K-means for Geo-location Clustering in Spark

Option: Project3

Course: CSE427S Cloud Computing for Big Data Application

Instructor: Marion Neumann

Student:

Start: 11/20/2018

End: 12/14/2018

**Milstone1:**

|  |  |  |  |
| --- | --- | --- | --- |
| Responsibility | Name | Student ID | Email |
| project manager | James Chen | 467360 | james.chen@wustl.edu |
| developer local |  |  |  |
| developer cloud |  |  |  |
| key user | Jason Zhu | 467138 | jingxuan@wustl.edu |

**Milstone2: Midterm Discussion and Report formation**

* project management (*every* team member is required to *participate* in the discussion)
  + conceptual understanding (5%)

project objects clarification(Goal):

data preprocessing:

What kind of data are we preparing for k-mean(step1-3)?

What is output result from preprocessing program?

What is input for k-mean?

k-mean:

What k-mean actually does?

How do we start with initialization(k-mean data visualization)?

What tool will we use for data visualization?

What is dimensionality of k-mean?

k-mean limitation(pseudo-cluster):

Dimensionality of data

Size of data

* + understanding of roles and individual responsibilities (5%)

project manager:

developer local:

developer cloud:

key user:

* technical (*check project instructions for more details*)
  + data preprocessing and other steps completed (10%)

**Milestone3: Final Report**

1. motivation (5%)

2. documentation of approach (5%)

3. small data/pseudo cluster

1) [code] implementation (20%)

2) [write-up] results/discussion (10%)

4. Big data application/dataset

1) which dataset/application – creativity (5%)

2) [write-up] description (5%)

3) [code] implementation/execution (10%)

4) [write-up] results/discussion (10%)

5. [write-up] final conclusion/lessons learned/future work (5%)

**Team Member Roles and Responsibilities**

* **project manager**
  + communication/coordinator among **all team members**
  + manages submission to SVN repository
  + set up and maintain team repo (optional)
  + [report/write-up] template
  + [report/write-up] motivation and introduction – coordinate with **key user**
  + [report/write-up] documentation of approach – coordinate with **developers**
  + [report/write-up] conclusion – coordinate with **key user**
* **developer local**
  + implementation
  + [report/write-up] documentation of implementation
  + testing locally
  + testing pseudo-cluster together with **developer cloud**
* **developer cloud**
  + assists **developer local**
  + testing pseudo-cluster
  + cloud execution
  + [report/write-up] documentation of cloud execution
  + assists **key user** with documentation of results
* **key user**
  + data preprocessing
  + find and preprocess Big data application/real-world dataset
  + [report/write-up] documentation of real world data
  + assists **developer cloud** in executing implementation on real-world data
  + [report/write-up] documentation of results