Kickstarter Dataset

Scala Files

I have totally five scala files, which you should run in the following order.

1. Kickstarter_data_ingest.scala
2. Kickstarter_etl.scala
3. Kickstarter_profiling.scala
4. Kickstarter_act_rem_code.scala
5. Kickstarter_app_code.scala

Under the data_ingest folder, I have two more Python files which are used to scrape more campaign information from the Kickstarter website.

Details of Scala Files

1. Kickstarter_data_ingest.scala

   It is used to read the dataset from HDFS to Spark.

2. Kickstarter_etl.scala

   After reading the file, I will start to parse the data in a json format to a dataframe.

3. Kickstarter_profiling.scala

   The files will show the summary of each feature of this dataset, such as the maximal length of a string-type feature and the range of an integer-type variable.

4. Kickstarter_act_rem_code.scala

   In this file, I explore each feature and use Tableau at the same time to do feature engineering before I build the model.

5. Kickstarter_app_code.scala

   I set up the pipeline of encoding, scaling, machine learning models and cross validation techniques in this file.

Dataset

I put my dataset, Kickstarter_20191017.json in the HDFS. By reading the dataset from HDFS to spark, I will run Kickstarter_data_ingest.scala. The file path is:

/user/kll482/final*project/Kickstarter*20191017.json