Preprocessing the smartphone behavior data

——320180942181 zhangxin

Homework description:

There are three files in my homework10 directory. They are ‘data_preprocess.ipynb’, ‘crawling.py’, and ‘documentation.pdf’.

‘data_preprocess.ipynb’: The main function implementation code file, run the code in it will get the preprocess result.

‘crawling.py’: A module will be called in ‘data_preprocess.ipynb’, its function is to crawl data we need from web page and save them in different directories on disk.

‘documentation.pdf’: Describe the homework.

Prepocess procedure:

1. Use custom function crawl required data from web page.
2. Get path of files from disk for pandas to read.
3. Read json files from disk, show information about every file. Finding out whether there are empty files. Finding out whether there are missing values in the data of every file.

4. Delete files that have unreasonable time. Draw graphs to get the time every data cost in the test and delete files that have unreasonable time.

5. Delete files that have unreasonable data.

(1) Draw the standard deviation change chart of each type of data. If standard deviation is too small, means that data change is too small. The data is unreasonable data.

(2) After last step, for more intuitive judgment, draw four types of graphs about all data to determine whether the data in each file is reasonable or dirty due to equipment or operation reasons: Whether the range of data changes reasonable. Then delete files that have unreasonable data.

6. Check how many files have been changed after process.

In the end:

During doing this homework, for so much data, I don’t know how to process them better. I hope that teacher could give us a demonstration. Because I want to learn from the correct way to deal with the data. Hope the teacher can criticize and correct!