Lab 02 Specification – Implementing data collection modules using the Cloud.
Due Monday, 12th Feb 2018 1PM (after 1PM submission will not be accepted)
50 points

Lab Goals

• Implementing Java application using Jsoup Library for data collection modules.
• Implementing Java application using Amazon AWS SDK and Jsch Library for cloud interaction modules.

Assignment Details

This lab exercise is going to be a team based lab. So you are allowed to collaborate with your team members and do the lab work together as a team. You are expected to submit your lab work on or before the due date by sending an email to the Instructor. See more details about submission in the last section. There is also an option for teams to collaborate through discussions and do individual implementations.

Preliminary Requirement:

1. You should implement the task using Java programming.
2. You should create a folder called cmpsc250-lab02-YOURFirstInitialLastName. So for example cmpsc441-lab02-amohan
3. You should create four sub folders namely: src, jar, dataset, and classes. All your source code ”.java” needs to be inside the src folder. All your classes files ”.class” needs to be in your classes folder. All your jar files ”.jar” needs to be in your jar folder. All your data files such as ”.txt” needs to be in your dataset folder.
   A simple technique to do this organization is by using the following command during compilation:
   javac -d ../classes -cp ./*.
4. We adhere to the group assignment that was done in our earlier class and the group sheet is provided in the course webpage. This lab work will be done within your group and all the team members are going to contribute to the lab in a fair manner.

If you fail to adhere to the requirements, your lab work will not be evaluated and thereby you and your team will not receive any points for the lab.

The Distributed Web Scraper

Problem Definition: Data collection is a tedious and time consuming process. How to get large scale data from web sources using the least amount of time? More specifically, how do we collect the information about different dog breeds, their characteristics and the rating associated with it.

Approach: We discussed during the lecture session, how to use Jsoup Library for collecting web data and how to use Amazon AWS SDK and Jsch Library to automate the Cloud interaction. Can we use both to solve this problem?

Requirement:
1. You should create a class called DistributedWebScraper that provides the implementation of your solution to the problem above.

2. You are required to implement the solution to collect the data from the dog breed profile site given below: http://dogtime.com/dog-breeds/profiles

3. You are required to implement a method called ShowAllDogBreeds, which should get the list of dog breeds and classify them alphabetically. For example: A-F (Class 01), G-L (Class 02), M-R (Class 03), S-Z (Class 04). You can work on top of the code provided to you by the Instructor.

4. You are required to implement a method called GenerateBCR, which should generate the Breed Characteristics Rating report for all the breeds in a file called DOGNAME-bcr.txt inside the dataset folder. For example: in class 01 AmericanBulldog.txt should look like:


5. You are required to implement a method called AWSPro, that will use the Amazon AWS SDK Library provision 5 number of machines in the EC2 Cloud with Java installation done in those instances.

6. You are required to implement a method called AWSShell, that performs three major duties. Firstly, the function moves the entire folder with the compiled classes from the source machine to all the Cloud instances. Compile your code first without AWSPro.java and AWSShell.java and make it ready in all the Cloud instances. Secondly, the function logs in to the shell of all the Cloud instances and execute the GenerateBCR class with a command line argument class01 (for instance 01), class02 (for instance 02), class03 (for instance 03), class04 (for instance 04), class05 (for instance 05). Thirdly, the functions moves all the data residing in the Dataset folder from all those Cloud instances back to the Dataset folder in your code executed on the client side.

7. Each machine is supposed to work on a particular dog breed data and at the end all the machines will consist of all the dog breed datasets that are generated in a distributed manner.

**Submission Details**

1. You are required to submit this lab by sending an email with a zipped version of your cmpsc441-lab02-YOURFirstInitialLastName folder.

2. Subject of your email should say "CMPSC441: Team X Lab 02 Submission". Here X needs to be replaced with your group number. Refer the group sheet in the course webpage.

3. One email should be sent for your teams lab submission.

4. Send the email to amohan@allegheny.edu // CC the email to all your team members.

5. You should add the following statement in the body of your email
   By doing this submission, I understand that I and my team members are subject to the Honor Code policy. Lab submitted by: X1, X2, and X3 (Here X1, X2, and X3 are the name of your team members)

6. Shifting team members is not ideal and not allowed, unless there is an extreme situation which is discussed with the instructor prior to your submission.