CKME136 Project Abstract

Shannon Dyck

May 7, 2016

Which TripAdvisor Reviews Are Actually Helpful?

***Overall Problem***

Online review sites typically get a considerable amount of reviews posted on them daily, but not all reviews are actually helpful to someone else considering making a purchase (in this case a vacation). In some cases, a very positive review might not actually contain any useful details in it to determine if you should book your vacation or not. I would like to study TripAdvisor reviews from a text classification and sentiment angle.

I would like to classify reviews as helpful or unhelpful using text classification that will score a review based on certain text features. This would minimize the number of reviews that would have to be manually classified by site moderators.

***Data***

The TripAdvisor JSON dataset from the following paper;

* **Hongning Wang**, Chi Wang, ChengXiang Zhai and Jiawei Han. *Learning Online Discussion Structures by Conditional Random Fields*. The 34th Annual International ACM SIGIR Conference (SIGIR'2011), P435-444, 2011.

***Techniques***

Text classification/Naïve Bayes -> obtaining a training set by classifying reviews first as helpful or non-helpful based on details contained within the review (does it contain the location names, specific names of services, etc). Training a model and then sending live reviews to be classified.

***Tools***

Using Python to build a classification model in combination with Hadoop and specifically Spark to read in the JSON files, parse and analyse.