Model Description

I use Content-Boosted Collaborative Filtering to build the model.

In the content-boosted part,

1. I generate business-profile based on the keywords of businesses' categories in "business.json."
2. I compare the Jaccard Similarities between each business pair and generate similarity-vector of each business.
3. I use similarity-vector of each business and rating-vector of each user to generate a new rating score for some business, so the training data are augmented.

In the CF part, I implement ALS Algorithm in Spark's mllib library to get the model. The model is trained on the augmented training data.

In the prediction part, I weight the prediction from ALS prediction and augmented training data. For the users and businesses that don't appear in the training data, I use the average business rating in "business_avg.json."