University of Augsburg, Institute of Computer Science SS 2019
Prof. Dr. P. Fischer 24. May 2019
J. Kastner, L. Rudenko Sheet 5

Analyzing Massive Data Sets

Exercise 1: Spark RDDs (homework)

Let us go back to the Sheet 1, Exercise 3. We had performed some basic data analysis using Pandas which provides Dataframes and SQL-Like operations for Python. As a dataset we used a part of the TPC-H benchmark. Now you should express the same queries with Spark's RDDs (not Dataframe/Datasets - see the next exercise).

a) Find 25 suppliers with the lowest account balance.

b) How many suppliers have a positive account balance?

c) Find all brands produced by the same manufacturer and calculate the number of items as well as the total sales price for each brand of each manufacturer.

d) How many items have 3 words in their name?

e) How many different items does each supplier have?

Can you express all requests with the RDDs? Do you think there is a better way to express these queries with Spark?

Exercise 2: Spark DataFrames (live)

In this Exercise you should get to know another concept of Spark - **DataFrames**. Solve the subtasks a)-e) from Exercise 1 of this sheet with the Spark's DataFrames this time.

You can test your solutions using Databricks **Community** Edition (to register, follow this link Databricks). The documentation for Databricks can be found here. The part about Data Loading could be particularly interesting.

Other useful resources are the Programming Guide and the API Reference.

Exercise 3: MinHashing (live)

The following documents are given:

- D_1 : "your new red bag lies on my table"
- D_2 : "my new red cat lies on my table"
- D_3 : "my old red bag lies on my table"

a) create the boolean input matrix with **columns = documents** and **rows = words**. To simplify the exercise, use each word individually (no k-grams/shingles) and do not remove any stop words. The words should not be numbered or hashed.

b) compute the signature matrix M using the following set of permutations Π:

- $\{1, 7, 9, 2, 3, 10, 8, 5, 6, 4\}$
- $\{3, 10, 1, 6, 9, 5, 2, 4, 8, 7\}$
- $\{1, 9, 6, 3, 5, 8, 2, 4, 7, 10\}$
- $\{5, 6, 4, 7, 10, 8, 3, 9, 2, 1\}$

c) Compute the Jaccard similarities for all document pairs using once columns from the **input matrix** and once columns from the **signature matrix**. Compare and assess the results.

d) Do the subtasks a)-c) again. This time don't use the single words, but **shingles/k-grams** with $k = 3words$. Again, use the shingles as is, do not use any hashing/numbering.

The following set of permutations Π is given:

- $\{7, 9, 1, 6, 10, 5, 3, 12, 11, 8, 4, 2\}$
- $\{12, 9, 5, 6, 2, 1, 7, 10, 3, 8, 11, 4\}$
- $\{3, 1, 9, 5, 10, 4, 6, 7, 8, 11, 12, 2\}$
- $\{10, 12, 2, 9, 4, 6, 8, 1, 3, 5, 7, 11\}$

What do you notice when you compare the results for words and shingles?