DATASET ANALYSIS

App identification based on data volume

Huber Flores

huber.flores@helsinki.fi

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Helsinki, Finland.

Guidelines

- Answer each question
  - The goal of each question is to give you insights about how to solve the problem
- Justify each question
  - Ultimately, each question (if applicable) should turn into an assumption that is part of the pipeline
- Add your questions
  - Let your mind discover new questions and add them for later discussion

Problem?

- Create a program that given as input the data volume transmitted of a particular app, and it returns back the name of the app that was used (id).

Solution?

- Create a main pipeline file (pipeline.R)
- Split your code into steps
  - Aka pipeline components
  - Each component may be associated with one question (justify)

```
1
2   #
3   # Step 1: load data
4   #
5   mydata <- read.csv("dataset.csv", sep=",")
6
7
8   #
9   # Step n: remove apps that do not have enough samples
10  #
11  .
12  .
13  .
14
15
```

Questions

- How many different apps contain the dataset?
- What is the max and min data volume interval of the dataset?
  - E.g., appX [500,505], appY[100, 10000]
- How many apps share similar data volume interval?
- If two apps have similar data volume interval, how can you make them different during the analysis [Tip: discretization]

Questions

- What is the max and min amount of samples for a particular app in the dataset?
- What is the min amount of samples needed for an app, such that it can be considered in the analysis?
- How many apps fulfil the previous requirement?
- Is there any relation between app usage and location (base station)?
  - If so, it could be possible to determine which app is used by combining location and volume

Questions

- How many different users has the dataset?
- Is there any relation between app usage, location (base station) and a particular?
  - users using a particular app in the same location would give some insights, e.g., pokemon go
- Is there any relation between app usage, location and time?
  - Time field format: 21143849 means: 2017 August 21, 14h 38m 49s