Задание

Найти иностранного пользователя, сообщения которого набрали больше всего ретвитов.

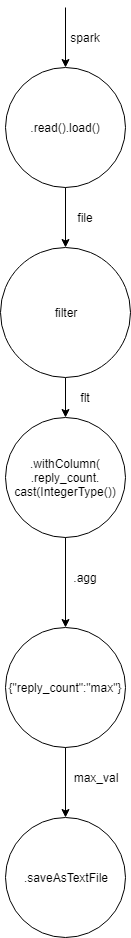
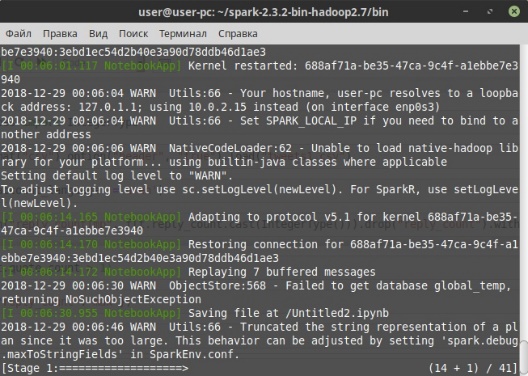


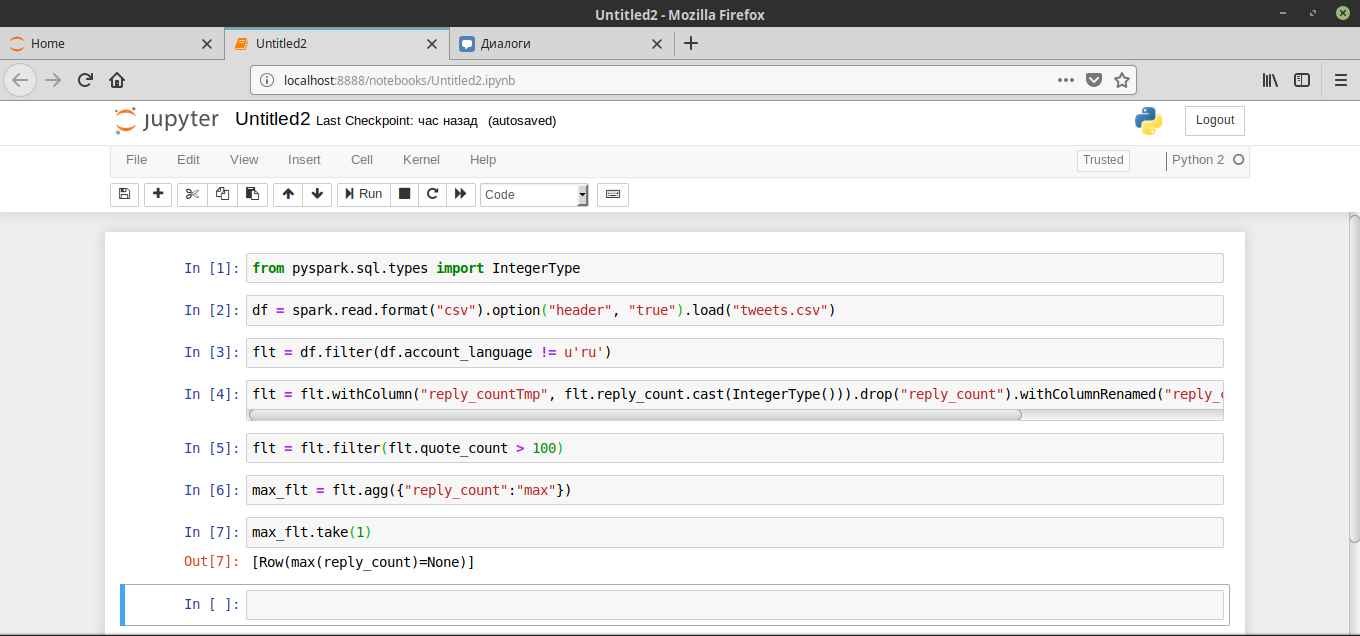
