Inf553 – Foundations and Applications of Data Mining

Summer 2018 Assignment 2

Deadline: 06/11 2018 11:59 PM PST

1 Overview of the Assignment

This assignment contains one main algorithm. You will implement the SON algorithm using the Apache Spark Framework. You will use three different datasets, ranging from very small to very large. This will help you to test, develop and optimize your algorithm given the number of records at hand. More details on the structure of the datasets and instructions on how to use the input files will be explained in details in the following sections. The goal of this assignment is to make you understand how you can apply the frequent itemset algorithms you have learned in class on a large number of data and more importantly how you can make your implementation more performant and efficient in a distributed environment.

1.1 Environment Requirements

Python: 2.7 Scala: 2.11 Spark: 2.2.1

Student must use python to complete both Task1 and Task2.

There will be 10% bonus if you also use Scala for both Task1 and Task2 (i.e. 10 - 11; 9 - 9.9).

IMPORTANT: We will use these versions to compile and test your code. If you use other versions, there will be a 20% penalty since we will not be able to grade it automatically.

You can only use Spark RDD.

In order for you to understand more deeply of the Spark, use RDD only, you won't get any point if you use Dataframe or Dataset.

1.2 Write your own code!

For this assignment to be an effective learning experience, you must write your own code! I emphasize this point because you will be able to find Python implementations of most or perhaps even all of the required functions on the web. Please do not look for or at any such code!

TA will combine some python code on Github which can be searched by keyword "INF553" and every students' code, using some software tool for detecting Plagiarism.

Do not share code with other students in the class!!

Submission Details

For this assignment you will need to turn in a Python, Java, or Scala program depending on your language of preference. We will test your code using the same datasets but with different support thresholds values. This assignment will surely need some time to be implemented so please plan accordingly and start early!

Your submission must be a .zip file with name: **Firstname_Lastname_hw2.zip**. The structure of your submission should be identical as shown below.

The Firstname_Lastname_Description.pdf file contains helpful instructions on how to run your code along with other necessary information as described in the following sections.

The OutputFiles directory contains the deliverable output files for each problem and the Solution directory contains your source code.

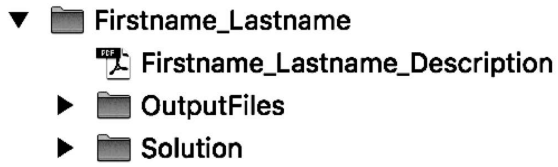

Figure 1: Submission Structure

1.3 SON Algorithm

In this assignment we implement the SON Algorithm to solve every problem (Problems 1 and 3) on top of Apache Spark Framework. We will rely on the fact that SON can process chunks of data in order to identify the frequent itemsets. You will need to find **all the possible combinations of the frequent itemsets** for any given input file that follows the format of the Amazon Review Datasets. In order to accomplish this task, you need to read Chapter 6 from the Mining of Massive Datasets book and concentrate on section 6.4 – Limited-Pass Algorithms. Inside the Firstname_Lastname_Description.pdf file we need you to describe the approach you used for your program. Specifically, in order to process each chunk which algorithm did you use, A-Priori, MultiHash, PCY, etc. . .

At the end of the assignment, Appendix A provides some more guidelines that will help you with the implementation and Appendix B specifies how to organize your Description pdf file.

For assignment 1 you used the Spark framework and most probably at this point you have a better understanding of the MapReduce operations. You can write your program in Python, Java or Scala. For this assignment you will need to find the collection of frequent itemsets of rated products using the MovieLens dataset with which you are already familiar from homework 1.

You will need to compute the frequent itemsets using SON algorithm, initially for a **synthetic testing** dataset (Problem 1) which resembles the MovieLens ratings.csv file, then for the **ml-latest-small/ratings.csv** dataset (Problem 2) and finally for the **ml-20m/ratings.csv** (Problem 3).

The MovieLens datasets can be found in the following link: MovieLens

You will download two datasets. The first one is the ml-20m.zip and the second is the ml-latest-small.zip.

Once you extract the zip archives you will find multiple data files. From those files for this assignment we only need the ratings.csv file from each zip archive.

