

Getting Start with Weibo Data Collection

This tutorial is to help you setup the environment to collect data from Weibo using a Java program. The program follows OAuth2.0 authorization mechanism to get access to the Weibo data. It fetches data through Weibo API and stores the data to the MySQL database. A simple analysis tool for data processing is also developed in the program.

What to you need to get started?

1. A Weibo developer account.

Please go to <http://open.weibo.com/> to create (update) your developer account and create an application (站内应用). After creating your application, you need to mark down your 'appkey' and 'app_secret'.

2. A Java IDE (e.g., Eclipse. In the following, I will use Eclipse IDE as an example)

3. MySQL installed.

How to setup?

1. Import the project "WeiboCrawler" in the attached package to your Java IDE (Eclipse, here).

2. Connect Java with MySQL

- a) Using MySQL commands create a database named 'sinaweibo' from the terminal.

- b) Open Eclipse. Go to the project property page. Under 'Java Build Path' -> 'Libraries' tap, click 'Add External JARs'. Add the jar 'mysql-connector-java-5.1.18-bin.jar' (available in the attached package as well).

- c) From the 'Package Explore' in the Eclipse, locate the file 'examples/(default package)/WeiboAnalysis.java'. Find the function 'InitialMysql()'. Set up the proper configuration values for 'database_name', 'url', 'user' and 'password'.

- d) Test the connection to MySQL. Run the function InitialMysql() only [e.g., you disable all the other functions in main()]. If success, you will see the message "**Succeeded connection to the Database.**"

Ref: Please find in <http://zetcode.com/databases/mysqljavatutorial/> for a simple tutorial of connection MySQL from Java.

3. Get the Weibo access token

We follow OAuth2.0 policy provided by Sina Weibo (<http://code.google.com/p/weibo4j/>). The simple tutorial will guide you to get your access token.

Note that, you do not need to download the Java SDK as mentioned in the tutorial, since our program was built from that SDK and we kept all the necessary files. The two files mentioned in the tutorial are located under 'src/config.properties' and 'examples/weibo4j.examples/OAuth4Code.java'.

4. Start the data collection process

a) Locate the file 'examples/(default package)/WeiboAnalysis.java'.

b) Configure the main() function

Specify the value 'access_token' and 'Entrance_user_id' in the main() function. The 'Entrance_user_id' is the UID, a unique identifier for each Weibo user. You can randomly specify one. Please read the Weibo API documents for more details. (<http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5>)

** A suggestion is that to choose the user with a small number of followings and followers (around hundreds) to start with.

** Note that the data collection process uses a recursive manner. It starts from the root user, and add the root user to the table (tblIndex = 1). It then adds all the followers and followings (** the followings may not be included due to performance issue in this program. Please also have a check) at the end of the table with increasing tblIndex's. The process keeps going by exploring users (users' profiles and their relationships) from tblIndex = 2 user and adds new users to the end of the table.

You should now see the data collection process is running. Good luck.