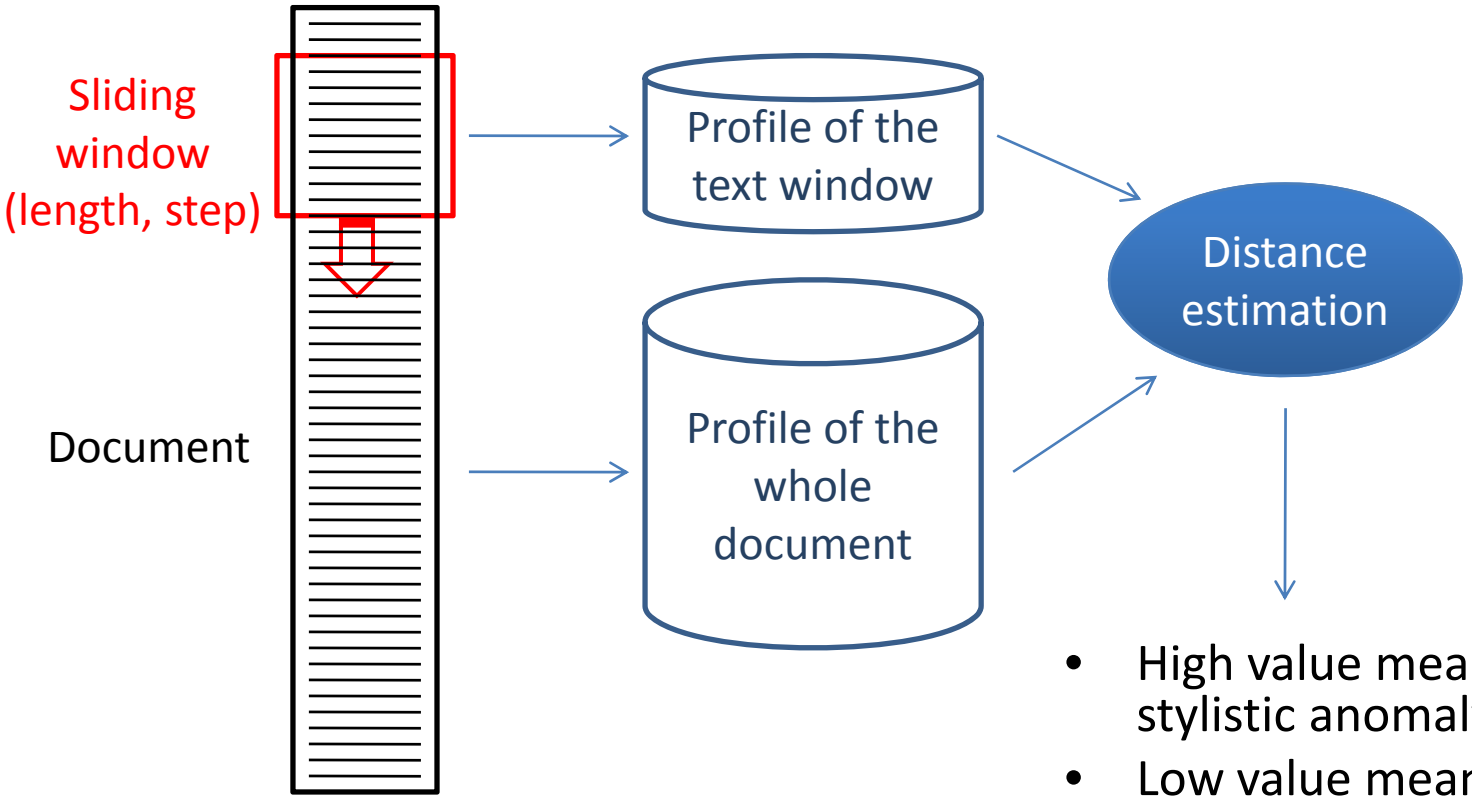
Character *n*-grams

- 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



- 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:

