Information Retrieval

Term Project Draft Report

“Information Retrieval System on IMDB”

Abstract

This draft report aims showcase the details of our project over the course of the semester. It will include descriptive definitions of the project, its functionality, a brief introduction to the project, our workload division and the platforms and technicalities we plan on using as we are working on the project.

**Table of Contents**

[Abstract 2](#_bookmark0)

[Project Description 4](#_bookmark1)

[Functionality 4](#_bookmark2)

[Introduction to the Project 4](#_bookmark3)

[Workload Division 5](#_bookmark4)

[API, Library & Platforms 5](#_bookmark7)

Project Description

In this project we are going to design classic information movie database like mostly known as “IMDB”. The working process includes some several steps which can be represented as making unstructured data to structured data. In this system, the movie database will contain information about some movie aspects such as actor/actress/director names, genres, titles, country, and year that the movie belongs to.

Functionality

This projects functionalities includes some steps. First of all, we are going to use a web crawler which will be the Selenium for reaching some data about movies and creating connection to the programming language (Java or Python) to it. After reaching some documents, we will use some data preprocessing operations which will be N-gram and Lemmetizer for separating and combining the data from the documents. We will transfer the data into the database which will be PostgreSQL. After reaching some structured data, we will use regex library for indexing process. This process aims to do finding and replacing operations to the data and checking the validation of it. Then, we are going to create vector space. Last but not least, we are going to normalize the data that we processed. In the demo session, we will show you the live data processing operation that we mentioned in the functionality part.

Introduction to the Project

In this project, we will create a classic movie database, commonly referred to as "IMDB." The working process consists of numerous processes that may be described as the transformation of unstructured data into structured data.

Workload Division

We as a group decided that two group members will be responsible with setting up the database for the project (Doğa & Çağla) while other two members will be responsible for coding and implementing (Can Derin & Can). But as project grows these divisions of work might change between group members as they wish to help one another on their sections.

API, Library & Platforms

* + The WebCrawler will be set-up using Python or Java.
  + Vector Space Model will also be implemented through Python.
  + A database will be created using a fitting language.
  + Normalization step will be done over the database.