Plagiarism Candidate Retrieval Using Selective Query Formulation and Discriminative Query Scoring
Osama Haggag, Samhaa El-Beltagy

Data Preparation

- Preprocessing
- Tokenization
- Frequency Distribution
- Keyphrase Extraction
- Sentence Tokenization

Word frequencies
Freq = 1
Freq > 1

Keyphrase
Query has to be < 10 keywords

Query Generation

Word frequencies

4-sentence chunk

Segment

Query

4-sentence chunks

Query Generation

Query has to be < 10 keywords

Search

Query

Snippet

> 50%?

Skip to next Query

Consider document a source
Add to list

Search

Query

Snippet

> 50%?

Skip to next Query

Consider document a source
Add to list