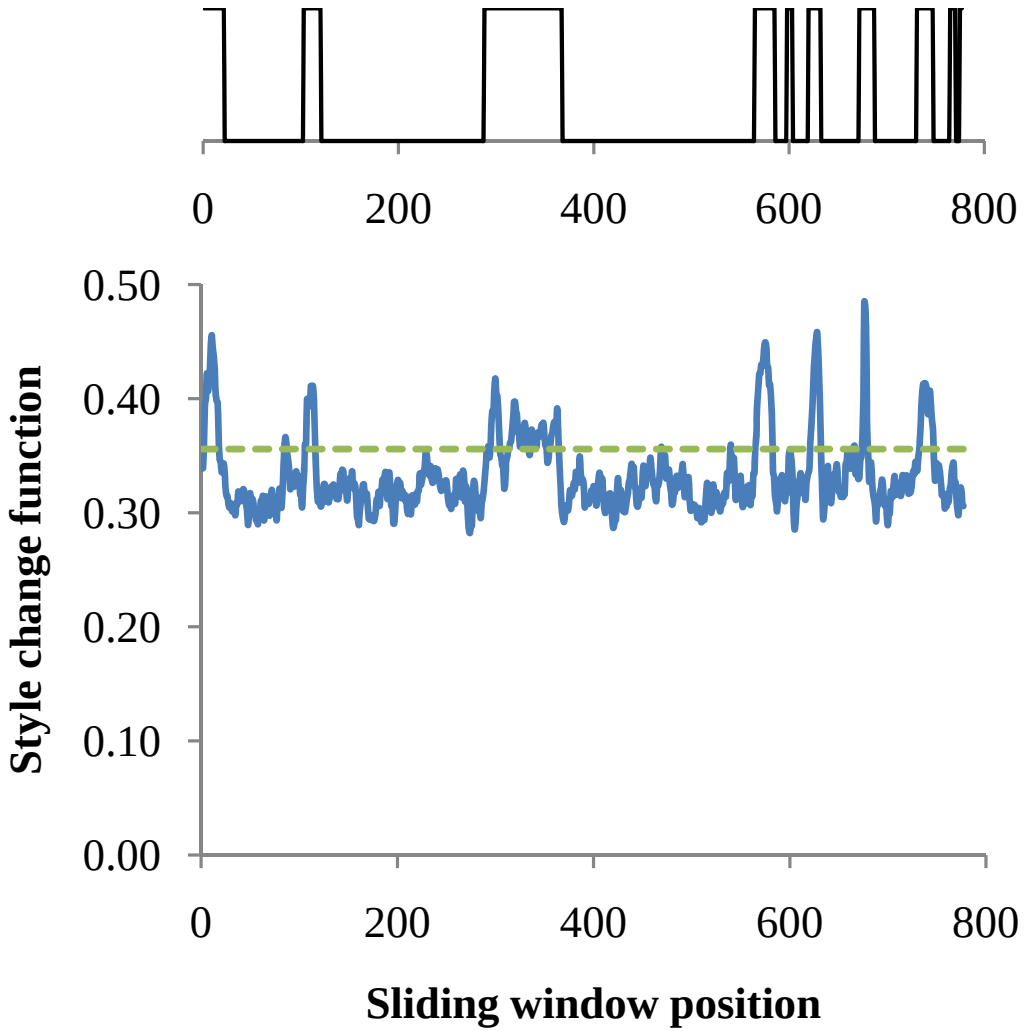
$$nd_1(A,B) = \frac{\sum_{g \in P(A)} \left(\frac{2(f_A(g) - f_B(g))}{f_A(g) + f_B(g)} \right)^2}{4|P(A)|}$$

- 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: $|w| = \left\lfloor 1 + \frac{x - l}{s} \right\rfloor$
  - x: text-length
  - l: sliding window length
  - s: sliding window step