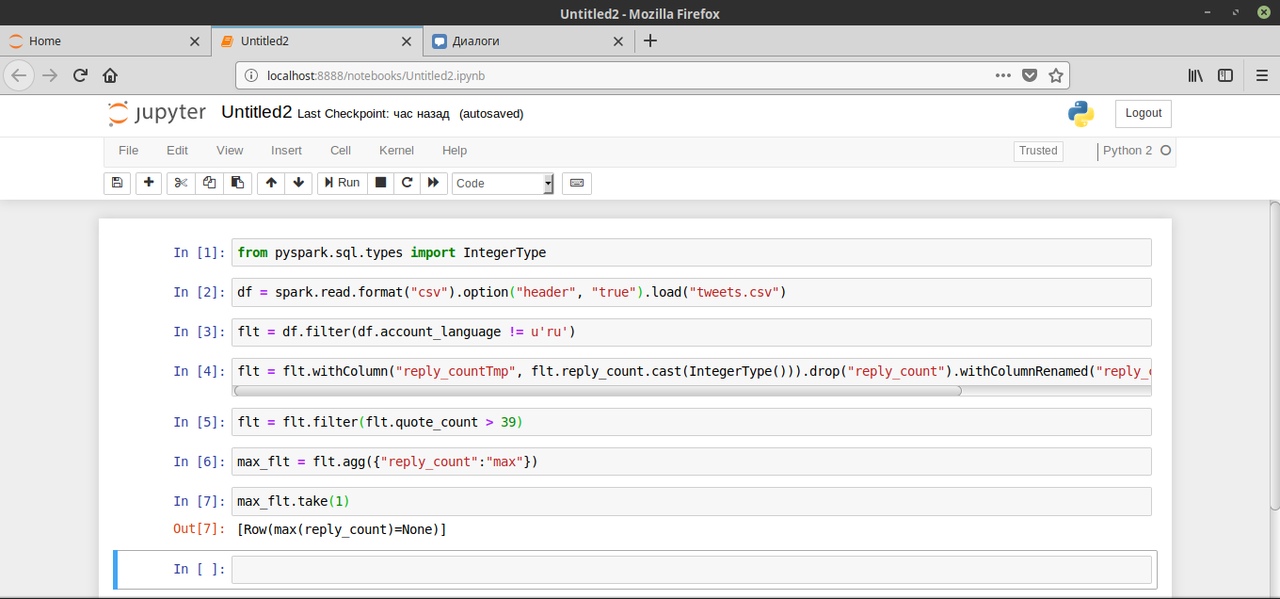
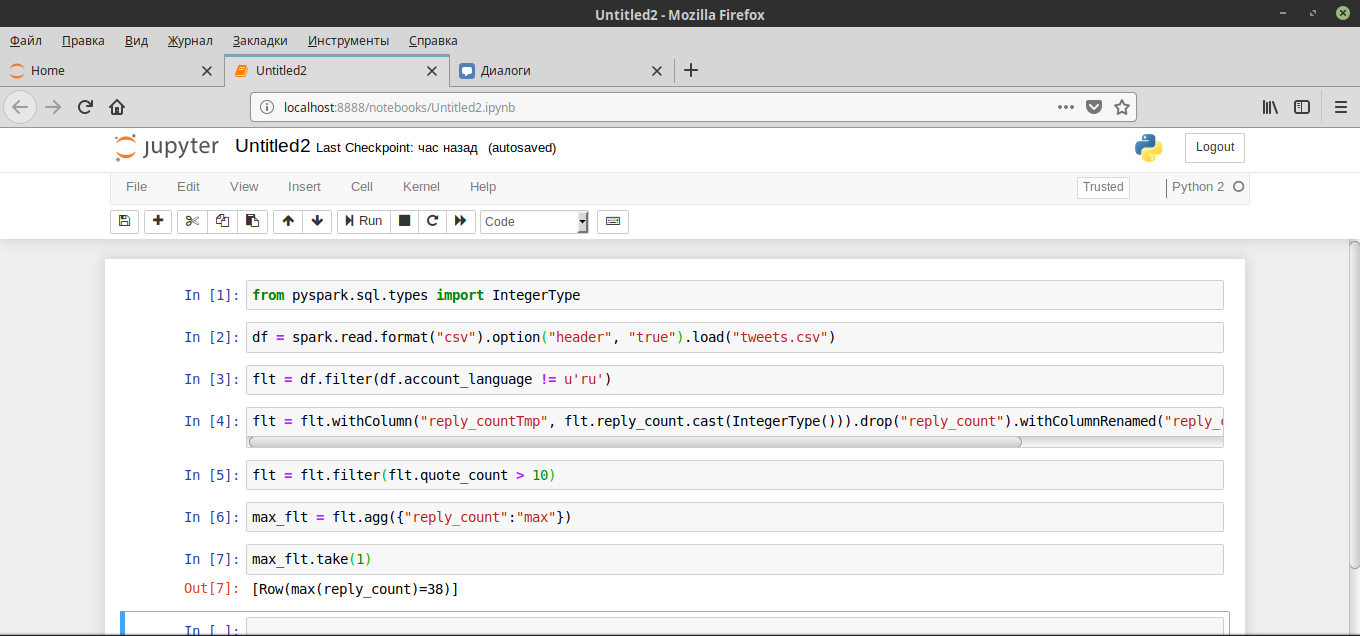
