

Assignment 2

CSE 406 – Software Engineering

Due: Nov. 1. *Submit in Class*

Programming Assignment (Individual assignment, version 1)

You will get a chance to use a few open source projects called:

- Apache Spark (<http://spark.apache.org/>)
- Pandas (<http://pandas.pydata.org/>)

Download `assign2.tar.gz` to start this assignment.

***Task 1.**

You will use Spark and Pandas to read the json data file (`data-300k.json`) and create separate data files grouped by distinct `'uuid'`. Each file should be **placed under the data directory** (`./data/`), **exported to file type 'pickle'**, and named as `'uuid.pickle'`. Write you python program, `'task1.py'`

`data-300k.json` schema (you can simply ignore what they mean):

```
-- RTT: double
-- SSID: string
-- Strength: long
-- WiFiStatus: string
-- _id: struct
    |   |-- $oid: string
-- latitude: double
-- longitude: double
-- timestamp: long
-- type: string, 'WiFi' or 'Mobile'
-- uuid: string
```

Example:

```
>> spark-submit ./task1.py
```

***Task 2. a)**

You will use the given `'d30c14b3-4039-3ad8-9cc3-025485863b7c-61939.pickle'` file to complete Task 2. Read this sample pickle file and count how many times `'WiFi'` and `'Mobile'` appear under `'type'` schema.

***Bonus.** Task 2. b) Compute the longest consecutive ‘WiFi’ appearance in this pickle file, likewise for ‘Mobile’

Example:

```
>> spark-submit ./task2.py
```

WiFi: 500

Mobile: 1000

Longest WiFi: 35

Longest Mobile: 20

SUBMISSION:

1. Print out your source code – “task1.py” and “task2.py”
2. Write one paragraph explaining your program and any difficulties you had. (No less than 250 words)

NOTE:

1. Python is needed.
2. Command Line base is recommended.
3. Any platform of your choice is fine, but I’d like to see many Linux as possible (e.g., Ubuntu).

IMPORTANT:

Do your best. Even if you can’t do the whole assignment, submit as much as you can (with explanation why you can’t do this). And to make sure that you do your own assignment, I will randomly select a few students in class and ask them to explain their code.

TRIVIA:

Google is your friend and teacher. Search!

Discuss with your classmates! (DO NOT send me an email first)

I will go over the assignment in class, so don’t worry too much.