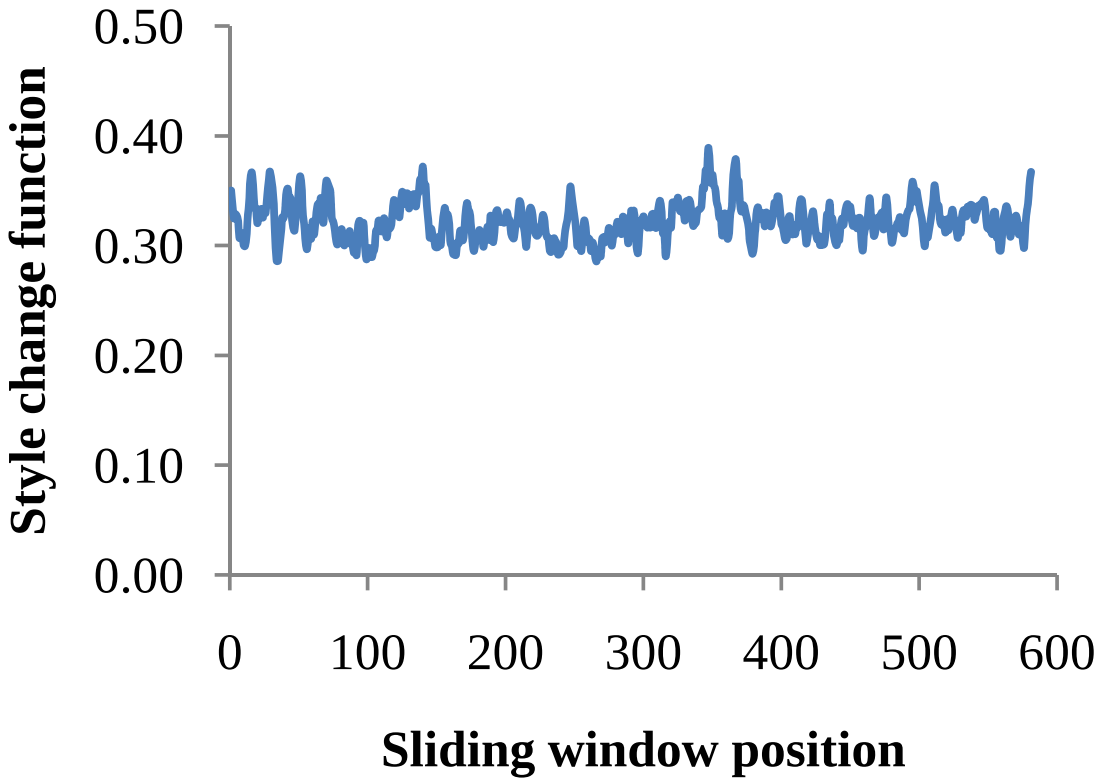
An Example



IPAT-DC

document #5

A Plagiarism-free Example



IPAT-DC
document #17

Detecting Plagiarism on the Document Level

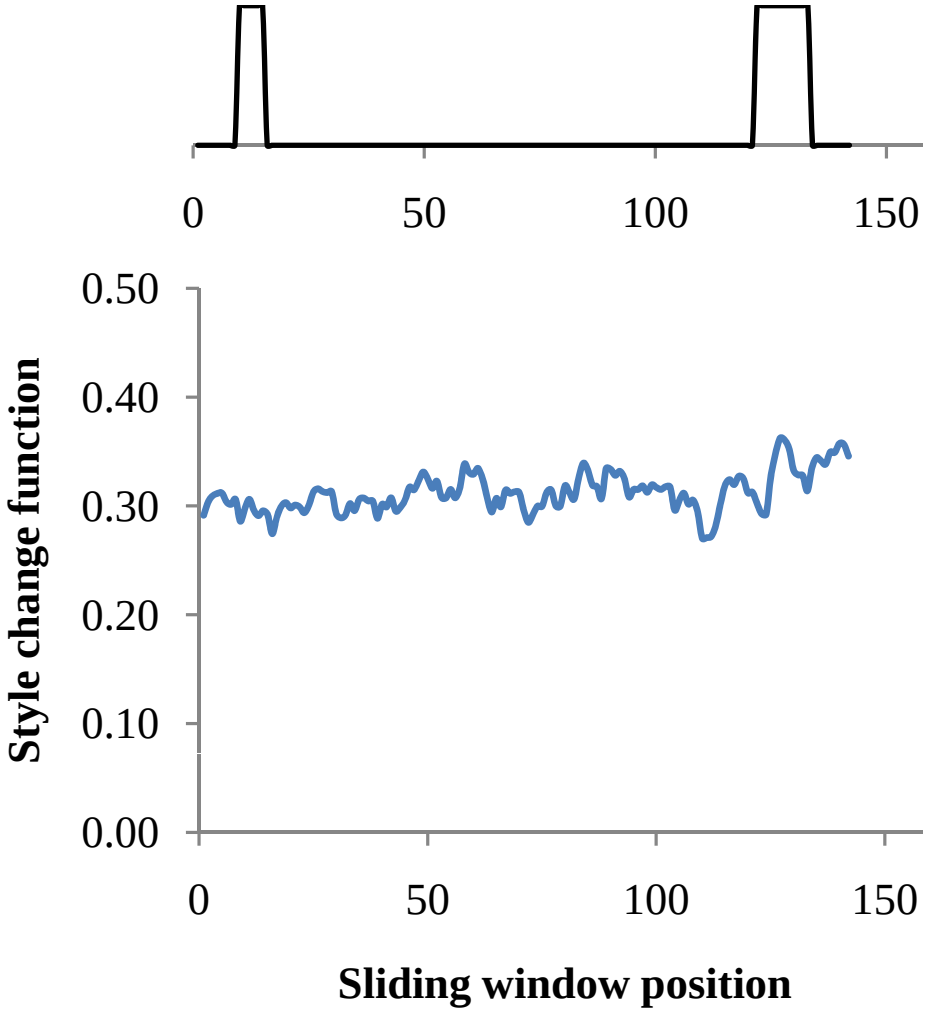
- This is crucial to keep precision high
- Two options:
  - Pre-processing
  - Post-processing
- Plagiarism-free criterion: $S<t_1$

