INF553 Foundations and Applications of Data Mining

Spring 2019

Assignment 4

Deadline: Apr. 1st 11:59 PM PST

1. Overview of the Assignment

In this assignment, you will explore the spark GraphFrames library as well as implement your own Girvan-Newman algorithm using the Spark Framework to detect communities in graphs. You will use the ub_sample_data.csv dataset to find users who have a similar business taste. The goal of this assignment is to help you understand how to use the Girvan-Newman algorithm to detect communities in an efficient way within a distributed environment.

2. Requirements

2.1 Programming Requirements

a. You must use Python to implement all tasks. There will be **10% bonus** for each task if you also submit a Scala implementation and both your Python and Scala implementations are correct.

b. You can use the Spark DataFrame and GraphFrames library for task1, but for task2 you can ONLY use Spark RDD and standard Python or Scala libraries. (ps. For Scala, you can try GraphX, but for the assignment, you need to use GraphFrames.)

2.2 Programming Environment

Python 3.6, Scala 2.11 and Spark 2.3.2

We will use these library versions to compile and test your code. There will be a 20% penalty if we cannot run your code due to the library version inconsistency.

2.3 Write your own code

Do not share code with other students!!

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

TAs will combine all the code we can find from the web (e.g., Github) as well as other students' code from this and other (previous) sections for plagiarism detection. We will report all detected plagiarism.

2.4 What you need to turn in

Your submission must be a zip file with the name convention: **firstname_lastname_hw4.zip** (all lowercase, e.g., xin_yu_hw4.zip). You should pack the following required (and optional) files in the zip file (see Figure 1):

a. [REQUIRED] two Python scripts, named: (all lowercase)

firstname_lastname_task1.py

firstname_lastname_task2.py

b1. [OPTIONAL] two Scala scripts, named: (all lowercase)

firstname_lastname_task1.scala

firstname_lastname_task2.scala

b2. [OPTIONAL] one jar package, named: (all lowercase)

firstname_lastname_hw4.jar

c. [OPTIONAL] You can include other scripts to support your programs and also name it with the prefix: "**firstname_lastname_"** (e.g. xin_yu_Graph.py)

d. You don't need to include your results. We will grade on your code with our testing data (data will be in the same format).

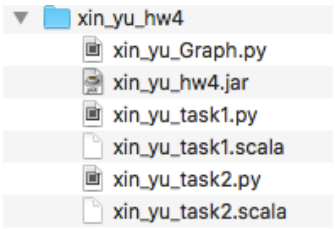

Figure 1: Submission Structure

3. Datasets

You will continue to use Yelp dataset. We have generated a sub-dataset, ub_sample_data.csv, from the Yelp review dataset containing user_id and business_id. You can download it from Blackboard.

4. Tasks

4.1 Graph Construction

To construct the social network graph, each node represents a user and there will be an edge between two nodes if the number of times that two users review the same business is **greater than or equivalent to** the filter threshold. For example, suppose user1 reviewed [business1, business2, business3] and user2 reviewed [business2, business3, business4, business5]. If the threshold is 2, there will be an edge between user1 and user2.

If the user node has no edge, we will not include that node in the graph.

NOTICE: In this assignment, the filter threshold is 7.

4.2 Task1: Community Detection Based on GraphFrames (2 pts)

In task1, you will explore the Spark GraphFrames library to detect communities in the network graph you constructed in 4.1. In the library, it provides the implementation of the Label Propagation Algorithm (LPA) which was proposed by Raghavan, Albert, and Kumara in 2007. It is an iterative community detection solution whereby information "flows" through the graph based on underlying edge structure. For the details of the algorithm, you can refer to the paper posted on the Blackboard. In this task, you do not need to implement the algorithm from scratch, you can call the method provided by the library. The following websites may help you get started with the Spark GraphFrames:

https://docs.databricks.com/spark/latest/graph-analysis/graphframes/user-guide-python.html

https://docs.databricks.com/spark/latest/graph-analysis/graphframes/user-guide-scala.html

4.2.1 Execution Detail

The version of the GraphFrames should be **0.6.0**.

