The assessment of the quality of the exhaustive cosine similarity search for similar documents retrieval

Rafał Frączek, Agnieszka Dąbrowska-Boruch, Maciej Wielgosz,
Andrzej Dorobisz, Marcin Pietroń, Michał Karwatowski,
Sebastian Koryciak, Paweł Russek,
Ernest Jamro, Kazimierz Wiatr

Academic Computer Centre Cyfronet AGH

Text search methods

- Exhaustive Cosine Similarity (ECS)
  - Python text retrieval system aimed at finding similar texts
  - full-text search, pdf handling
  - web-based user interface
- Solr
  - Java open-source text search platform provided by the Apache Lucene
  - full-text-search, real-time indexing, dynamic clustering, database integration, pdf handling
  - REST-like HTTP/XML and JSON API

Mathematical Models

- Local documents repositories
- Text preprocessing
  - removal of all special characters,
  - removal of all redundant words,
  - Lemmatization
  - keywords extraction
- Numeric model construction
  - VSM, TF-IDF
- Classification
  - Nearest neighbour (cosine similarity measure)
- Jensen-Shannon divergence

Experiments