Assignment 1

Ayush Tiwari (17CS10056)
ayushtiwari@icloud.com

October 2, 2020

Note: All my programs are self contained and can be built using **sbt**. I have provided the `build.sbt` file.

Question 1

(a)

- The `GitLog.scala` file contains the case class for storing logs.
- `processLine` contains the valid regex for parsing input.
- A flatmap has been used to map each line to the corresponding **GitLog** object.
- `rdd.count` gives the size of data.

(b)

- Filter by `debugLevel=="WARN"`

(c)

- Filter with `retrievalStage=="api_client"`.
- Map to the repository name and count distinct.

(d)

- Filter with key `"api_client"`
- Map and Reduce therafter
- For failed accesses, Same as above, just check for `"Failed"` in `rest`.

(e)

- For most active time use key as hours and then map and reduce.
- For most active repository use the repository rdd extracted in (c) and then map and reduce.

(f)

- Filter by finding `"Failed" and "Access"` as substring in `rest`.
- Map to the Key and the reduce.

Question 2

(a)

- Use `rdd.productElement(column_no)` to create a column iterator.
- We can then traverse the column downloadId with this iterator.

(b)

- Use the same logic as in Question 1 c), just filter by downloadId beforehand.

(c)

- Get the iterator and traverse it, the print the count of unique entries.

Question 3

(a)

- Process in a similar fashion as in 1 a).
- Then print the count.

(b)

- key infoRdd and logsRdd by repo name
- For logsRdd first extract the repo name (similar to 1 c)) Use join function

(c)

- Check for `"Failed"` string in URL and then Map and Reduce.