For Python:

- In PyCharm, you need to add the sentence below into your code

  pip install graphframes

  os.environ["PYSPARK_SUBMIT_ARGS"] = (

  "--packages graphframes:graphframes:0.6.0-spark2.3-s_2.11")

- In the terminal, you need to assign the parameter "packages" of the spark-submit:

  --packages graphframes:graphframes:0.6.0-spark2.3-s_2.11

For Scala:

- In Intellij IDEA, you need to add library dependencies to your project

  "graphframes" % "graphframes" % "0.6.0-spark2.3-s_2.11"

  "org.apache.spark" %% "spark-graphx" % sparkVersion

- In the terminal, you need to assign the parameter "packages" of the spark-submit:

  --packages graphframes:graphframes:0.6.0-spark2.3-s_2.11

For the parameter "maxIter" of LPA method, you should set it to 5.

4.2.2 Output Result

In this task, you need to save your result of communities in a **txt** file. Each line represents one community and the format is:

'user_id1', 'user_id2', 'user_id3', 'user_id4', …

Your result should be firstly sorted by the size of communities in the ascending order and then the first user_id in the community in **lexicographical** order (the user_id is type of string). The user_ids in each community should also be in the **lexicographical** order.

If there is only one node in the community, we still regard it as a valid community.

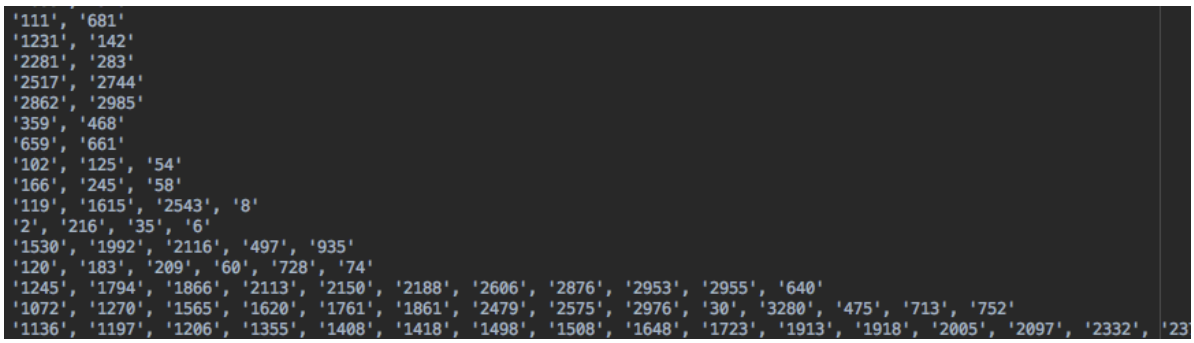

Figure 2: community output file format

4.3 Task2: Community Detection Based on Girvan-Newman algorithm (5 pts)

In task2, you will implement your own Girvan-Newman algorithm to detect the communities in the network graph. Because you task1 and task2 code will be executed separately, you need to construct the graph again in this task following the rules in section 4.1. You can refer to the Chapter 10 from the Mining of Massive Datasets book for the algorithm details.

For task2, you can ONLY use Spark RDD and standard Python or Scala libraries.

4.3.1 Betweenness Calculation (2 pts)

In this part, you will calculate the betweenness of each edge in the **original graph** you constructed in 4.1. Then you need to save your result in a **txt** file. The format of each line is

('user_id1', 'user_id2'), betweenness value

Your result should be firstly sorted by the betweenness values in the descending order and then the first user_id in the tuple in **lexicographical** order (the user_id is type of string). The two user_ids in each tuple should also in **lexicographical** order. You do not need to round your result.

