<table>
<tr><th rowspan="2">cases</th><th>doc_1</th><th>doc_2</th><th>decision</th><th>id</th></tr>
<tr>
<td>
<table>
<tr><th>authors</th><td>Xintao Chu
Jianping Liu
Jian Wang
Xiaofeng Wang
Yingfei Wang
Meng Wang
Xunxun Gu</td></tr>
<tr><th>title</th><td>CSDR-BERT: a pre-trained scientific dataset match model for Chinese Scientific Dataset Retrieval</td></tr>
<tr><th>publication_date</th><td>2023-01-30 07:12:38+00:00</td></tr>
<tr><th>source</th><td>SupportedSources.ARXIV</td></tr>
<tr><th>journal</th><td>None</td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>http://arxiv.org/pdf/2301.12700v3
http://arxiv.org/abs/2301.12700v3
http://arxiv.org/pdf/2301.12700v3</td></tr>
<tr><th>id</th><td>id-5652472759003822637</td></tr>
<tr><th>abstract</th><td>As the number of open and shared scientific datasets on the Internet increases under the open science movement, efficiently retrieving these datasets is a crucial task in information retrieval (IR) research. In recent years, the development of large models, particularly the pre-training and fine-tuning paradigm, which involves pre-training on large models and fine-tuning on downstream tasks, has provided new solutions for IR match tasks. In this study, we use the original BERT token in the embedding layer, improve the Sentence-BERT model structure in the model layer by introducing the SimCSE and K-Nearest Neighbors method, and use the cosent loss function in the optimization phase to optimize the target output. Our experimental results show that our model outperforms other competing models on both public and self-built datasets through comparative experiments and ablation implementations. This study explores and validates the feasibility and efficiency of pre-training techniques for semantic retrieval of Chinese scientific datasets.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>
<table>
<tr><th>authors</th><td>Xintao Chu
Jianping Liu
Jian Wang
Xiaofeng Wang
Yingfei Wang
Meng Wang
Xunxun Gu</td></tr>
<tr><th>title</th><td>CSDR-BERT: a pre-trained scientific dataset match model for Chinese Scientific Dataset Retrieval</td></tr>
<tr><th>publication_date</th><td>2023-01-31 00:00:00</td></tr>
<tr><th>source</th><td>SupportedSources.INTERNET_ARCHIVE</td></tr>
<tr><th>journal</th><td></td></tr>
<tr><th>volume</th><td></td></tr>
<tr><th>doi</th><td></td></tr>
<tr><th>urls</th><td>https://web.archive.org/web/20230206013603/https://arxiv.org/pdf/2301.12700v2.pdf</td></tr>
<tr><th>id</th><td>id4508891993084394901</td></tr>
<tr><th>abstract</th><td>As the number of open and shared scientific datasets on the Internet increases under the open science movement, efficiently retrieving these datasets is a crucial task in information retrieval (IR) research. In recent years, the development of large models, particularly the pre-training and fine-tuning paradigm, which involves pre-training on large models and fine-tuning on downstream tasks, has provided new solutions for IR match tasks. In this study, we use the original BERT token in the embedding layer, improve the Sentence-BERT model structure in the model layer by introducing the SimCSE and K-Nearest Neighbors method, and use the cosent loss function in the optimization phase to optimize the target output. Our experimental results show that our model outperforms other competing models on both public and self-built datasets through comparative experiments and ablation implementations. This study explores and validates the feasibility and efficiency of pre-training techniques for semantic retrieval of Chinese scientific datasets.</td></tr>
<tr><th>versions</th><td></td></tr>
</table>
</td>
<td>DUPLICATES</td>
<td>6</td>
</tr>
</table>