We would like to compute two cases of possible frequent itemsets using the testing and ratings.csv files.

Case 1

We would like to calculate the combinations of frequent **movies** (as singletons, pairs, triples, etc...) that were **rated** and are **qualified as frequent given a support threshold value**.

In order to apply this computation, we will need to create a basket for each user containing the ids of the movies that were rated by this user. If a movie was rated more than one time from a user, we consider that this movie was rated only once. More specifically, the movie ids are unique within each basket are unique. The generated baskets are similar to:

$$user_1 = (movie_{11}, movie_{12}, movie_{13}, \dots)$$
$$user_2 = (movie_{21}, movie_{22}, movie_{23}, \dots)$$
$$user_3 = (movie_{31}, movie_{32}, movie_{33}, \dots)$$

Case 2

In the second case we want to calculate the combinations of frequent **users** (as singletons, pairs, triples, etc...) who can be **qualified as frequent given a support threshold value**.

In order to apply this computation, we will need to create a basket for each movie containing the ids of the users who rated this movie. If a movie was rated more than one time from a user, we consider it was a rated only once by this user. More specifically, the user ids are unique within each basket. The generated baskets are similar to:

$$movie_1 = (user_{11}, user_{12}, user_{13}, \dots)$$
$$movie_2 = (user_{21}, user_{22}, user_{23}, \dots)$$
$$movie_3 = (user_{31}, user_{32}, user_{33}, \dots)$$

Finally, in the section Problem 1, we will describe explicitly how you should run your program, and what should be the format of your expected output. Everything that is described in section Problem 1 must be applied to the subsequent sections as well (i.e., Problem 2 and Problem 3)

2 Problem 1 (20 Points) Implementing SON using Spark with the Testing Dataset

Under the **/Data** folder of the assignment you will find two small sample datasets. The **Data/Small1.csv** dataset can be used to verify the correctness of your implementation. We will also require you to submit, for each of the two above cases, one output for evaluation for the **Data/Small2.csv** dataset, as described in the following Deliverables section.

2.1 Execution Requirements

2.1.1 Input Arguments

1. Case Number An integer value specifying which case from the ones we just described we want to compute the frequent itemsets. The input is an integer value. **1 for case 1 and 2 for case 2**.

2. Input.csv This is the path to the input ratings file containing all the transactions. Each line corresponds to a transaction. Each transaction has items that are comma separated. For Problem 1 you can use the Data/movies.small1.csv file to test the correctness of your algorithm.

3. Support Integer that defines the minimum count to qualify as a frequent itemset.

2.1.2 Output

A file in the format shown in the snapshot of the Execution Example section below. In particular, for each line you should output the frequent itemsets you found for the current combination followed by an empty line after each combination. The printed itemsets must be sorted in ascending order. A higher level description of this format is:

(frequent_singleton1), (frequent_singleton2), . . . , (frequent_singletonK)
(frequent_pair1), (frequent_pair2), . . . , (frequent_pairM)
(frequent_triple1), (frequent_triple2), . . . , (frequent_tripleN)
. . .

2.2 Execution Example

The first argument passed to our program in the below execution is the case number we want to run the algorithm against. The second input is the path to the ratings input file and the third is the support threshold value. Following we present examples of how you can rsun your program with spark submit both when your application is a Java/Scala program or a Python script.

A. Example of running a Java/Scala application with spark-submit

Notice that the argument class of the spark-submit specifies the main class of your application and it is followed by the jar file of the application.