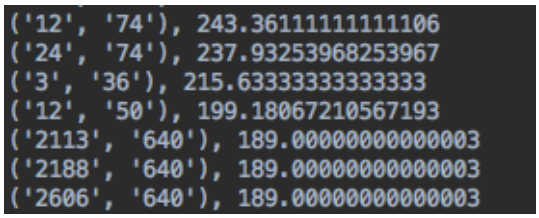

Figure 3: betweenness output file format

4.3.2 Community Detection (3 pts)

You are required to divide the graph into suitable communities, which reaches the global highest modularity. The formula of modularity is shown below:

Modularity of partitioning S of graph G:

- $Q = \sum_{s \in S}$ [(# edges within group s) – (expected # edges within group s)]
- $Q(G,S) = \frac{1}{2m} \sum_{s \in S} \sum_{i \in s} \sum_{j \in s} \left(A_{ij} - \frac{k_i k_j}{2m} \right)$

Normalizing cost.: -1<Q<1

A_{ij} = 1 if i connects j, 0 else

According to the Girvan-Newman algorithm, after removing one edge, you should re-compute the betweenness. The "m" in the formula represents the edge number of the **original graph**. The "A" in the formula is the adjacent matrix of the **original graph**. (Hint: In each remove step, "m" and "A" should not be changed).

It is positive if the number of edges within groups exceeds the expected number,
0.3 - 0.7 < Q means significant community structure

If the community only has one user node, we still regard it as a valid community.

You need to save your result in a **txt** file. The format is the same with the output file from task1.

4.4 Execution Format

Execution example:

Python:

spark-submit --packages graphframes:graphframes:0.6.0-spark2.3-s_2.11 firstname_lastname_task1.py <filter threshold> <input_file_path> <community_output_file_path>

spark-submit firstname_lastname_task2.py <filter threshold> <input_file_path> <betweenness_output_file_path> <community_output_file_path>

Scala:

spark-submit --packages graphframes:graphframes:0.6.0-spark2.3-s_2.11 –class firstname_lastname_task1 firstname_lastname_hw4.jar <filter threshold> <input_file_path> <community_output_file_path>

spark-submit –class firstname_lastname_task2 firstname_lastname_hw4.jar <filter threshold> <input_file_path> <betweenness_output_file_path> <community_output_file_path>

Input parameters:

1. <filter threshold>: the filter threshold to generate edges between user nodes.

2. <input file path>: the path to the input file including path, file name and extension.

3. <betweenness output file path>: the path to the betweenness output file including path, file name and extension.

4. <community output file path>: the path to the community output file including path, file name and extension.

Execution time:

The suggested overall runtime of your task1 (from reading the input file to finishing writing the community output file) is **200** seconds.

The overall runtime of your task2 (from reading the input file to finishing writing the community output file) should be less than **200** seconds.

If your runtime is between 200 seconds and 300 seconds, there will be 50% penalty.

If your runtime exceeds 300 seconds, there will be no point for this task.

5. Grading Criteria

(% penalty = % penalty of possible points you get)

1. You can use your free 5-day extension separately or together.
2. There will be 10% bonus if you use both Scala and Python.
3. If you do not apply the Girvan-Newman algorithm in task2, there will be no point for this task.
4. If we cannot run your programs with the command we specified, there will be 80% penalty.
5. If your program cannot run with the required Scala/Python/Spark versions, there will be 20% penalty.
6. If the outputs of your program are unsorted or partially sorted, there will be 50% penalty.
7. The total runtime of this assignment should not exceed 15 minutes or there will be no point for this assignment.
8. We can regrade on your assignments within seven days once the scores are released. No argue after one week. There will be 20% penalty if our grading is correct.
9. There will be 20% penalty for late submission within a week and no point after a week.
10. Only when your results from Python are correct, the bonus of using Scala will be calculated. There is no partially point for Scala. See the example below:

Example situations

Task	Score for Python	Score for Scala (10% of previous column if correct)	Total
Task1	Correct: 3 points	Correct: 3 * 10%	3.3
Task1	Wrong: 0 point	Correct: 0 * 10%	0.0
Task1	Partially correct: 1.5 points	Correct: 1.5 * 10%	1.65
Task1	Partially correct: 1.5 points	Wrong: 0	1.5