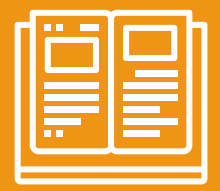
Indexing and Search Solutions in Java and Python



A Comparative Analysis

Aștilean Rareș | Dan Rareș | Dănilă Miruna | Olteanu Teodora | Vasiu Andrei

Implementing IR Solutions

Exploring Java Lucene and Python Whoosh in Information Retrieval

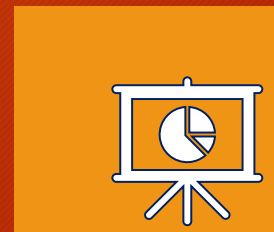
- This project delves into the development and comparison of two distinct Information Retrieval systems. Utilizing Java's Lucene and Python's Whoosh, we aimed to enhance the efficiency and accuracy of data retrieval processes. Our approach involved indexing and querying Wikipedia pages, focusing on optimizing search results and handling various data challenges.

Methodological Approach

Exploring Dual IR Systems: Java Lucene & Python Whoosh

- **Java Lucene:**
  Employed for its robust indexing and search capabilities. We designed an indexer in Java to process Wikipedia pages, optimizing search queries for precision and efficiency. Lucene's EnglishAnalyzer method was used for effective term processing.
- **Python Whoosh:**
  Utilized for its simplicity and effectiveness in data retrieval. The project leveraged Whoosh for indexing Wikipedia content, experimenting with both naive and improved indexing techniques, including lemmatization and stop-word removal.

Project Results and Insights



Measuring Success in Information Retrieval

- **Lucene Results:**
  Showcased high effectiveness in indexing and querying Wikipedia pages. Key metrics like Precision at 1 (P@1) and Mean Reciprocal Rank (MRR) indicated strong performance, particularly in handling straightforward, single-answer queries.

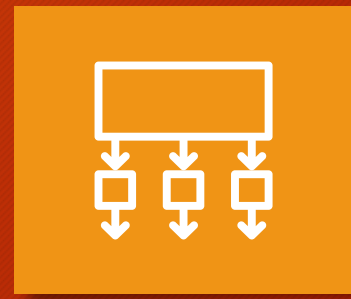
- **Whoosh Results:**
  Demonstrated a simpler yet effective approach. The naive indexer outperformed the improved indexer in certain cases, highlighting the importance of query structure and content processing techniques.

Project Challenges and Solutions

- **Key Challenges:**
  We faced issues like ambiguous queries and inconsistent Wikipedia formats. These challenges required innovative approaches to ensure the reliability and accuracy of our IR systems.

- **Adopted Solutions:**
  To address these challenges, we optimized our query processing techniques and experimented with different indexing strategies, including the use of Lucene's StandardAnalyzer and Whoosh's lemmatization features.

Conclusion and Future Directions



Summarizing Key Achievements

- **Key Takeaways:**
  Our project successfully demonstrated the capabilities of Java Lucene and Python Whoosh in handling information retrieval tasks. We achieved notable success in query accuracy and processing efficiency, providing valuable insights into IR system optimization.

- **Looking Ahead:**
  Future work could explore the integration of AI and machine learning techniques to further enhance query understanding and result relevance. Additionally, experimenting with more diverse data sets could provide broader insights into the adaptability of IR systems.

Thank You!