JAVA LABELING SYSTEM

Software Requirement Analysis

V1.1

Lead Software Developers

Mahmut Hilmi ARIKMERT

Muhammed Enes AKTÜRK

Yunus Emre ERTUNÇ

Rabiul ISLAM

Asaf Talha GÜLTEKİN

Hamza TÜRKMEN

Kerim BOYACI

Customers

Murat Can GANİZ

Lokman ALTIN

Introduction

The purpose of this project is creating a data labeling system for different Natural Processing Language problems using a console based program. Industry standard Object Oriented Designing and Programming will be implemented for this project. The chosen language is Java and agile software development methodology will be followed throughout the project.

Overview

Data labeling means assigning the several predetermined labels (class labels, categories, tags) to a group of instances (samples, examples, records, documents). Firstly, for implementing this system, there will be user information and we will use a data set in configuration file format for labeling. The data set will contain set of labels, instances, labeling properties of instances and users information. We will use random labeling mechanism that users will label instances randomly. A user can label an instance more than once. The instances can be labeled by different users with different labels. The system should provide opportunity for configuring number of users and adding data sets. After necessary operations we will produce an output file and a resulting report that shows us the quality of the data labeling and the quality of the users. The simulation can be stopped at any time so the output file and resulting report should be produced correctly.

Functional Requirements

- **Usability Requirements**

  a. The program starts with reading a configuration file that includes user information and data sets.
  b. The starting data set will be set at the beginning of the program,
  c. The data set should be readable and easily accessible.
  d. More users or data sets can be added to configuration file.
  e. The user information will be predefined in configuration file.
  f. The instances will be assigned to users randomly but in some cases previously labeled instances can be assigned the same user which labeled them before.
  g. Labeling is randomly chosen.
  h. The final label for instance can change at each label assignment.
  i. Each label assignments will be collected and used for calculating metrics.
  j. The actions should be logged and printed on the command line.
  k. The output file and the resulting report will be recorded in json files and also can be seen on the command line.

  l. The output file and resulting report should be produced correctly when the program is terminated.
  m. The program will be work in English language.
  n. Type of error handling will be error message.

- **Implementation Requirements**

  a. The program will be written in Java with Object Oriented Programming concept.
  b. The program will take a configuration file as input
  c. The input file will be in json file format.
  d. There will be no database integrity for this project

- **Physical Requirements**
  a. The program can run without login.
  b. The program can be configured for adding users or data sets by the customers.
  c. The program doesn’t work on web browser.
  d. The program should be light and easy to operate through console.

Non-Functional Requirements

- **Performance Requirements**

  a. The program should give the correct results for different cases as excepted. The program should response quickly.
  b. After configurations program should run without any error.
  c. If any problem or exception detected, it should be handled in a short time, and recover the system.
  d. Memory usage should be organized well and unnecessary parts in the code should be eliminated so the program can run faster.

- **Supportability Requirements**

  a. The program should be extensible and changeable according to customer's expectations.
  b. The program should be adaptive for different kind of situations.
  c. The program should have compatibility for different Operating Systems.
  d. The program should be testable and if any error or bug is detected in the system, it should be caught in test method and fixed.
  e. The program should be reusable.

