Напишите программу на PySpark согласно вашему варианту для датасета с данными твитов. Ссылка на датасет и его описание: https://www.kaggle.com/rgupta09/world-cup-2018-tweets. Альтернативно, вы можете выбрать другой датасет с твитами самостоятельно (например, с этого же веб-сайта kaggle.com), но убедитесь, что он содержит признаки, необходимые для выполнения задания.

Для выполнения задания необходимо освоиться с параметрами загрузки .csv-файла в Pyspark, а также проанализировать содержимое файла с данными. Для успешной сдачи лабораторной работы настоятельно рекомендуется воспользоваться операциями работы с RDD, описанными в книге «Изучаем Spark. Молниеносный анализ данных -2015» (стр. 42). Необходимо оформить отчёт, в котором описываются шаги выполнения работы и представлены скриншоты работы программы.

Варианты задания выбираются согласно вашему номеру журнала.

Вариант 1

Получить десять наиболее упоминаемых хештегов. Получить десять столиц государств, из которых наиболее часто посылались твиты. Список столиц нужно хранить в отдельном файле. Определить наиболее часто употребляющийся хештег для каждой столицы из топ-10. Определить столицы, в которых преобладающие хештеги из десяти наиболее упоминаемых.

Вариант 2

Получить пять пользователей, которых упоминали в твитах чаще всего. Для каждого из этих пользователей определить десять наиболее часто употребляемых вместе с ними хештегов. Выбрать из них пользователя, у которого хештеги наиболее отличается от четырёх остальных. Выбрать два пользователя, у которых наиболее близкое друг к другу соотношение частоты встречаемости выделенных хештегов.

Вариант 3

Выбрать в датасете пять временных промежутков. Определить десять наиболее упоминаемых хештегов для каждого выбранного временного промежутка. Определить, какие из выделенных хештегов уникальны для каждого промежутка. Определить, для каких двух промежутков наиболее близкое друг к другу соотношение частоты встречаемости хештегов.

Вариант 4

Получить двадцать пользователей с наибольшим количеством фолловеров, двадцать пользователей с наибольшим количеством друзей, двадцать наиболее часто упоминаемых пользователей. Определите топ-5 уникальных хештегов для каждой из выделенных групп. Выделить пользователей, которые попадают в несколько из этих списков, а также пользователей, которые попали только в один список.

Вариант 5

Создать отдельный файл со списком столиц государств. Среди твитов, которые отправлялись из столиц государств, сделать десять случайных выборок произвольного объёма. Определить две выборки, в которых наибольший и наименьший разброс по количеству уникальных пользователей. Определить две выборки, в которых наибольший и наименьший разброс по количеству уникальных хештегов. Определить две выборки с наиболее близким друг к другу соотношением частоты встречаемости хештегов из топ-5 по выборке.