PD01 Developing a Plagiarism Detection System

Background

Recent developments in language models and services such as chatGPT have allowed people to make legitimate-looking copies of texts without knowing the sources. Using ideas, and concepts without citing the sources could lead to plagiarism. A Plagiarism Detection System (PDS) helps in finding instances of copied elements in a document from potential source documents. In this project, you will work on developing a PDS. Specifically, you will learn about how documents are handled in a PDS and similarity in document pairs is calculated. Along with textual reuse detection, you will also get to work with the detection of non-textual reuse such as mathematical formulae, images, etc.

Goal

- Developing a Plagiarism Detection System.

Tasks

- Building a system interface to select a document under inspection and potential source documents.
- Integrate big data analytics platforms to handle a large number of documents.
- Implementing a document retrieval algorithm.
- Highlight detected reuse(potentially plagiarized) instances.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

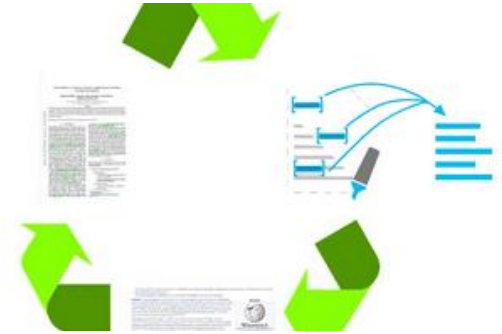
Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de

PD02 Do Plagiarism Detection Systems Really Detect Plagiarism?

Background

Existing plagiarism detection systems (PDS) detects slightly altered copies of the text. However, plagiarism occurring in scientific documents is highly disguised. In this project, you will evaluate if the existing PDS works on naturally occurring cases of plagiarism or not. There exist artificial corpora of plagiarism such as PAN but they don't represent naturally occurring cases of plagiarism because the plagiarism is artificially created. Hence, you will utilize corpora with naturally occurring cases of plagiarism such as Vroniplag, Dissernet, etc. Eventually, you would build a PDS of tomorrow that detects highly disguised cases.



Goal

- Evaluating plagiarism detection systems (open source) on naturally occurring cases of plagiarism.

Tasks

- Studying plagiarism detection approaches.
- Using corpora representing naturally occurring cases of plagiarism as test cases.
- Recording character positions of detected reuse.
- Working on non-textual reuse detection approaches.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de

RS01 Recommender system for math-heavy scientific documents

Background

Do you also experience that the recommendations you see are not fulfilling your information needs? Especially when you are looking for information on a scientific topic and would like to understand the topic more. If yes, then we are in the same boat. These days it is easy to get “Helmet” as a recommendation when buying a “bike” online, but it is hard to get relevant scientific recommendations, especially in math-heavy STEM (Science, Technology, Engineering, Mathematics). In this project, you work on the problem of generating recommendations for scientific documents with mathematical contents. You develop methods and perform experiments to generate relevant recommendations. You will utilize a dataset that represents ideal recommendation.

Goal

- Generating recommendations for math-heavy scientific documents using non-textual features.

Tasks

- Analyzing citation patterns to generate recommendations.
- Generating recommendations by finding similar math formulae.
- Identifying and formulating non-textual features from scientific documents.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de