EURECOM INSTITUTE

Semester Project Proposal

RATING PREDICTION BASED ON TEXT REVIEWS

Supervisor: Prof. Pietro Michiardi
Student: Kehe CAI, Vinh-Tuong MAI
September 17, 2014

1 CONTEXT

Yelp introduced a deep dataset for research-minded academics from their wealth of data (`http://www.yelp.com/dataset_challenge`). This dataset includes business, review, user, tip and check-in data in the form of separate JSON objects. A review object has a rating, review text, and is associated with a specific business id and user id. Our objective is to answer such a question: How and how well can you guess a review's rating from its text alone ?

2 PROBLEM

How to apply natural language processing methods to classify the text reviews and extract the sentiment information as well as the user's level of satisfaction, which are considered as a scalable problem.

3 GOAL

Using machine learning approach to solve real life problem which is a challenge published by Yelp, a US company. Apply MapReduce programming model to implement the source code in order to make it scalable when the dataset is massive.

4 Expected Result

- Open source implementation of code: Implement the project by using some famous models and public libraries to make it satisfy the requirements proposed.
- Critical assessment predict proposed: Propose some methods to evaluate the results in order to know whether the model, the algorithm and the implementation used are good or not.
- Problem statement, the techniques used and main results: Write weekly report to explain the problems we encounter and the solutions to solve them, by which techniques and the results as well as the comparison between the techniques used and others.
- Optional: design and implementation can be adjusted to address large scale problems. If the design is quite promising, the implementation can be extended and reused to tackle other scalable problems. Since those are the real life problems which are very important nowadays, it is necessary to build adaptable systems to deal with them.