Write a program for PySpark according to your variant, utilizing a dataset with tweet data. Link to dataset description and download is provided here: https://www.kaggle.com/rgupta09/world-cup-2018-tweets. Alternatively, you can select your own dataset (for example, taken from Kaggle website), but make sure that it contains the features necessary for task completion.

To complete the task, you need to get acquainted with .csv-file load parameters in PySpark, and also analyze the contents of the file. To successfully defend the lab work, it is strongly recommended to use the RDD operations described in book "Learning Spark. Lightning-fast data analysis" (chapter 2). Make a report describing the steps necessary for task completion and provide screenshots of program results.

The task variant is chosen according to your number in the journal.

Variant 1

Get the top-10 most used hashtags. Get top-10 city capitals by the number of tweets. The list of capitals must be stored in a separate file. Determine the most used hashtag for every capital of top-10. Determine which capitals have the most used tweets out of top-10 hashtags.

Variant 2

Get five most mentioned users. For each of these users: (1) Find 10 most used hashtags with these mentions (hashtag list). (2) Find a user with the most difference in hashtags. (3) Find two users which have the closest hashtag occurence distribution rate within the top-10 list.

Variant 3

Select five different time intervals within the dataset. Determine 10 most mentioned hashtags for each interval. Determine which of these hashtags are unique for each time interval. Determine which 2 of the intervals have the closest hashtag occurence distribution rate within the top-10 list.

Variant 4

Get top-20 users by the amount of followers, by the amount of friends and 20 most mentioned users. Determine top-5 distinct hashtags for each group. Determine which users are contained within multiple of those 3 lists, and which users are contained in just a single list.

Variant 5

Create a separate file with a list of capitals. Among the tweets sent from capitals, make 10 random samples of arbitrary size, containing users and hashtags. Get two samples with the highest and lowest amount of unique usernames. Get two samples with the highest and lowest amount of the unique hashtags. Find two samples which have the closest hashtag occurence distribution rate within the top-5 list.