  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

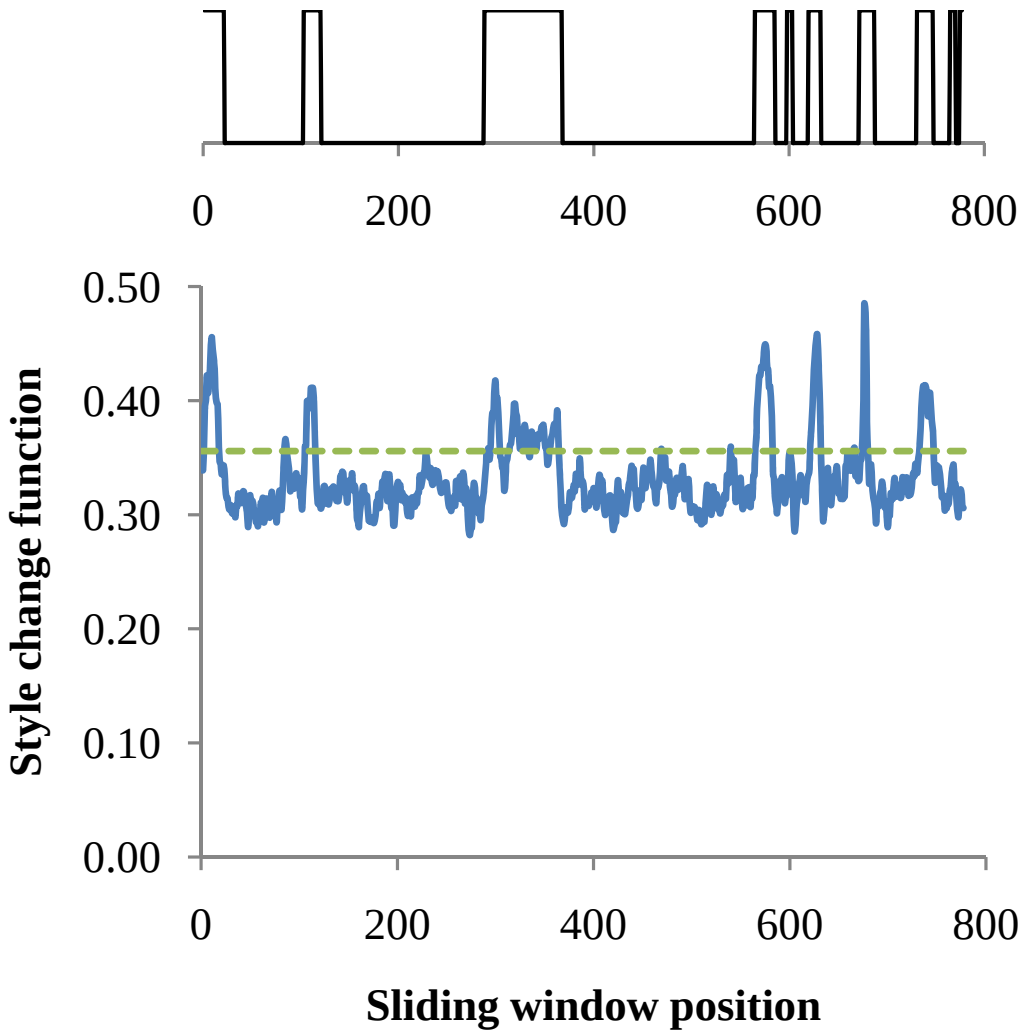


IPAT-DC
Document #34

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)