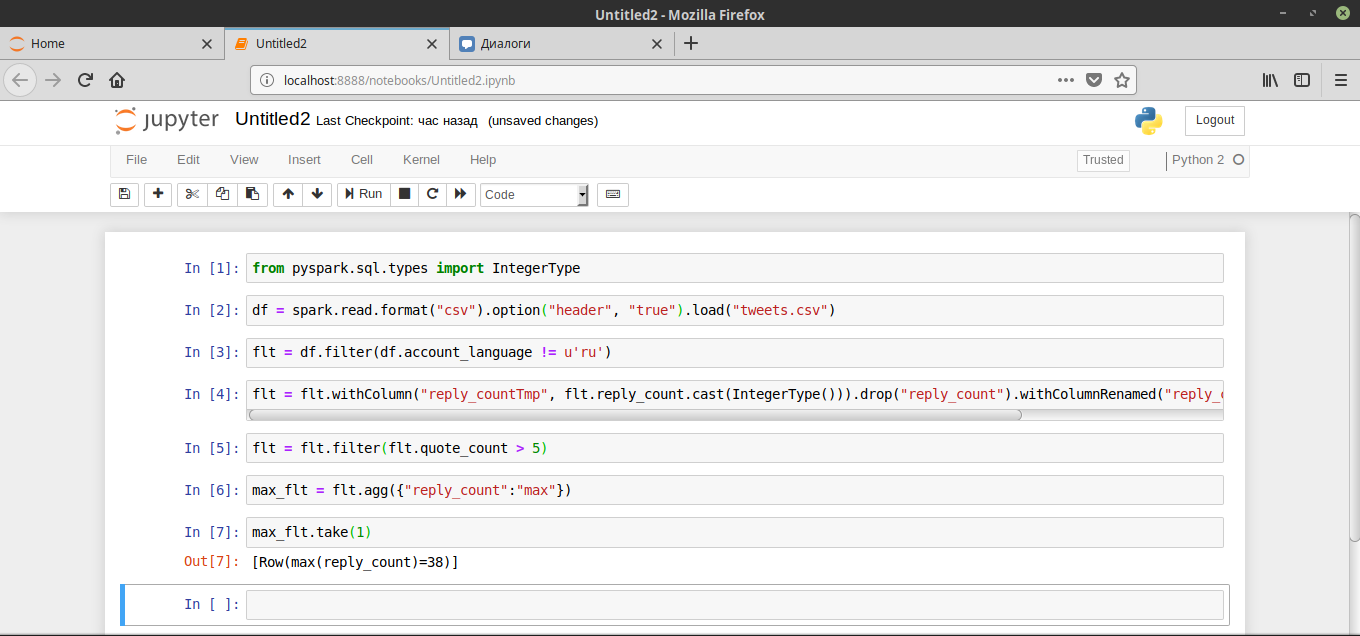
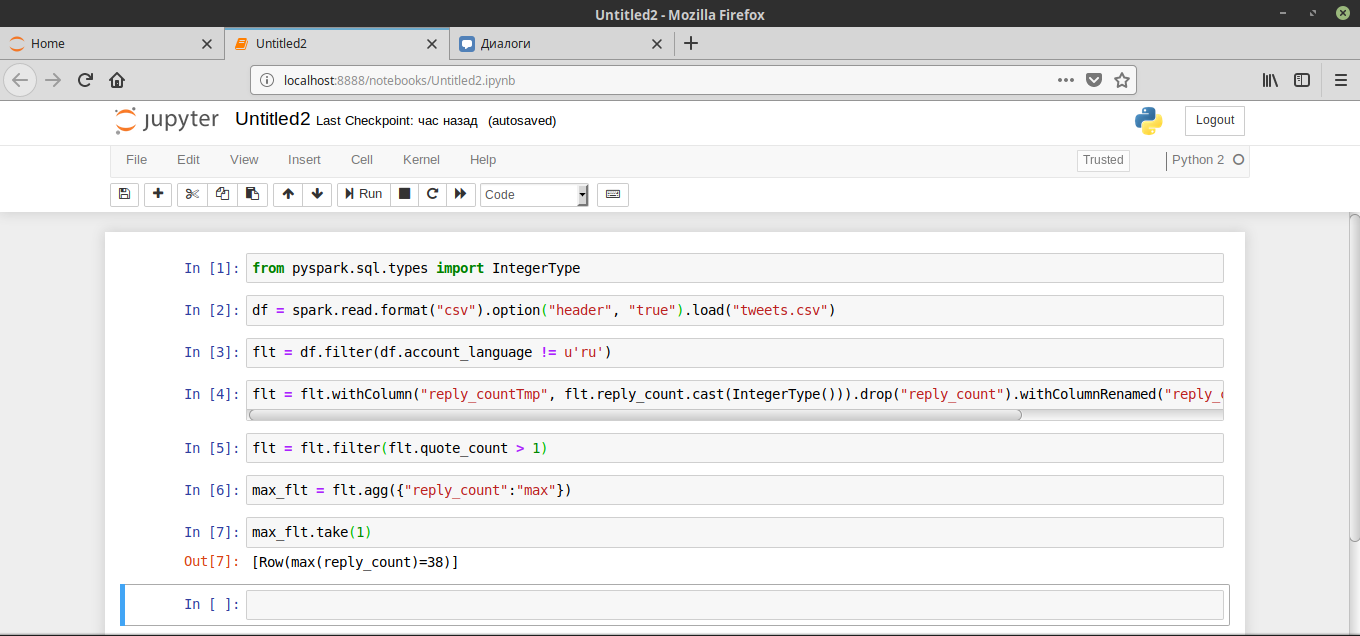
Рисунок 1 Граф потока данных