Please use FrequentItemsetsSON as class name

```
bin/spark-submit --class FrequentItemsetsSON FirstName_LastName_SON.jar <case number> <csv file path> <support>
```

Figure 2: Command Line Format for Scala

```
bin/spark-submit --class FrequentItemsetsSON FirstName_LastName_SON.jar 1 movies.small1.csv 4
```

Figure 3: Command Line example for case1

```
bin/spark-submit --class FrequentItemsetsSON FirstName_LastName_SON.jar 2 movies.small1.csv 8
```

Figure 4: Command Line example for case2

B. Example of running a Python application with spark-submit

```
bin/spark-submit FirstName_LastName_SON.py <case number> <csv file path> <support>
```

Figure 5: Command Line Format for Python

```
bin/spark-submit FirstName_LastName_SON.py 1 movies.small1.csv 4
```

Figure 6: Command Line example for case1

```
bin/spark-submit FirstName_LastName_SON.py 2 movies.small1.csv 8
```

Figure 7: Command Line example for case2

The solution of the above execution for case 1 is similar to the following snapshot. Since both the movie ids and the user ids are integers the format of the output will be the same in both cases:

Solution of A.Case1 and B.Case1 Snaphosts, with input case number 1 , input file Small1.csv and support threshold equal to 4:

```
(97), (98), (99), (100), (101), (102), (103)
(97, 98), (97, 99), (97, 101), (97, 102), (98, 99), (98, 100), (98, 101), (98, 102), (99, 101), (99, 102), (99, 103), (100, 101), (101, 102), (102, 103)
(97, 98, 99), (97, 99, 101), (98, 100, 101), (99, 102, 103)
```

Figure 8: Solution example

2.3 Deliverables for Problem 1

2.3.1 Script or Jar File and Source Code

Please name your Python script as: FirstName_LastName_SON.py.
Or if you submit a jar file as: FirstName_LastName_SON.jar.

The python script or the .jar file of your implementation should be inside the Solution directory of your submission.

2.3.2 Output Files

We need two output files for Problem 1.
For case 1, run your program against Data/Small2.csv dataset with support 3.
For case 2, run your program against Data/Small2.csv dataset with support 5.

The format of the output should be exactly the same as the above snapshot for both cases.
The names of the output files should be as:
FirstName_LastName_SON_Small2.case1-3.txt
FirstName_LastName_SON_Small2.case2-5.txt

You don't need to specify the path of the output file in the commandline, you only need to save the file with the format FirstName_LastName_SON_DATASET.case1or2-SUPPORT.txt in the same path your program run.

The above output files should be placed inside the OutputFiles directory of your submission.

2.3.3 Description

Inside the Firstname_LastName_Description pdf document please write the command line that you used with spark-submit in order to run your code. Specify also the Spark version that you use to write your code. If it is a jar file, please specify the name of the main class of your app as shown in the above snapshots. We will use this in order to rerun your code against different support values if needed.

3 Problem 2 (60 Points) Implementing SON using Spark with the MovieLens Small Dataset

The requirements for Problem 2 are similar to Problem 1. However, here we would like to **check for the performance of our implementation for a larger dataset**. We would like to find the frequent itemsets among a larger number of records. For this purpose, a good indicator of how well our algorithm works is the total execution time. In this execution time **we take into account also the time of reading the files from the disk**. Following, we provide a table of execution time for two threshold values for each case described in the first section. You can use this array as an evaluation metric of your implementation.

CASE 1		CASE 2	
Support Threshold	**Execution Time**	**Support Threshold**	**Execution Time**
120	<100 secs	180	<600 secs
150	<70 secs	200	<500 secs

Figure 9: Time Threshold for Problem 2

3.1 Deliverables for Problem 2

3.1.1 Output Files

We need four output files for Problem 2.
For case 1, run your program against the ml-latest-small/ratings.csv dataset with support 120 and 150.

For case 2, run your program against the ml-latest-small/ratings.csv dataset with support 180 and 200.

The format of the output should be exactly the same as the above.
The names of the output files should be as:
FirstName_LastName_SON_MovieLens.Small.case1-120.txt
FirstName_LastName_SON_MovieLens.Small.case1-150.txt
FirstName_LastName_SON_MovieLens.Small.case2-180.txt
FirstName_LastName_SON_MovieLens.Small.case2-200.txt

You don't need to specify the path of the output file in the commandline, you only need to save the file with the format FirstName_LastName_SON_DATASET.case1or2-SUPPORT.txt in the same path your program run.
You can change the ratings.csv file of the small dataset to MovieLens.Small.csv and use the filename as the DATASET in the output filename

The above output files should be placed inside the OutputFiles directory of your submission.