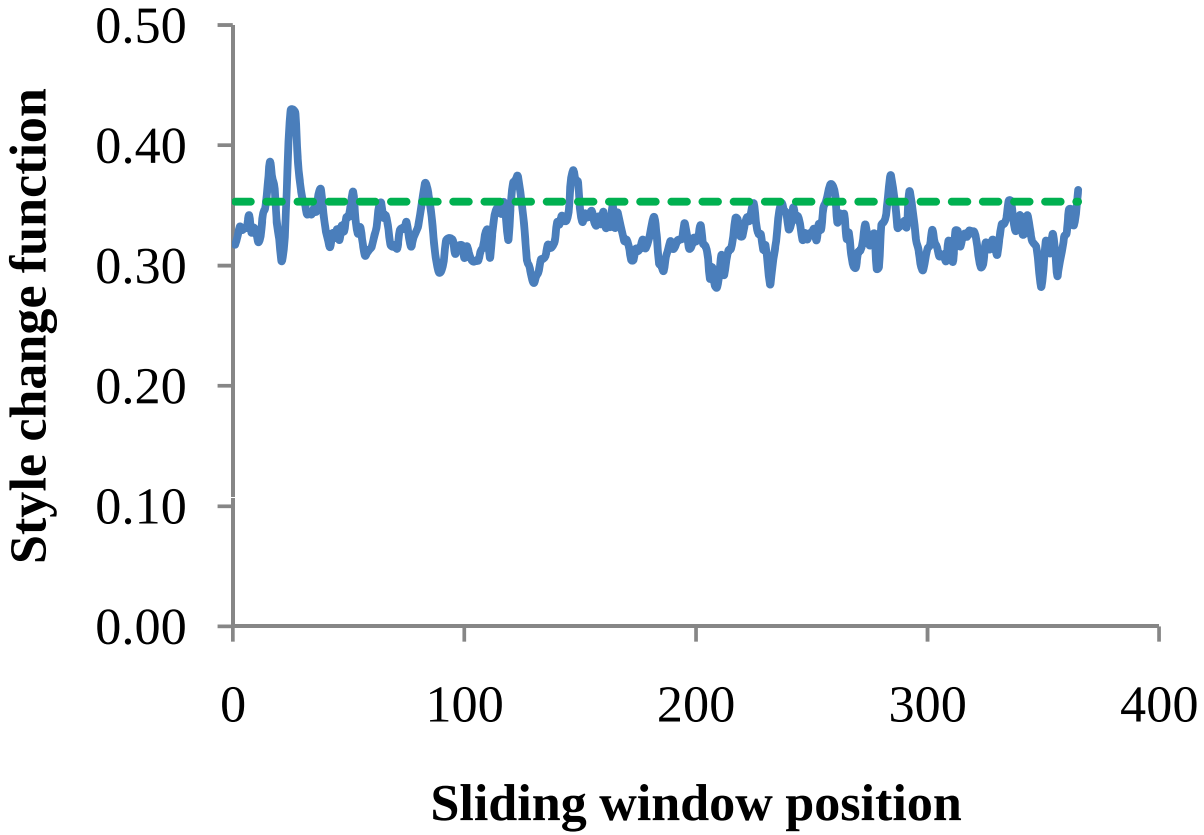
An Example



IPAT-DC

document #5

Another Example



IPAT-DC
Document #22

Detecting Irrelevant Style Changes

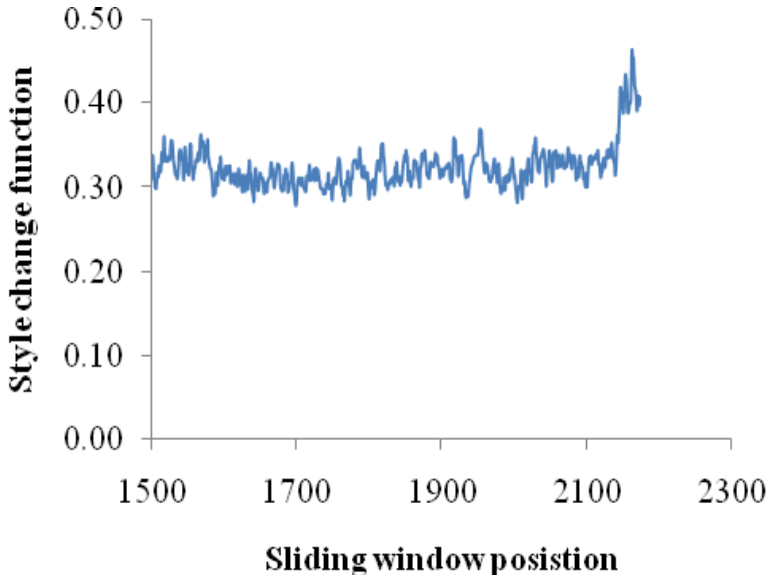
- Not all stylistic changes are caused by plagiarism
  - Text formatting affects style
  - Genre affects style
  - …
- To reduce the formatting factor:
  - All text is transformed to lowercase
  - Every character *n*-gram that contains no letter characters (a-z) is removed from the profile
