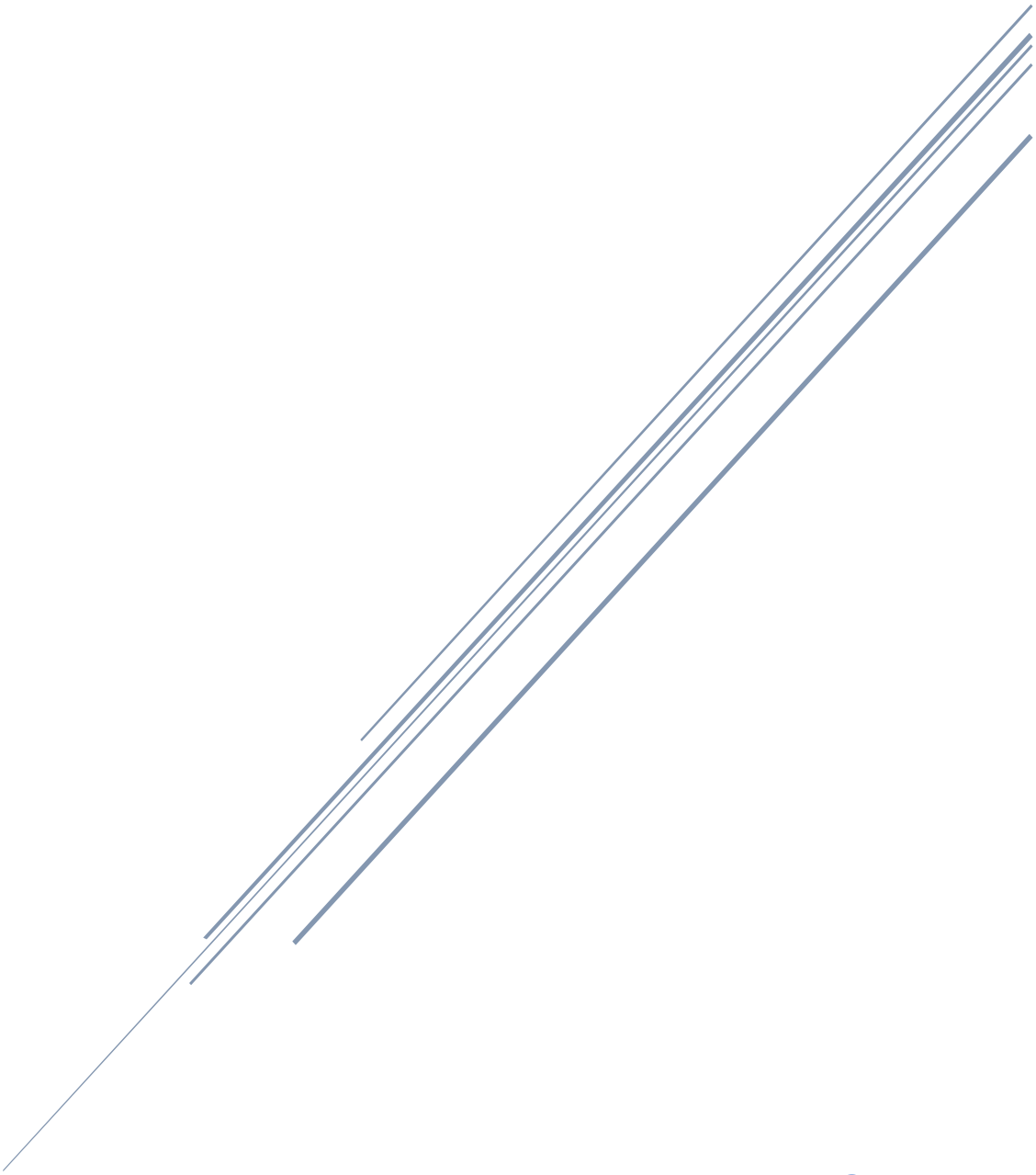
INFORMATION RETRIEVAL

Programming Assignment 05



Use Case 1
SEARCH RESULT CLUSTERING

TEAM MEMBERS

Supriya Pandurangacharya Upadhyaya

244815
supriya.upadhyaya@st.ovgu.de

Aarathi Vijayachandran Chettiar Bhagavathi

244267
aarathi.vijayachandran@st.ovgu.de

Abhinav Pareek

244209
abhinav.pareek@st.ovgu.de

Karthikeyan Muthukumar

239802
karthikeyan.muthukumar@st.ovgu.de

OVERVIEW

A Rest API was created using Java SpringBoot to index and retrieve documents, which were then clustered, and the clusters visualized with the help of Angular.

OBJECTIVE

- Analyzing and indexing at least 1000 documents from Wikipedia.
- Implementation of a clustering algorithm based on similarity score.
- Visualization of the retrieved clustered results.

TECHNOLOGIES

BACKEND

- Lucene
  - One of the most efficient Information Retrieval tools
  - Recommended for the Information Retrieval course
- Springboot
  - To power the rest API
  - Selected mainly to keep the main language of the backend as Java
- Weka library
  - For Clustering (Unsupervised Machine Learning)
  - Selected mainly to keep the main language of the backend as Java
- Gradle
  - As a packaging and building tool

FRONTEND

- Angular

  For the front end, angular framework was used with its many helpful tools it made the development of the front-end straightforward.
- HTML CSS

  The UI was developed with the help of HTML & CSS.
- Highchart.js

  The clusters are visualized with the help of highchart.js angular library for interactive charts.

INSTRUCTIONS FOR PROGRAM EXECUTION

FRONTEND

- The project runs with npm, angular and typescript and therefore install them
- Unzip the project zip file
- Run “npm install” in the command line after changing directory to the front-end folder (data-fetch-app-main) and wait for it to install the dependencies
- Run “npm start” and that will launch a development server running on localhost, port 4200 (http://localhost:4200/).
- If running scripts is disabled on the pc, then open a cmd prompt as Administrator. Pass the command Set-ExecutionPolicy -ExecutionPolicy RemoteSigned, and then ‘A’ (Yes to All)

BACKEND

- This project requires gradlew 7.0 or above to be installed
- Unzip the project zip file
- While the front-end is running on a separate command line instance, start a second command line instance and change directory to the back-end folder (folder name: complete)
- Run “gradlew bootrun” command
- The command line may get stuck at ‘80% Executing’ - it’s normal and it means program is running.

PROJECT DESCRIPTION

- This project was started by crawling 1000 Topics from Wikipedia
- An analyzer was created to index the documents - EnglishAnalyzer which uses Lowercase porter stemming and stopword removal filters among other filters
- Then the query that entered by user was used to find the top 10 hits using BM25 Similarity
- After that, the top 10 documents were clustered to the number of clusters entered by the user using the K Means Algorithm.
- Then, the user is redirected to the Wikipedia page of the Topic chosen by the user.
- After completing the back end, front-end development was started, while maintaining the performance of the project
- Dynamic and static website pages were combined for best performance
- Possible future improvements:
  - Feature for the user to add a specific Topic from Wikipedia
  - Feature for the user to choose the similarity method to be used