ID2221 – Lab 3
Spark Streaming, Kafka and Cassandra

Dmytro Siniukov siniukov@kth.se 951228-2436
Miquel Larsson miquell@kth.se 920614-5998

October 7, 2018

1. **Introduction**

   In this lab, we are given the task of implementation a Spark Streaming application that reads streaming data in Kafka (in a key, value format), and stores the average of each key in Cassandra, calculating and updating this value continuously.

2. **Code explanation**

   The code ("KafkaSpark.scala") is pretty straightforward to understand. The main function of the KafkaSpark object:

   - Initializes the Cassandra Scheme and table (if it does not exist yet) in order to store the results.
   - Connects to Kafka in order to read the stream of data generated by the generator script.
   - Does a mapping word $->$ count.
   - Averages up all the counts by words.
   - Records the results to Cassandra.

3. **Running the script**

   In order to test the solution:

   (1) Create the Kafka topic "avg".

   (2) Launch kafka and cassandra (as it is mentioned in the assignment description).

   (3) Execute "`sbt run`" in both generator and SparkStreaming folders in parallel.

   (4) Check the content of Cassandra table "avg_space.avg", by entering the Cassandra command line and doing: `use avg_space;` and then `select * from avg;`

4. **Results**

   We got the following results after executing the generator for about 30 seconds:

```
cqlsh:avg_space> select * from avg;

 word | count
------+----------
    z | 15.90423
    a | 15.91878
    c | 15.93994
    m | 15.94357
    f | 15.34126
    o | 15.93955
    n | 15.91628
    q | 15.83854
    g | 15.91728
    p | 15.90199
    e | 15.83402
    r | 15.78686
    d | 15.35929
    h | 15.78637
    w | 15.94006
    l |  15.9487
    j | 15.92539
    v | 15.93233
    y |  15.9392
    u | 15.99458
    i | 15.97127
    k | 15.91599
    t | 15.98311
    x | 15.93112
    b | 15.82516
    s | 15.88258

(26 rows)
```