Zero-Knowledge Plagiarism Detection

Background

Chunking and fingerprinting are widely-used methods for identifying copy-paste plagiarism. The advantage is confidentiality, i.e. the original text cannot be recovered from the fingerprint. The disadvantage is the inability of these methods to identify translation or paraphrase plagiarism.

Goal

Devise an approach to decide if two text fragments mean the same without revealing the meaning.

Tasks

- Explore the usability of fuzzy hashes for this task.
- Explore features which may be extracted to form semantic fingerprints.
- Try to formalize a trade-off between confidentiality and comparability.

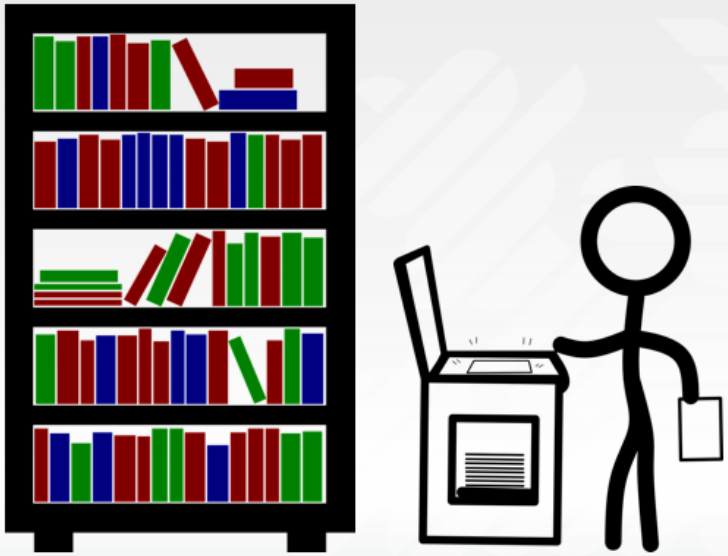
Tomáš Foltýnek
foltynek@uni-wuppertal.de

Norman Meuschke
meuschke@uni-wuppertal.de

Creating a Corpus for Plagiarism Detection

Goal

Create a corpus for the evaluation of plagiarism detection methods that includes text, mathematical formulae, citations, images and possibly other elements. The corpus should contain both original and plagiarized documents that have been obfuscated using various techniques.



Tasks

- Review the literature on previously developed corpora for plagiarism detection and the metrics for corpora quality assessment.
- Create your own corpus.
- Evaluate the suitability of your corpus.

Tomáš Foltýnek
foltynek@uni-wuppertal.de

Norman Meuschke
meuschke@uni-wuppertal.de