Plagiarism Detection

CS533 - Information Retrieval
Project Proposal

İsmail Uyanık - Deniz Kerimoğlu

Problem Description

Plagiarism: the act of presenting another's work or ideas as your own.

Availability of digital documents provides good chances for plagiarism.

Types of plagiarism:

- Copy - paste plagiarism
- Paraphrasing
- Translation
- Idea plagiarism

Motivation

Plagiarism has turned into a serious problem for publishers, researchers and educators.

The main motivation is to protect the property rights of the owner.

The task is to find all text passages in the suspicious document which have been plagiarized.

Methodology

The algorithms consists of 3 steps

- Selection: Reduces the search space by selecting a small number of suitable candidates for plagiarism
  - Word length compression
  - n-gram distance
- Matches: Performs detailed analysis on selected texts looking for matches longer than a fixed threshold (e.g. 15 characters)
  - T9 encoding
  - Longest common substring algorithm
- Squares: Joins the set of plagiarised passages

PAN Plagiarism Corpus

The PAN-PC-09 is a new large-scale dataset for the controlled evaluation of plagiarism detection algorithms.

Corpus overview:

- 41 223 text documents
- 94 202 plagiarism cases
- 70% is dedicated to external plagiarism
- 30% is dedicated to intrinsic plagiarism
- Types of cases: monolingual with and without obfustucation, and cross-lingual

Expected Results

- We will test the 3 steps algorithm on PAN dataset.
- We expect to reveal most of the low level obfuscations.
- Performance improvement via windowing method in matches step of the algorithm.

Conclusion

- Due to the very tight schedule of the PAN competition, authors state that many improvements are possible.
- Matching problem can be formulated either via Windowing or Hidden Markov Models.
- It may worth to see how standard clustering algorithms perform in this case.