Примеры выполнения программы

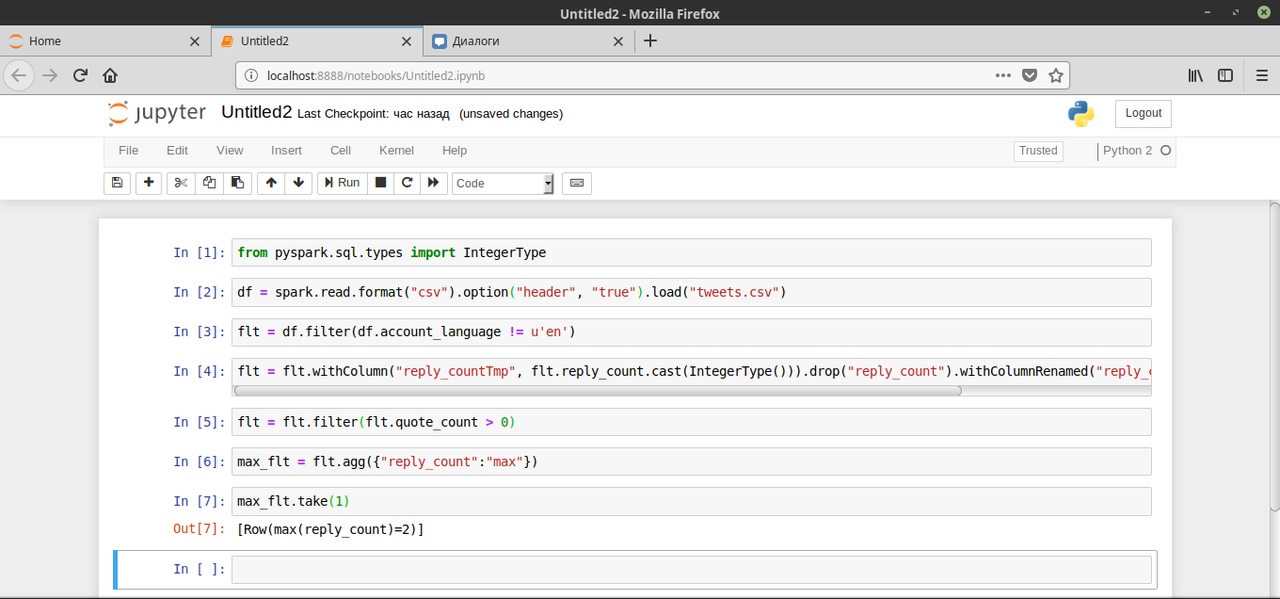
Лог спарка



Различные тесты на поиск максимального ретвита, если ищем не иностранного пользователя с условием что DataFrame содержит не все твиты с оригинальной БД, а только часть из них. Файл tweets.csv



