Documentation for Twitter Crawler

Table of contents

[Documentation for Twitter Crawler 1](#_Toc338220168)

[1 Installation 2](#_Toc338220169)

[1.1Preparation 2](#_Toc338220170)

[1.2Getting the crawler procedure 2](#_Toc338220171)

[2 Configuration 3](#_Toc338220172)

[2.1Two configuration files 3](#_Toc338220173)

[2.1.1Config\_mongo.properties 3](#_Toc338220174)

[3 Run the crawler 4](#_Toc338220175)

[3.1Check whether the crawler is run successfully 4](#_Toc338220176)

[4 Crawler design 5](#_Toc338220177)

[5 issues and challeges in this crawler 6](#_Toc338220178)

# 1 Installation

## 1.1 Preparation

#### JAVA 1.6 or later (since the crawler is written in JAVA)

#### Twitter4j (just download it from <http://twitter4j.com>, it is an optional step since the twitter4j library is included in the crawler procedure)

#### Apply twitter account as well as twitter developer account from the link: <http://dev.twitter.com>, register an application (just fill in a form in the site above, application name and other description can be anything), and finally get four string segments (consumer key, consumer secret, access key, access password) which will be put into the configuration file *twitterToken.txt,* further configuration details will be put in configuration section.

## 1.2 Getting the crawler procedure

Download it from SVN <https://svn.cigi.uiuc.edu/socialmedia/TwitterCrawler> , and unzip it if compressed in a proper directory in your own system.

Let’s assume the directory of crawler procedure is $TWITTER\_CRAWLER, we will take the assumption in the whole documentation.

# 2 Configuration

## 2.1 Two configuration files

### 2.1.1 Config\_mongo.properties

Find the configuration file at $TWITTER\_CRAWLER/config/config\_mongo.properties

**DBUser** -- user name of the mongoDB

**DBPassword** -- password of the user

**DBName** – database name

**DBUrl** –- mongoDB server IP/domain name

**DBPort** – port for the mongoDB service, default value is 27017

**DataDirectory** – directory of storing data file

BufferSize – batch size for the single update in database.

**Bbox** – the region which you are concerning, with the format *(sw\_lat,sw\_lon),(ne\_lat,ne\_lon)*

#### 2.1.2 twitterToken.txt

This file is created to store the authentication information for twitter.

Find the configuration file at $TWITTER\_CRAWLER/config/twitterToken.txt

The file format can be seen in Fig 1

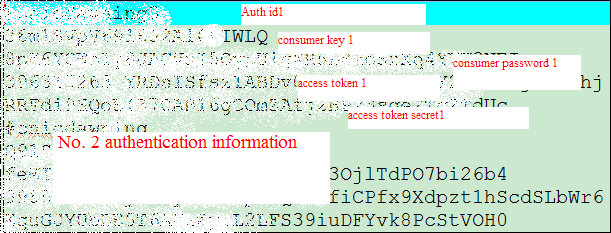


Fig 1. twitterToken.txt sample

# 3 Run the crawler

Go to the directory $TWITTER\_CRAWLER, just use *./crawler.sh command* to start, for background run mode, use *nohup ./crawler.sh &* instead.

**Hint**: change the file access level first to make sure you can execute the file.

## 3.1 Check whether the crawler is run successfully

* Find the log file from the $TWITTER\_CRAWLER/log, and see if there are any exceptions
* Go to MongoDB database, to check whether data is stored in the proper collection and database. For MongoDB usage documentation, please refer to “**Documentation for current**

**MongoDB instance**“

# 4 Crawler design

This twitter crawler is using twitter steaming API to get tweets posted by twitter users. With passing geo boundary as the parameter to streaming API, we can get geo-related tweets; with passing flu related key words to the API, we can get tweets talking about flu.

The twitter status is monitored by a StatusListener which is implemented by us, once there is new update in the concerned region, the listener can get to know immediately and return the update(tweet) object. We cache the data into data file first with a batch, and finally put data into MongoDB with the attributed described in the other documentation.

Workflow of the twitter crawler is like Fig. 2

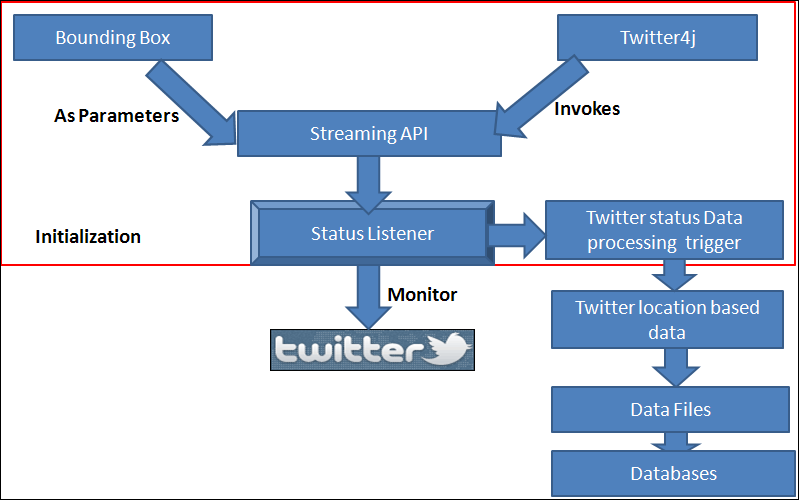


Fig.2 workflow of the twitter crawler

This crawler can also support multi thread crawling from the twitter platform, with one single thread taking charge of one separate region. The framework is as Fig. 3

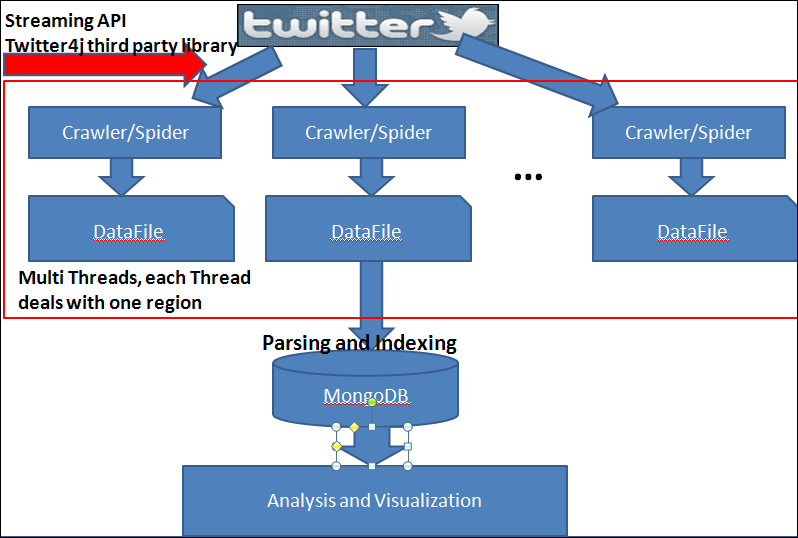


Fig.3 multi-thread twitter crawler framework

# 5 Issues and challenges in this crawler

Owing to Twitter’s policy, we can only get a sample of the whole dataset at our current account level.