Tags Prediction of Text Questions from StackExchange

Po-Ning Tseng, Li-Yuan Hung, Tai-Lun Tseng, Yu-chun Huang
b00902067, b99609035, r01922094, d98922047

Description

The goal of this project is to predict tags of text questions from StackExchange[2]. The 7-GB dataset is provided by StackExchange in the competition held by Facebook on Kaggle.com[1]. Each record in the training data contains 4 columns: Id, Title, Body and Tags. The prediction of tags on test data should only be based on the given Id, Title and Body but not any snooping on the corresponding question on StackExchange.

Figure 1: Stack Exchange Logo

Motivation

The problem of tagging StackExchange is definitely a big data problem since it matches the 3V priciples of big data:

1. *Volume.* 7.6 million questions and 13.6 million answers.
2. *Velocity.* Over 64 million monthly unique visitors
3. *Variety.* 110 sites ranging from technology to art.

In this competition, we only deal with a subset of the data, so the traditional algorithm is still applicable. But in reality it may be infeasible or not cost-effective. Due to these reasons we want to develop not only a good algorithm but also a scalable implementation in terms of big data.

Evaluation Criteria

The performance of this project will be evaluated in F1 score and the time spent on training the tagging model. The rank of our team after each submission will also be recorded as a reference.

Functionality

Part 1: algorithm

Machine learning, natural language processing, statistical inference or language models are possible techniques under our current survey. Hopefully the algorithm will select the best hypothesis from the hypothesis set with the training data and the ten-fold validation set, and can correctly predict the number and name of the tags in the test data.

Part 2: parallelization

We will refactor the program developed above in order to fit into the MapReduce or Spark model. We plan to use both Hadoop and Spark for acceleration depending on the differnt properties of the algorithm. The environment settings will be well configured by our knowledge of the dataset and algorithms in order to reach the best parallelism and performance.

Possible Challenges

The following list is the possible challenges while designing and implementing the tag-prediction model:

1. The algorithm to successfully predict tag number and name may already be hard enough.
2. The training corpus is limited to what is provided by the competition; internet crawling is not allowed.
3. The data source contains corrupted records.
4. Spark is a new programming model; the documentation is not abundant, thus increasing the difficulty on debugging and testing.
5. It takes much time and efforts for team members to get familiar with MapReduce and Spark.

Schedule

1. Dataset format analysis and pre-processing (11/4 - 11/10).
2. Literature survey (11/11 - 12/01).
3. Implementation (11/25 - 12/15).
4. Beta test (12/16 - 12/22).
5. Project documentation (12/16 - 12/25).

References

[1] http://Kaggle.com. Kaggle: Go from big data to big analytics.

[2] http://StackExchange.com. Stackexchange - free, community-powered q&a.