Spilios Spiliopoulos : 4495
Kostanitnos Chatzopoulos : 1796

Link to github Repository: https://github.com/spilios4495/Anaktisi-Pliroforias-Phase_1

@README.md:
Spilios Spiliopoulos : 4495
Kostanitnos Chatzopoulos : 1796

Project Programming Language: Python | Java

General Project Description:

The project's goal is to create a search engine that will be able to find the best results based on user's inputs.

The system will be able to achieve the best search results based on key words, relevancy, search history, synonym extension, typo correction and acronym extension.

Machine learning could be integrated into the system in order to achieve the feature of autocomplete sentences, search propositions abd more.

This system will be based on the open-source library "Lucene":

"Lucene Core is a Java library providing powerful indexing and search features, as well as spellchecking,
hit highlighting and advanced analysis/tokenization capabilities. The PyLucene sub project provides Python
bindings for Lucene Core."

In particular the dataset that will be used for this project is in the following format:
['source_id']['year']['title']['full_text']
Where
'source_id' : The unique id of each Paper
'year' : The publication year of the Paper
'title' : The title of the paper
'full_text' : The contents of the paper
Additional information might be added later for best results

In order to collect our data we used the "All NeurIPS (NIPS) Papers" from Kaggle at @ https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated/data?select=papers.csv
With the python "papers_selection.py" script we collect the exact number of the Papers for our dataset (500 samples)

Introduction: The goal of the system is to provide an efficient mechanism to search for Paper's articles and related information about authros and potentially other Papers.

Text analysis and index construction: For pre-processing the data, various techniques will be used such as removing stop words, stemming for root word analysis, and creating indexes for fields such as Paper's title, publication year and the contents. The indexes will allow various search modes, such as keyword search, field search (e.g., Paper's title), and others.

Search: Search will be implemented using the available Lucene features, such as query string search and field search as mentioned above.

Results Presentation: results will be presented in pages with a list of Paper’s matching the search query. Each Paper will include the title, year, and a snippet of the contents(articles), with keywords highlighted. In addition, the user will be able to group the results in a batch of 10 at a time based on a criterion such as artist or other based on other labels/fields.

In depth details will be included at the final Report as well
Precise information will be available soon, as the project starts
This README will be updated soon