PD01 Document Similarity Web-App using Python-Django

Background

Reuse detection allows us to find similar content between two or more documents. This help us in detecting plagiarism. To develop algorithms for reuse detection we need dataset that consists legitimate reuse cases. We have developed an web-App to annotate the cases of reuse between two documents under consideration. You aim would be to get to know the developed app in Django and integrate some features that assist users in finding the overlap of longest common tokens. This will allow users of the App to find the already present reuse in the documents under comparison.

Goal

- Developing a web-App to find similar textual-tokens and Math-tokens between two documents.

Tasks

- Preprocessing of scientific documents using python NLP library Spacy.
- Implementing a script to highlight similar shared token in two documents.
- Extending implementation to Math-token.
- Allowing user to view “n” numbers of longest common tokens.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

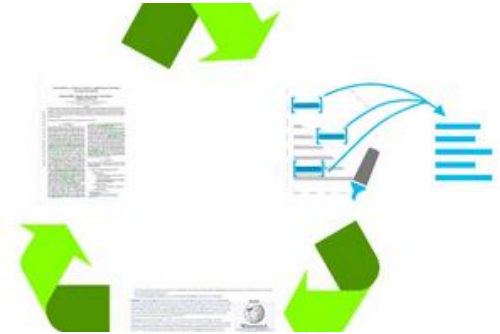
Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de

PD02 Do plagiarism detection systems (PDS) detect plagiarism?

Background

Existing plagiarism detection systems (PDS) detects slightly altered copies of the text. However, the plagiarism occurring in the scientific documents is highly disguised. In this project, we will check if the existing PDS algorithms works on naturally occurring cases of plagiarism or not. There exists artificial corpora of plagiarism such as PAN but they don't represent the actual problem we would like to solve since the plagiarism in those cases is artificially introduced. Hence, we first look for the naturally occurring cases of plagiarism and then check if the state of the art PDS systems performs well in real plagiarism cases or not.



Goal

- Analyzing state of the art plagiarism detection systems (open source).

Tasks

- Evaluating state of the art algorithms of Plagiarism detection.
- Finding corpora representing naturally occurring cases of plagiarism..
- Applying state of the art algorithms on naturally occurring cases of plagiarism and checking their performance effectiveness in detecting real plagiarism.

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 sometimes 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 areas. In this project, you work on the problem of generating recommendations for math-heavy scientific documents. You develop methods and perform experiments to generate relevant recommendations.

Goal

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

Tasks

- Exploring existing recommender system research focused on scientific documents.
- Generating recommendations for math-heavy scientific documents.
- Experimenting with textual and non-textual similarity measures.
- Evaluating developed approaches.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de