CSCB07 – Software Design

Lab 3

Topics Covered

- Input/Output
- Regular Expressions

Logistics

- This lab will be supervised by your TA during the tutorial session of Week 5 (June 6 - 10, 2022). If you encounter any problem while doing the steps listed in the following sections, ask the TA for help.
- The lab should be done individually.

Instructions

The "Tweets2020" folder provided with this lab includes tweets/retweets/replies posted by 10 influential people in 2020. Using regular expressions, you are required to develop Java code that extracts the hashtags used by each person and stores them in separate files (one file for each person). The following should be taken into consideration:

1. You need to unzip *Tweets2020.zip*.

2. The paths of the folder containing the tweets and the one that would be used to store the hashtags should be provided by the user at runtime using the keyboard.

3. Storing duplicate hashtags is allowed.

4. You should not make any assumptions about the number of people or their names. That is, your code should work **without modification** for other folders having a different number of people and/or different names.

5. You can make the following assumptions about hashtags:

   a. Apart from the first character which should be a '#', only letters, digits, and the underscore character are allowed within a hashtag.

   b. A hashtag is preceded by a whitespace character if it doesn't occur at the beginning of the file, and it is followed by a whitespace character if it does not occur at the end of the file.

Submission

Upload your code to “Lab 3” on Quercus by June 12th at the latest.