Data-intensive Computing | Exercise 2

Markus Kiesel | 1228952

May 26, 2021

1 RDDs

When we compare the output of two implementations we notice that most of top selected tokens are the same. Especially the first ranking tokens by each category match exactly. But the computed chi-square values differ in their magnitude which leads to differences in the ranking of the tokens. The differences most likely result from the deprecated scala JSON parser and the split on the regex.

2 Datasets/DataFrames: Spark ML and Pipelines

Although the process of selecting the "most relevant" tokens is similar the process can not be compared directly. Further, for part 1 the output_rdd.txt contained 2595 distinct tokens in the last line while in part 2 we selected the top 4000 values. Over 75% of the tokens selected with the RDDs approach are the same as for the spark.ml approach which shows that we can archive similar results while using the DataFrames is much more readable and understandable.

3 Text Classification

For the last part a text classification setup was implemented to classify a given category by the review text. The implementation uses a TrainValidation split for Hyperparameter optimization with feasible parameters provided in a ParameterGrid.

The overall F1 measure on the test set (train-test-split of 0.9 to 0.1) is 0.7228 which is rather exceeding my expectations.

Unfortunately I was not able to extract the best model after loading the model from the file again which means I am not able to make any explanations which parameters lead to this result. Further, I did not find a way to extract the F1 measures by each category.