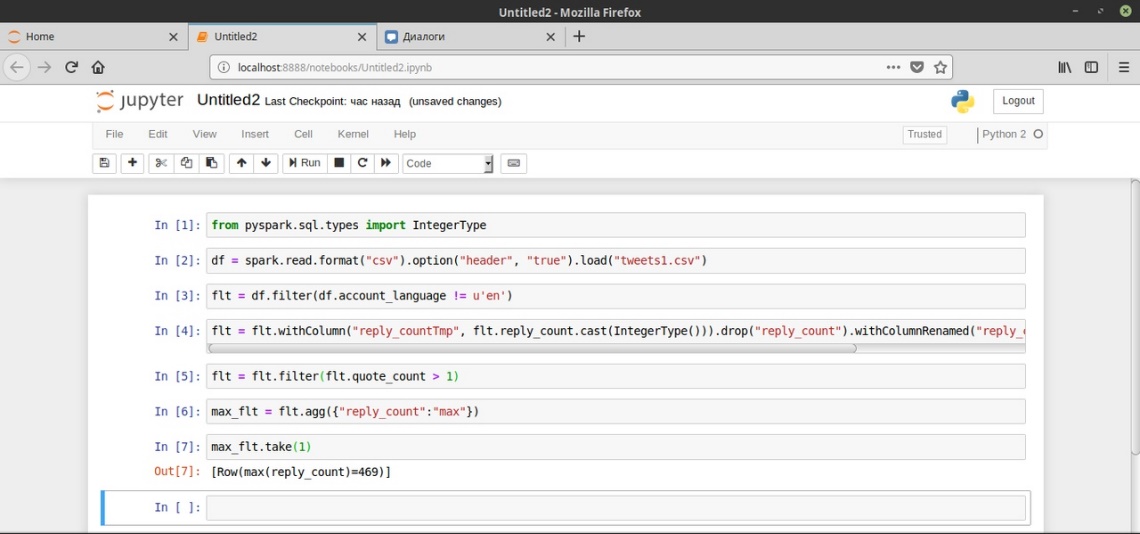
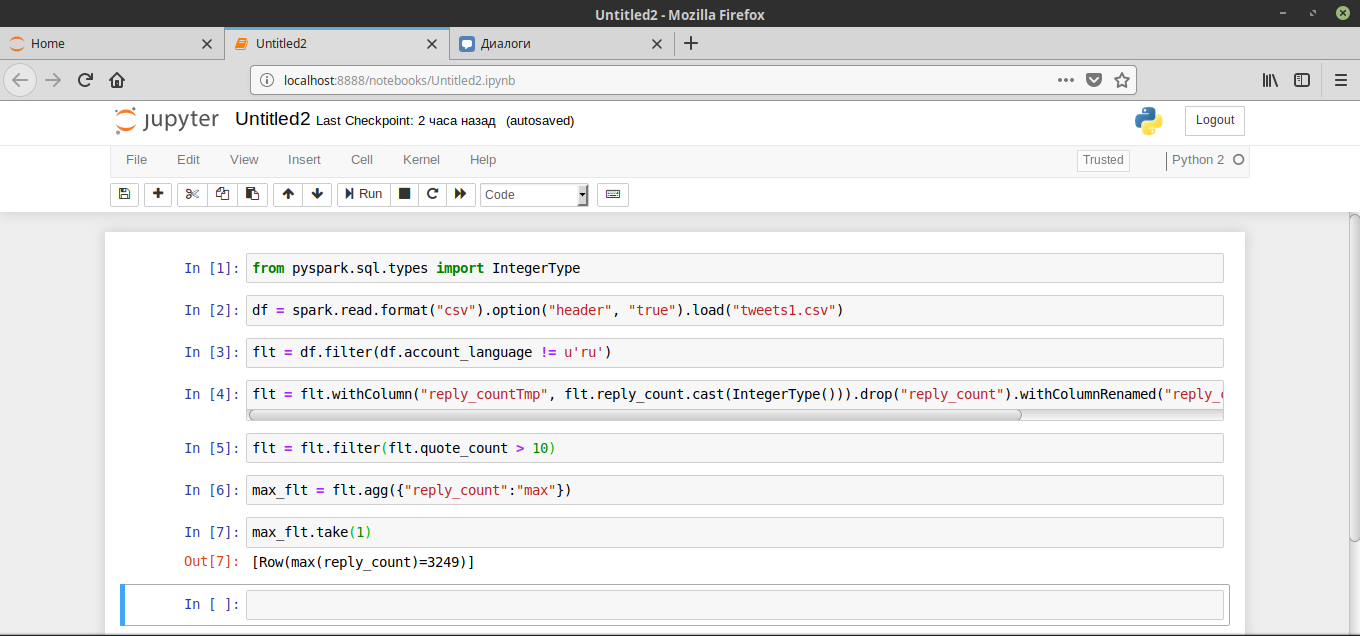
Результат. Максимальный ретвит 38. Если сделать ограничение, что твитов должно быть не меньше 100 или 39 для выборки, то результат будет 0, так как максимальное значение ответов после решения было 38.



Проведем проверку, изменив язык аккаунта на русский.

Результат работы: Максимальное количество ответов 2.

Заменим DataFrame исходным, содержащим все данные, и посчитаем максимальное количество ретвитов для иностранного и неиностранного пользователям. Файл tweets1.csv. Результат работы 3249 для иностранных и 469 неиностранных



Приложение

Полный DataFrame. Файл tweets1.csv.Выборка иностранных пользователей.

from pyspark.sql.types import IntegerType

df = spark.read.format("csv").option("header", "true").load("tweets1.csv")

flt = df.filter(df.account\_language != u'ru')

flt = flt.withColumn("reply\_countTmp", flt.reply\_count.cast(IntegerType())).drop("reply\_count").withColumnRenamed("reply\_countTmp", "reply\_count")

flt = flt.filter(flt.quote\_count > 10)

max\_flt = flt.agg({"reply\_count":"max"})

max\_flt.take(1)