School of Computer Science

COMP30770

Project 2
Spark

Teaching Assistant:	Thomas Laurent
Coordinator:	Anthony Ventresque
Date:	Friday 19th February, 2021
Total Number of Pages:	3

General Instructions

- You are encouraged to collaborate with your peers on this project, but all written work must be your own. In particular we expect you to be able to explain every aspect of your solutions if asked.
- We ask you to hand in an archive (zip or tar.gz) of your solution: code/scripts, README.txt file describing how to run your programs, a short pdf report of your work (do not include your code in it but explain what you did).
- The report should list your answers and should also contain a short introduction and conclusion. The report should also contain (see Section 3 of this document): a short (1 or 2 pages) description of one of the research papers you can find on Brightspace.
- The report should not be longer than 10 pages.
- The breakdown of marks for the project will be as follows:
  - Exercise 1: 30%
  - Exercise 2: 35%
  - Exercise 3: 35%
- **Due date: 07/03/2021**

1 Spark

In this exercise you will explore a small dataset of the 100 most starred Github projects in the "big-data" topic. Download the dataset using `wget`: `http://csserver.ucd.ie/~thomas/github-big-data.csv`. Each line, delimited by commas, contains a project's name, its description, its main language and the number of stars it received.

Launch the Spark shell and then create an RDD or a DataFrame from the input file. For each of the following tasks, write Scala code to solve it. You can use operations on DataFrames (see lab 5 and here, in the Scala API, DataFrame is simply a type alias of Dataset[Row]), spark SQL on Dataframes (see lab 5), or operations on RDDs (see lab 5, here, and here).

1. Determine which project has the most stars. If multiple projects have the same number of stars, list all of them.
2. Compute the total number of stars for each language.
3. (a) Determine the number of project descriptions that contain the word "data".
   (b) Among those, how many have their language value set (not empty/null)?
4. Determine the most frequently used word in the project descriptions.

2 Graph Processing

In this exercise you will explore a dataset extracted from the DBLP co-authorship dataset, representing a graph of scientists connected if they have authored a paper together. This dataset has been built to contain only authors with an Erdős number under a certain value. Download the dataset using `wget`: `http://csserver.ucd.ie/~thomas/dblp_coauthorship.csv`. Each line in the csv file contains a pair of names *author1,author2* that have co-authored a paper together. Each line is mirrored (*author2,author1*) as GraphX only considers directed graphs but this graph is undirected.

1. Write a Spark GraphX program to read the dataset and build a graph that represents it. The `zipWithIndex` function might be useful to create ids for the authors.
2. Write a GraphX program to find what is the maximum Erdős number of authors in the dataset, i.e. the maximum value of the minimum distance between an author and Erdős in the Graph.

   You can use the `ShortestPaths.run()` method to compute the length of the shortest path between Erdős and the author authors. This method returns a graph in which each vertex has the id of an author and contains a map containing the shortest distance to the target (called "landmark") authors.

3 Reflection

Write a short report (1 or 2 pages) on one of the research papers that are available on Brightspace. The following list of sections is an indication of how to write your paper. Some of the items might not be relevant for all papers, and you might want to add some sections in your report (e.g., evaluate the posterity of a solution etc.). We will have an open mind when reading your report and we just want to see how you analyse a research paper and are able to discuss it - in short there is no one single perfect report, everything that shows you made an effort to understand and focus on the important parts (research methodology, hypotheses, etc.) will be welcome.

- Identify the question/challenge the paper addresses. Explain in your own words what the motivation for the research is.
- Describe briefly the related work, i.e., the other (related) solutions that the authors compare themselves to. Show the limitations of these related solutions.
- Give an outline of the solution proposed by the authors (no need to go into details) showing the main components
- describe their scientific method: what are the research questions they evaluate? How do they evaluate?
- Describe briefly their results.
- Give your impression on the idea, what you liked about the paper and whether you see any limitations etc.