

Text comparison: similarity detection with determination of possible source of plagiarism

Aleksandra Byczyńska, Aleksandra Gontarz, Włodzimierz Funika

AGH-UST, Fac. of Comp. Science, Electronics and Telecommunication, Dept. of Computer Science
emails: al.byczynska@gmail.com, aleksandra.m.gontarz@gmail.com, funika@agh.edu.pl

CGW Workshop '16
Kraków, Poland, October 24-26, 2016

Agenda

1. Introduction and problem statement
2. The idea behind our program
3. Comparison of algorithms
4. Example
5. Conclusions and future work


Popularisation of plagiarism and why similarity detection is crucial for its recognition?

Problem statement

comparing a set of different texts
and finding a possible source of plagiarism



Using 4 different algorithms

- Common Words Method
 - Longest Common Subsequence
 - Cosine Similarity
 - Levenshtein Distance
- 

Common Words Method

Treats texts as sets of words.

Produces a vector of words and their occurrences for each text and computes a correlation coefficient to describe the similarity between them.

Complexity: $O(\frac{m(m-1)}{2} \cdot n + mn) \approx O(m^2n)$

m – number of texts, k – number of distinctive words in a text, n – number of all words in a text, $k \cong n$

Longest Common Subsequence

Works on texts as sets of characters.

Seeks LCS of two texts and returns its length divided by the length of the shorter text.
Implemented with Dynamic Programming to reduce complexity.

Complexity: $O\left(\frac{m(m-1)}{2} \cdot n^2\right) \approx O(m^2 n^2)$

m – number of texts, n – number of characters in a text

Cosine Similarity

Treats texts as sets of words.

Produces a vector of words and their occurrences for each text and computes a correlation coefficient as a scalar product of the above.

Complexity: $O(\frac{m(m-1)}{2} \cdot 3n + mn) \approx O(m^2n)$

m – number of texts, k – number of distinctive words in a text, n – number of all words in a text, $k \cong n$

Levenshtein Method

Works on texts as a sets of characters.

Counts the number of changes that are needed for the two texts to become identical, as a coefficients, returns the number of changes divided by the length of the longer of texts. Implemented with Dynamic Programming to reduce complexity.

Complexity: $O(\frac{m(m-1)}{2} \cdot n^2) \approx O(m^2 n^2)$

m – number of texts, n – number of characters in a text

Output and results

Correlation coefficient – two texts are identical when the computed coefficient is 1, they have nothing in common when it is equal to 0.

The tool seeks for possible sources of plagiarisms – determined based on the amount of texts that are similar to it, as well as their coefficients.

E.g. in a set of N texts, if $N-1$ texts have a high similarity coefficient with the remaining one, and these $N-1$ texts have lower coefficients between each other than with the remainder, it is the N -th text that can possibly be the source on which the rest was based.

Example – comparison of two texts

TEXT 1

Appropriate action of the novel is a flashback of three days and two nights in **mid December** 1949 **years**. Holden, expelled from a private school Pencey Prep for a few days before the end of the first **half, looking** from the hill in the **team** last season friendly **football match** school. Then visit Professor Spencer, a history teacher, with whom he wanted to say goodbye before leaving home, but **it irked talk** about his poor performance in science quickly end the visit.

At night, after returning to the **bedroom** Holden **learns** that his roommate, Stradlater, was on a date with Jane Gallagher, **an old friend** Caulfield. When Stradlater not respond to questions about the course of dating, Holden tries to hit him, but is quickly defeated. He decides to leave school in the middle of the night and return to New York. To avoid a meeting with the family rents a room **at** the Hotel Edmont. Holden visited two nightclubs, where **is**, among others, **the former** girlfriend of his brother, D. B., and three tourists from Seattle, **to dance with**. Still, he feels lonely; all **persons found** perceived as **superficial** and boring snobs. In the hotel elevator boyfriend Maurice **child prostitutes** proposed services; Holden agrees. When a girl, Sunny, comes to **her** room, the **child** loses **value**. **It** proposes to spend time to talk, but annoying Sunny comes out.

TEXT 2

The proper action of the novel is a flashback of three days and two nights in **mid-December** 1949 **year**. Holden, expelled from a private school Pencey Prep for a few days before the end of the first **semester, watching** from the hill in last season friendly **match football** school **team**. Then visit Professor Spencer, a history teacher, with whom he wanted to say goodbye before leaving **for** home, but **irritated talking** about his bad performance in science quickly end the visit.