3.1.2 Description

Inside the Firstname_LastName_Description.pdf document of your submission please include a table that is exactly the same with the one provided on the top of this section. You must use the same support threshold values as the table above and include the execution times of your implementation for each case. We will rerun your code so make sure the times you record on the table are the ones corresponding to your implementation.

Grade breakdown

a. Four correct output files (8pts each)
b. Your execution time needs to be smaller than the ones in the table (7pts each)

4 Problem 3 (20 Points) Implementing SON using Spark with the MovieLens Big Dataset

For the last part of this assignment we will run our application using the big MovieLens dataset, located inside the ml-20m/ratings.csv of the downloaded zip file. Since the purpose of this assignment is to test the efficiency and see how you can optimize your implementation we also provide some execution times

similarly as we did in Problem 2. In this execution time we take into account also the time of reading the files from the disk.

CASE 1		CASE 2	
Support Threshold	**Execution Time**	**Support Threshold**	**Execution Time**
30000	<600secs	2800	<700secs
35000	<450secs	3000	<580secs

Figure 10: Time Threshold for Problem 3

4.1 Deliverables for Problem 3

4.1.1 Output Files

We need four output files for Problem 3.
For case 1, run your program against the ml-20m/ratings.csv dataset with support 30000 and 35000.
For case 2, run your program against the ml-20m/ratings.csv dataset with support 2800 and 3000.

The format of the output should be exactly the same as the above.
The names of the output files should be as:
FirstName_LastName_SON_MovieLens.Big.case1-30000.txt
FirstName_LastName_SON_MovieLens.Big.case1-35000.txt
FirstName_LastName_SON_MovieLens.Big.case2-2800.txt
FirstName_LastName_SON_MovieLens.Big.case2-3000.txt

You don't need to specify the path of the output file in the commandline, you only need to save the file with the format FirstName_LastName_SON_DATASET.case1or2-SUPPORT.txt in the same path your program run.
You can change the ratings.csv file of the Big dataset to MovieLens.Big.csv and use the filename as the DATASET in the output filename

The above output files should be placed inside the OutputFiles directory of your submission.

4.1.2 Description

Inside the Firstname_LastName_Description.pdf document of your submission please include a table that is exactly the same with the one provided on the top

of this section. You must use the same support threshold values as the table above and include the execution times of your implementation for each case. Don't make up the numbers because we will rerun your program.

Grade breakdown

a. Four correct output files (2.5pts each)
b. Your execution time needs to be smaller than the ones in the table (2.5pts each)

General Instructions

1. Make sure your code compiles before submitting
2. Make sure to follow the output format and the naming format.

Grading Criteria

1. If your programs cannot be run with the commands you provide, your submission will be graded based on the result files you submit and 80% penalty for it.
2. If the files generated by your programs are not sorted based on the specifications, there will be 20% penalty.
3. If your program generates more than one file, there will be 20% penalty.
4. If you don't provide the source code and just the .jar file in case of a Java/Scala application there will be 100% penalty.
5. If your submission does not state inside the Description pdf file how to run your code, which Spark version you used and which approach you followed to implement your algorithm there will be a penalty of 30%.
6. There will be 20% penalty for late submission within a week and 0 grade after a week.
7. You can use your free 5-day extension.
8. There will be 10% bonus if you use both Scala and python for the entire assignment.
9. There will 0 grade if you use Dataframe or Dataset.

APPENDIX A

Pay great attention on the thresholds number for each case. The lower the threshold the more the computation. Do not try arbitrary threshold values. Try testing values within the given ranges.

- You need to take into account the Monotonicity of the Itemsets

• You need to leverage Spark capabilities of processing partitions/chunks of data and analyze the data within each partition.

• You need to reduce the support threshold according to the size of your partitions.

• You should emit appropriate (key, value) pairs in order to speed up the computation time.

• Try to avoid data shuffling during your execution.

APPENDIX B

Please include the following information inside your description document.

- Succinctly describe your approach to implement the algorithm.
- Command line command to execute your program
- Problem 2 execution table
- Problem 3 execution table