Hana Ćatić

Assignment 1

For this assignment, our task was to proccess a CSV file containing about 2.3 million applications on the Google Play Store. We were required to answer a few questions about the file.

The first task was to find the number of apps per category. I expected that task to be the easiest, because reading form a CSV file is very straightforward. We just split the line by commas. However, this file contained many errors and exceptions, so they needed to be handled, e.g. double quotes, empty fields or excess commas. I used Pattern and Matcher to handle those situations, then extracted the third element of the array for each line and stored it in a map, counting the number of occurrences of each category.

The second task was to find the top 100 companies with the most apps on the Play Store. That problem was solved by splitting the app ID by dots and extracting the first two words e.g. *com.google*, and then sorting the entries by the number of occurrences.

The third task was finding the top 3 developers with the biggest number of developed applications, but which do not work for the company which released the app. The developer's email was the key to solving this problem, because the domain showed us whether they were an employee of the company or not. If the domain didn't contain the company's name, the developer was stored in a map which counts the number of applications they developed.

The fourth and fifth task were the easiest - counting the number of applications we could buy given a certain budget and counting the number of free vs paid apps. I solved that using simple loops and conditions. The only thing I was not expecting was that I would have to handle special characters (like "+" or double quotes) in the installs column and the fact that I had to use the *Long* datatype because the number of installs for certain apps was too large for the *Integer* datatype.

If I were to do the assignment again, I would spend more time trying to figure out the regex for splitting the lines on my own, because I found that to be the most difficult part of the tasks. Maybe I would even separate the part of code where I scanned the file into a distinct function.