Twitter Data Fetching and Profiling Package

This is a simple package for getting tweets using Twitter Stream API.

Licence: GPLv3

How to compile

Prerequisites
Maven (https://maven.apache.org/)
Java Development Kit (http://www.oracle.com/technetwork/java/javase/downloads/index.html)

Download the package and compile using the following command:

```
mvn clean install
```

The package dependency should be resolved automatically. You will see no error if the package is installed successfully. A jar file is generated under a newly generated directory called “target”, and the name of the jar file looks like “BigDataProject-XXX.jar” where XXX is the version and build information.

How to use

First, create a file called twitter4j.properties in the root directory of the package with the following content:

```
debug=true
oauth.consumerKey=******
oauth.consumerSecret=******
oauth.accessToken=******
oauth.accessTokenSecret=******
```

Replace the asterisks with your corresponding twitter consumer key and application token. Be sure to grant “Read” access to your application.

To call the program, use the following command:

Profiling
The profiling code tests the field length, providing basic characteristics of the data.

```
hadoop jar target/BigDataProject-XXX.jar profiling.Profiling [path to the input file] [path to the output directory]
```

Fetching Tweets based on keywords (only lang=en)

```
nohup java -cp target/classes:target/lib/* twitter.TweetsFetcher [keywords] >>[output file] 2>[log file] &
```

Fetching Tweets based on locations (only lang=en)

```
nohup java -cp target/classes:target/lib/* twitter.TweetsFetcher [latitude, longitude] >>[output file] 2>[log file] &
```

Clean Fetched Tweets (only keep whitespace and Latin-Basic characters)

```
cat [tweets file] | java -cp target/classes/:target/lib/* cleaning.Cleaning > [output file]
```