01/23/2018

Second Measure Data Science Write-up – Ramin Tavassoli

Project Synopsis:

Stackoverflow.com is a collaborative platform for questions and answers on computer science topics. It provides an anonymized data dump on which I performed wrangling, analysis and machine learning using Spark with Scala. The goal was to predict the tag of a question from its body text.

The dump was comprised of three folders, users, posts and votes, each populated with chunked and gzipped XML files. Each XML row included elements corresponding to the respective schema of the three categories. I wrote a function which discarded malformatted XML rows, parsed the remaining rows and retuned RDDs of users, posts and votes. To test the integrity of the RDDs, I formed the hypothesis that a post with a higher favorite count has a higher upvote to downvote ratio. Taking the favorite count of the posts as the key, I summed the number of upvotes and downvotes and determined the ratio. The results corroborated my hypothesis.

To predict the tag from the body of a question, instead of training a multi-label classifier, I treated the problem as one vs. all such that I trained 100 logistic regression classifiers for the top 100 occurring tags in the posts RDD. After finding this tag set, I tokenized the body texts and used a hashing transformer to map the tokens to their term frequencies (TF). Using this feature set and associated tag labels, I trained the classifiers. I did the same pre-processing steps to the test set and predicted tags with 78% accuracy.

Discussion:

Deep learning and KNN would be excellent improvements by treating the problem as multi-label. The project meant to be an intermediate introduction to the Spark with Scala ecosystem and its emphasis on modeling optimization was secondary. I am putting together an IPython notebook for this project under the name “StackOverFlow Spark” on my GitHub page by 01/24/2018.