Instructions for ACL-2015 Proceedings

Shuwei Zhang
shuweiz@usc.edu

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

1 Project Domain & Goals

2 Related Work

3 Data

In order to build a text-based recommendation system, we need to find a dataset that records reviews and rating toward restaurants. Fortunately, Yelp has established a well-organized streamed database [1] that provides reviews and ratings from customers, and the information about the restaurants those reviews are about. In this database, all the data is stored in JSON format with consistent schema. The data set that we will mainly work on is the review dataset, which includes that following entries: business_id, which we will use the business dataset to match to the restaurant; text, which the review left to this restaurant; and stars, which is the rating left by the reviewer.

The review data set had over 8 million instances, and it is hard to load all the data at once due the limitation of our computation resource. Therefore, we will utilize the business data set and apply some data mining algorithms to split the data set to reduce the size and avoid the biased data set. Afterward, we will shuffle and split the data set into a 4: 1 train and test partition. For the text part, we will first do the data clean to remove the unnecessary characters and do the contractions for the text. Meanwhile, the vocabularies used in the review are not alway correct. Therefore, we will use the spell corrector library in python to correct the spelling and reduce the vocabulary size. Also, another challenge is raised by the length of each review, which varies from 1 word to a long paragraph. Therefore, we will find a threshold to truncate the long review to increase the efficiency of feature extraction.

4 Technical Challenge

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. The Theory of Parsing, Translation and Compiling, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. Publications Manual. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. Computing Reviews, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. Journal of the Association for Computing Machinery, 28(1):114–133.

Dan Gusfield. 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK.

---

[1] Yelp Dataset documentation url: https://www.yelp.com/dataset/documentation/main