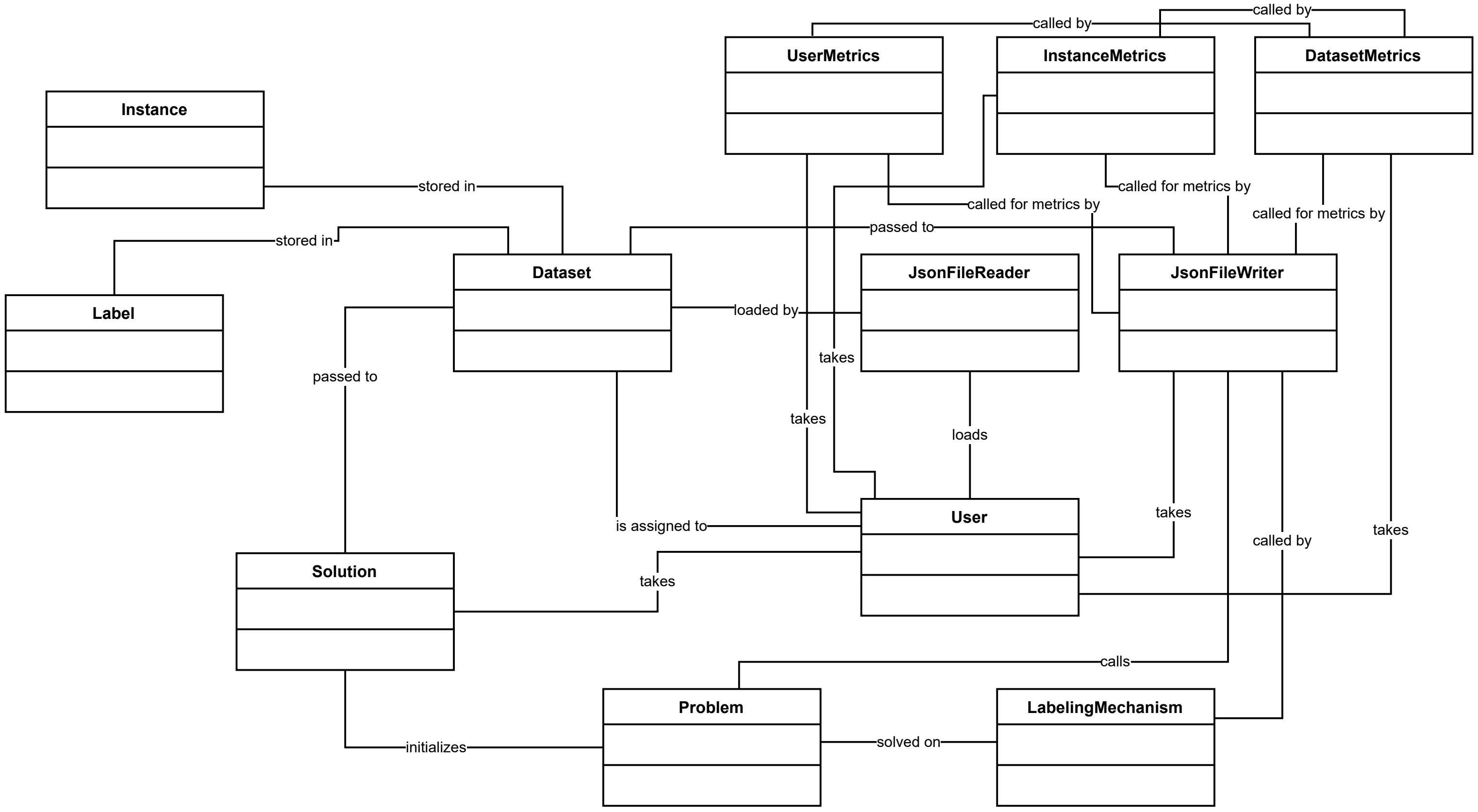
- **Reliability Requirements**

  a. The program should perform as expected from command line.
  b. The program should produce results as accurate as possible.
  c. The program shouldn't lag or throw error halfway of the operation.

Use – Case

Firstly, the program will read a configuration file as a data set and there will be user information in the file. Users will assign labels to instances randomly that are taken from data set. Users will have a possibility value. According to that value, user can label previous instances which are already labeled by the same user. These previous instances also will be selected randomly then shown to user. Program will collect statistics about these assignments and use them for calculations about performance metrics. Customer can stop the program at any time. When the program is stopped, it will give an output file for assignments and a resulting report for performance metrics. At the beginning of the each run, customer can change number of users and possibility value of them and also add different data sets for using. After doing these configurations program will run without any problem and produce results correctly.

Domain Model



Glossary

NLP – Natural Language Processing
Sentiment Analysis – Understanding human language's emotion
Label – Categorizing human language
Console – Command Line or a black window that is seen in windows/linux operating system. User needs to type through keyboard to provide necessary instruction to the program instead of normal mouse click and navigation button.
Json – JSON stands for JavaScript Object Notation. JSON is a lightweight format for storing and transporting data