TED ÜNİVERSİTESİ

CMPE 451 Project Specifications Report

Group 13

Göktuğ Yılmaz

Fatih Sultan Mehmet Çelik

Mehmet Anıl Akgül

Bensu Şeker

1. Introduction

Artificial intelligence and machine learning are topics that almost everyone discusses in today's world and wants to be a part of it in some way, but there is no doubt at the center of these topics "the data" takes place. Data, in its simplest definition, means "a little information".It is even possible to define it more simply and simply, and to call the data an "information particle" However, the position of this concept ,the data , in the new world order whose definition is so simple, is just as complex and important. This world and human beings with the developing technology produce more and more comprehensive data every day . Even on the internet, which is used by more than 4 billion people today, it is known that more than 2,500 petabytes of data are produced in every day. Making sense of this massive data is possible with artificial intelligence. It is possible to use the data in almost every field by making sense of it by processing it with artificial intelligence.It is possible to gain superiority in the military field or to take big steps towards energy efficiency, and even to have intelligence capabilities that are unprecedented in history, with data interpreted by artificial intelligence.Or, it is possible to facilitate human life with smart urbanization applications, and to gain great advantages in many areas from finance to cyber security, from business intelligence to information management.What matters is good use of artificial intelligence and data collaboration. There is a strong and indispensable relationship between artificial intelligence algorithms and data sets.Artificial intelligence cannot make meaningful and accurate inferences in instant situations without large amounts of properly prepared data sets.As for the acquisition of necessary data for artificial intelligence, most institutions do not share their data for the purpose of protecting corporate privacy, but despite this, it has become possible to obtain data from global platforms today.

1.1 Description

In data science and machine learning, datasets are vital. No matter how strong the use case or talented data scientists are, the Machine Learning model will fail if your dataset is insufficient. Datasets are employed at every stage of ML project development, from ML model training to model tuning to ML model testing which is the three Datasets used are the training set, the validation set, and the testing set. The training dataset prepares the machine learning algorithm to use ideas like artificial neural networks to learn and generate the desired output. The dataset includes both the input data and the anticipated output of the ML algorithm. The input data are contained in the test dataset, and the accuracy of the output is typically confirmed by human verification. Based on this, we decided to make a dataset website project using information retrieval. We prepare a repository website containing datasets for machine learning and data science, proceed through a computer as a server for the website and we will use the same computer as a database, going to perform the relevant datasets according to the search words made over the sites where artificial intelligence data is shared. On this website, we will present datasets that users mostly search on other sites (e.g., Kaggle, Google dataset search). We purpose to doing this with keyword extraction. A text analysis technique called keyword extraction mechanically extracts from a document the most frequently used and important words and expressions. It helps in recognizing the main topics presented in texts and extracting their essential concepts. The project "Data Cosmos" maintains "ML dataset repositories," or groups of datasets for machine learning or datascience projects. Many of the data sets are "open source," making them accessible to everyone in the globe. We will proceed through a computer as a server for the website and we will use the same computer as a database. We identified 10 sites with datasets. For example, we foresee creating the code in Java that can automatically pull these datasets from Google or Kaggle.

1.2 Constraints

1.2.1 User Interface Constraints

This system is straightforward and easy to use. All of the system's functionality should be clear to a user with a working knowledge of basic browser navigation.

1.2.2 Hardware Constraints

Most desktop and laptop computers that support Java and HTTP should be able to use the system.

1.2.3 Software Constraints

The system is designed to work with versions of Firefox 4 and higher, Google Chrome 10 and higher, and Internet Explorer 8 and higher.

1.2.4 Data Management Constraints

The system must be able to interface with other parts in accordance with their requirements.

1.2.5 Operational Constraints

The maximum number of users the system can sustain at a given moment is constrained by its operating server.

1.2.6 Site Adaption Constraints

At the conclusion of the system creation, the component will be adjusted to the overall system.

1.3 Professional and Ethical Issues

Datacosmos is a project that we are creating to help people with their data set research. In professional way, we did not encounter with another project like this. There are some websites which is similar in a way and collects data from other websites like Cimri.com and compare those data. The difference in our project is we will just search the data sets according to the users filters in pre-determined websites and present those data sets to the user. Those data sets have categories so our user can filter while searching. In ethical way, we are not violating any right by doing our project and we will make it easy for people to find necessary data sets, so it is beneficial for our users.

2. Requirements

2.1 Web Server

Spring Boot can run as a standalone server, but we planning to use with Apache because of that Apache putting it behind an Apache web server has several advantages, such as load balancing and cluster management.

2.2 Object Relational Mapping (ORM)

We will use Hibernate for converting data between relational databases and Java.

2.3 Building Microservice

Spring Boot offers non-functional features that helps creating fast production-ready apps. Simple to install embedded servers using containers.It helps in monitoring several component and in external component configuration.

2.4 Logical database

User accounts, dataset categories and other data will all be preserved in the database. A solid database architecture is necessary to ensure that the database supports concurrent access and is always maintained consistent. We will used Hibernate which is a foundation for mapping relational databases to object-oriented domain models is provided by the open source object relational mapping (ORM) program known as Hibernate.

2.5 Web crawling with Selenium

We are planning to collect dataset with web crawl which is web crawling is one of the most popular way of information gathering mechanism. We are focusing on a java application that can be used to crawl a Web on top of Selenium library.

2.6 Testing

- Unit Testing with JUnit
- Integration Testing using SpringBootTest
- End-to-End Testing
- User Testing

3. References

How to Identify and Manage Constraints in Project Management. (n.d.). Smartsheet. Retrieved October 27, 2022, from https://www.smartsheet.com/content/project-constraints

Alwis, R. (2018, April 28). *Web Crawling [Java][Selenium] - Tech Vision*. Medium. https://medium.com/tech-vision/web-crawling-java-selenium-8805fc006db1

Spring Boot Microservices: Building Microservices Application Using Spring Boot. (2022, March 28). Edureka. https://www.edureka.co/blog/microservices-with-spring-boot

Hibernate - ORM Overview. (n.d.). Retrieved October 31, 2022, from https://www.tutorialspoint.com/hibernate/orm_overview.htm