  - The sliding 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')

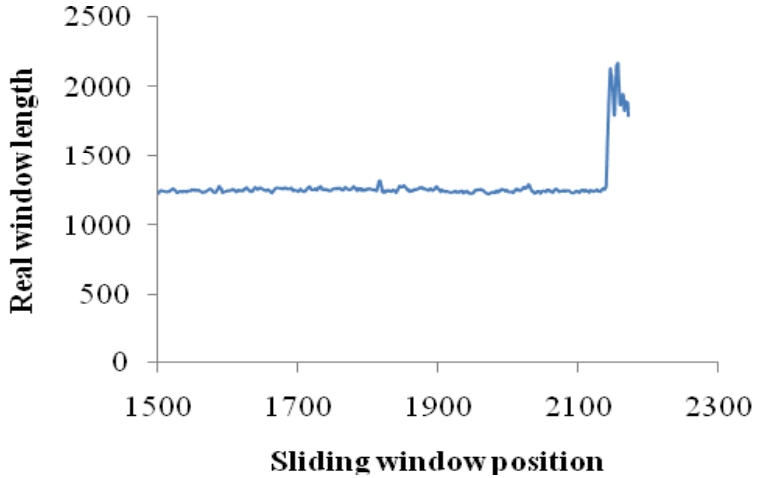
Detecting Irrelevant Style Changes

- 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

Description	Symbol	Value
Character *n*-gram length	n	3
Sliding window length	l	1,000
Sliding window step	s	200
Threshold of plagiarism-free criterion	t_1	0.02
Real window length threshold	t_2	1,500
Sensitivity of plagiarism detection	a	2

