File name: FID-024.txt

Result: PLAGIARISM NOT DETECTED

Plagiarism percentage: 0%

Text to analyze:

Cross-language plagiarism detection deals with the automatic identification and extraction of plagiarism in a multilingual setting. In this setting, a suspicious document is given, and the task is to retrieve all sections from the document that originate from a large, multilingual document collection. Our contributions in this field are as follows: (1) a comprehensive retrieval process for cross-language plagiarism detection is introduced, highlighting the differences to monolingual plagiarism detection, (2) state-of-the-art solutions for two important subtasks are reviewed, (3) retrieval models for the assessment of cross-language similarity are surveyed, and, (4) the three models CL-CNG, CL-ESA and CL-ASA are compared. Our evaluation is of realistic scale: it relies on 120,000 test documents which are selected from the corpora JRC-Acquis and Wikipedia, so that for each test document highly similar documents are available in all of the six languages English, German, Spanish, French, Dutch, and Polish. The models are employed in a series of ranking tasks, and more than 100 million similarities are computed with each model. The results of our evaluation indicate that CL-CNG, despite its simple approach, is the best choice to rank and compare texts across languages if they are syntactically related. CL-ESA almost matches the performance of CL-CNG, but on arbitrary pairs of languages. CL-ASA works best on â■■exactâ■■ translations but does not generalize well.

Sentence analysis:

Original sentence (file FID-024.txt):

'Cross-language plagiarism detection deals with the automatic identification and extraction of plagiarism in a multilingual setting.'

Original sentence (file FID-024.txt):

'In this setting, a suspicious document is given, and the task is to retrieve all sections from the document that originate from a large, multilingual document collection.'

Original sentence (file FID-024.txt):

'Our contributions in this field are as follows: (1) a comprehensive retrieval process for cross-language plagiarism detection is introduced, highlighting the differences to monolingual plagiarism detection, (2) state-of-the-art solutions for two important subtasks are reviewed, (3) retrieval models for the assessment of cross-language similarity are surveyed, and, (4) the three models CL-CNG, CL-ESA and CL-ASA are compared.'

Original sentence (file FID-024.txt):

'Our evaluation is of realistic scale: it relies on 120,000 test documents which are selected from the corpora JRC-Acquis and Wikipedia, so that for each test document highly similar documents are available in all of the six languages English, German, Spanish, French, Dutch, and Polish.'

Original sentence (file FID-024.txt):

'The models are employed in a series of ranking tasks, and more than 100 million similarities are computed with each model.'

Original sentence (file FID-024.txt):

'The results of our evaluation indicate that CL-CNG, despite its simple approach, is the best choice to rank and compare texts across languages if they are syntactically related.'

Original sentence (file FID-024.txt):

'CL-ESA almost matches the performance of CL-CNG, but on arbitrary pairs of languages.'

Original sentence (file FID-024.txt):

'CL-ASA works best on â■■exactâ■■ translations but does not generalize well.'