Project Data Labeling

Requirement Analysis Document

This Project is dedicated to build a data labeling software. Data Labeling can be defined as a process of describing and associating some texts (instances) with single or couple of words(labels) so that they can be classified and analyzed to use.

Functional Requirements

System can read a digital file entered and can parse the information included in it.

System is supposed to allow users to associate instances with labels which are entered to system in a digital file.

System can make a random selection from labels included in dataset and associate selected label with an instance.

System can produce an output file including labeled texts and their labels

System collect statistics for users.

System compare users in the context of a particular dataset or globally.

System calculate metrics for instances in the dataset that are labeled with many users.

Non-Functional Requirements

System must be user friendly, easy to use and secure.

The result of the system give an idea about the quality of the data labeling and the quality of the users.

Glossary

Instance:

An instance represents a piece of text which can consist of either a single word or couple of sentences. Instances take place in a dataset which is entered to the system as a file. Instances are labeled by Users.

Label:

A label is one or couple of words that will describe an instance or it’s feature briefly in a point of view. One or more labels can be associated with an instance.

Dataset:

A dataset is a group of information including instances and labels can be used for these instances. A dataset also includes information of maximum number of labels that can be associated with a single instance. Datasets are entered to system in digital files.

Label Assignment:

Label Assignment represents an associating activity of one or more labels with an instance.

Maximum Number Of Labels:

Maximum number of labels is the answer to the question such that how many labels can be associated with a single instance at maximum.

User:

A user is a person or a mechanism that will associate labels and instances in the dataset which is entered to the system.

Final Label:

If an instance is labeled more than once (by the same user or different users) then assign the most frequent class label as its final label

Labeling System:

The system which labeling process has been done.

Labeling Mission:

Labeling of the users entered into the system with the datasets entered into the system.

Dataset Performance Metrics:

Statistical data of the Dataset's information.

Instance Performance Metrics:
Data for the performance of the Instance.

User Performance Metrics:

Data for the performance of the Users.

Consistency Percentage:

e.g. %60 of the recurrent instances are labeled with the same class.

Entropy:

The number of unique labels in this particular instance as the log base.

Completeness Percentage:

What percentage of the instances are labeled.

Domain Model