- Empirically derived, not optimized

Evaluation on the Document Level

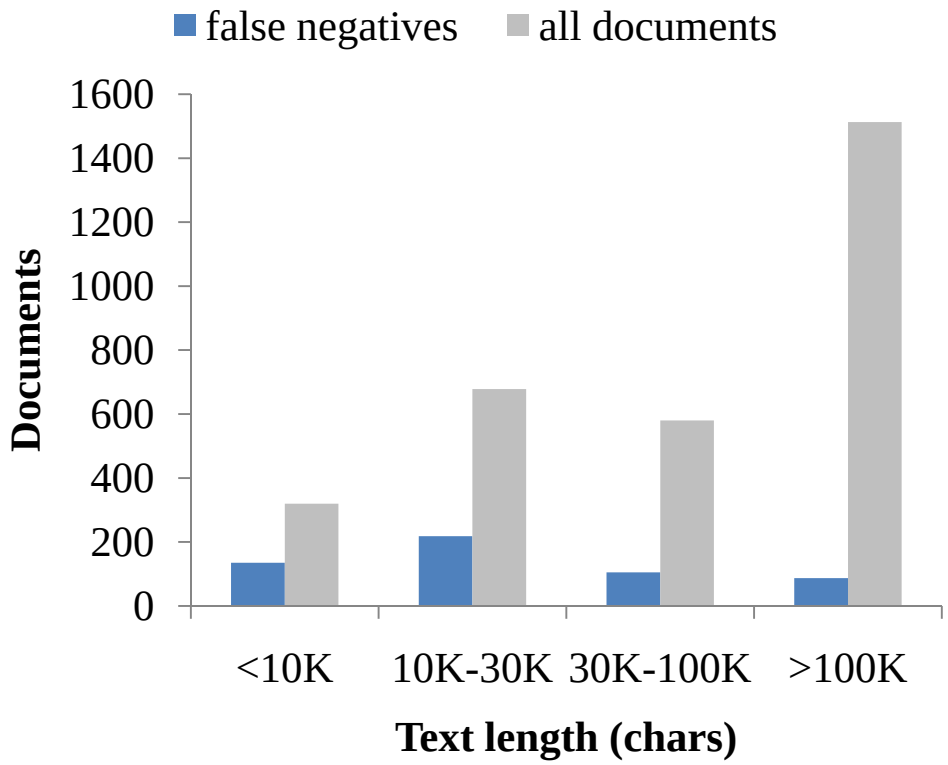
Guess	Actual	
	Plagiarism-free	Plagiarized
Plagiarism-free	1102	545 (22%)
Plagiarized	443	1001 (78%)

Plagiarized passages

Upper bound for Recall

- Results on IPAT-DC

False Negatives



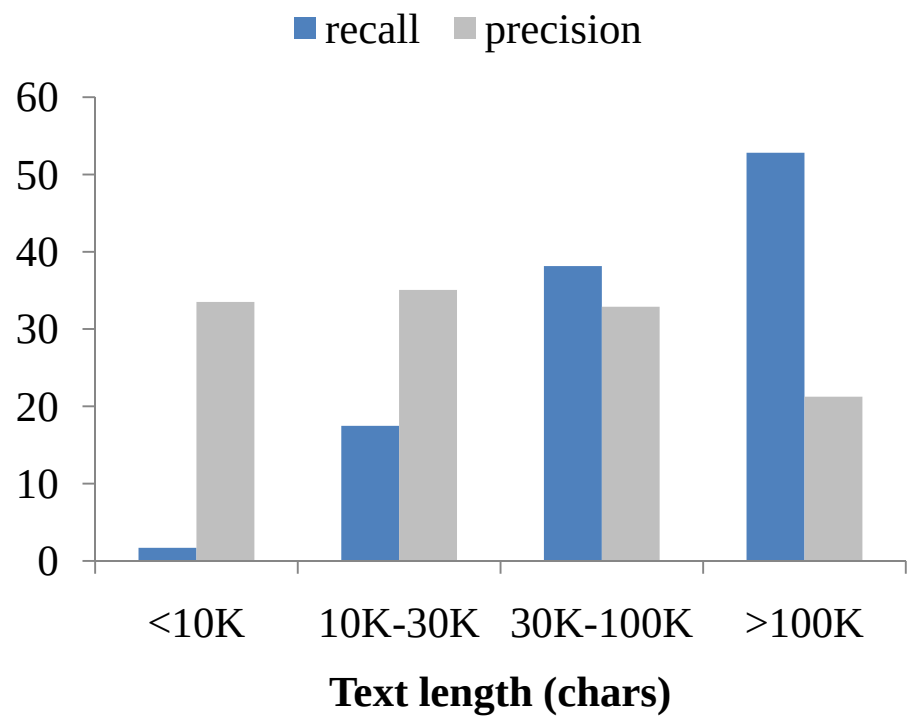
- The majority of false negatives are relatively short documents (<30K chars)
- The shorter a document, the more likely to false negative

Evaluation on the Passage Level

Corpus	IPAT-DC	IPAT-CC
Recall	0.4552	0.4607
Precision	0.2183	0.2321
F-score	0.2876	0.3086
Granularity	1.22	1.25
Overall score	0.2358	0.2462

- 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 *α* 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