Detailed Comparison Module in CoReMo 1.9 Plagiarism Detector

Diego A. Rodríguez-Torrejón¹,²
José Manuel Martín-Ramos¹

¹ Universidad de Huelva
jmmartin@dti.uhu.es

² I.E.S. José Caballero
dartsystems@gmail.com

Our attendance is sponsored and granted by ELIAS (ESF Research Networking Programme).
CoReMo is a classic PAN competitor since PAN2010. Its detection speed was a highlighted goal ever. Detailed Comparison module is new however, but also improved by Surrounding Context N-grams (SCnG), an extended concept of former CTnG (case folding, stopwords/short removal, stemming and internal sort) by including new special skip n-grams to the classic consecutive. It gets almost 3 times more n-grams than words in the text, having a discriminative magnifier effect: “The quick brown fox jumps over the lazy dog” brown_fox_quick, brown_jump_quick, fox_jump_quick
The highlighted runtime speed is due to:

- **C/C++ 64 bits** programming (single core however)
- **GNU Linux 64bits OS** and **ext4 file system** platform
- Internal sort of n-grams by **Bubblesort** algorithm
- n-grams into a document ordered by **Quicksort**
- **Modified Mergesort** algorithm to compare both docs
- **Local translations** by dictionary mapping
- **Taking the advantage** of suspicious document modelling when repeated in **consecutive comparisons**
Detection is reached when minimum length and distance between n-grams conditions are got in both suspicious (counting n-grams) and source (counting chars) sections.

\[ \text{maxNgramDist} = 2 \cdot \text{chunkLength} \]
\[ \text{maxCharDist} = \text{chunkLength} \cdot \text{wordLengthAverage} \]
\[ \text{minNgramLength} = (\text{monotony} - 1.5) \cdot \text{chunkLength} \]
\[ \text{minCharLength} = \text{minNgramLength} \cdot \text{wordLengthAverage} \]

\[ \text{chunkLength} = 4 \text{ trigrams} \rightarrow 6 \text{ words (monolingual)} \]
\[ \text{Monotony} = 2 \text{ chunks} \]

Joining distance for direct detections: 80,000 chars