In the evening, after returning to the **dormitory** Holden **finds out** that his roommate, Stradlater, was on a date with Jane Gallagher, **a former** friend Caulfield. When Stradlater not respond to questions about the course of dating, Holden tries to hit him, but is quickly defeated. He decides to leave school in the middle of the night and return to New York. To avoid a meeting with the family rents a room **in** the Hotel Edmont. Holden visited two nightclubs, where **he meets** among others, **ex-girlfriend** of his brother, D. B., and three tourists from Seattle, **with whom dancing**. Still, he feels lonely; all **encountered people** perceived as **shallow** and boring snobs. In the hotel elevator **boy** boyfriend Maurice proposes **services prostitutes**; Holden agrees. When a **young** girl, Sunny, comes to **his** room, the **boy** loses **courage**. **He** proposes to spend time to talk, but annoyed Sunny comes out.

Example – results

Common Words: text1 : text2 correlation coefficient = 0,8734

LCS: text1 : text2 correlation coefficient = 0,9092

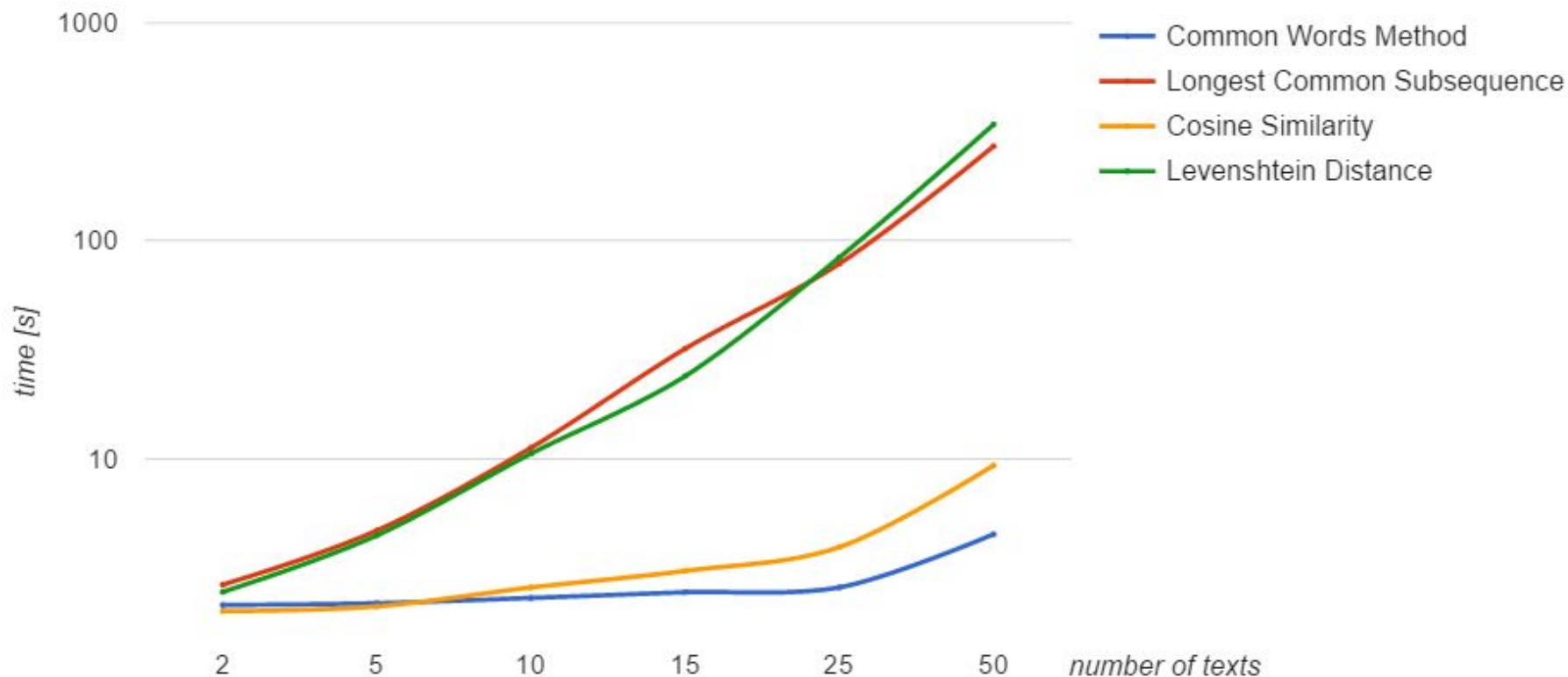
Cosine Similarity: text1 : text2 correlation coefficient = 0,7308

Levenshtein: text1 : text2 correlation coefficient = 0,8454

In every case CC is high enough to state that one of those two texts was a source of plagiarism for the other.


Performance issues

Testing results – relationships between the number of texts, time and used method





Conclusions and future work

- determining the minimum coefficient which indicates the possible plagiarism source
 - creating analysis tactics (i.e., to run less complex algorithms first and based upon the pre-computed results run more complex ones
 - enabling efficient work on inflected languages
- 



Thank you for attention!

