Heidelberg University
Winter semester 2019/2020

Group Parallel and Distributed Systems (PVS)
Artur Andrzejak, Tuyen Le

Problem Set 8 for lecture Mining Massive Datasets

Due December 16, 2019, 11:59 pm

Exercise 1 (1 point)

Prove that for every $N \geq 1$ the following holds (hint: you don't need to manipulate terms, focus on the interpretation of the binomial coefficient):

$$\sum_{k=0}^{N} \binom{N}{k} = 2^N.$$

Exercise 2 (4 points)

Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items {1, 2, 3, 4, 6, 12}, since these are all the integers that divide 12. Answer the following questions (without programming) and explain how you have obtained the solution:

a) If the support threshold is 5, which items are frequent?

b) If the support threshold is 5, which pairs of items are frequent?

c) What is the sum of the sizes of all the baskets?

d) What is the confidence of the following association rules $R_1 = \{5, 7\} \rightarrow 2$ and $R_2 = \{2, 3, 4\} \rightarrow 5$.

Exercise 3 (3 points)

Using the same setup as in the Exercise 1, apply the A-Priori Algorithm ("on paper", i.e. without programming) with a support threshold of 5. Consider itemsets of cardinality $k = 1, 2, 3$ (i.e. frequent items, pairs and triples) and submit as your solution the results of each pass of the algorithm.

Exercise 4 (3 points)

Let there be I items in a market-basket data set of B baskets. Suppose that every basket contains exactly K items. Assuming I, B, and Kas (integer) parameters, answer the following questions:

a) How much space does the triangular-matrix method take to store the counts of all pairs of items, assuming four bytes per array element?

b) What is the largest possible number of pairs with a nonzero count?

c) Under what circumstances can we be certain that the triples method will use less space than the triangular array?

Exercise 5 (5 points)

This exercise is the first in a series of tasks related to processing data with Spark dataframes. Your implementation will be reused in future exercises.

Take a look at a dataset[1] with prices of so-called spot instances from Amazon Elastic Compute Cloud (EC2) service. This dataset contains prices collected for *seven* availability zones, grouped in 28 zipped files. Inside of each compressed file there is a tab-delimited text file with the structure

<Type>\t<Price>\t<Timestamp>\t<InstanceType>\t<ProductDescription>\t<AvailabilityZone>.

a) Implement a subroutine in Python/Spark which takes as an argument a file name of a compressed (*.gz) file with the structure as described above, reads its content into a new Spark dataframe, and returns a reference to this dataframe. To this aim define an appropriate schema with a correct name and (memory-efficient) data type for each column (e.g. use *TimestampType*() and not *StringType*() for column *Timestamp*, see also https://goo.gl/rkxENJ). The resulting dataframe should **not** contain the column *Type* (as each row has the value "SPOTINSTANCEPRICE"). Submit your code as the solution.

b) To test your implementation, read the file *prices-eu-central-1-2019-05-24.txt.gz* into a dataframe, and then calculate and output the average price per each combination *InstanceType*/*ProductDescription*.

[1] Available on Moodle, Weekly problem sets: Datasets/dataset-EC2-series/