INF344, Télécom ParisTech

Twitter and Web Scraping

Pierre Senellart (pierre.senellart@telecom-paristech.fr)

23 June 2016

The purpose of this lab session is to combine the use of the Twitter API with Web scraping to build a Web application that combines data from Twitter with that of a regular Web site, Ethnologue. Ethnologue http://www.ethnologue.com/ is a Web site containing reference information on the languages of the world. The goal is to find references to the Ethnologue Web site on Twitter that relate to a specific language described on Ethnologue, and to associate the corresponding tweets with information about the language referred to on the Ethnologue Web site, using a Web scraping system.

1 Twitter API

You will need a Twitter account for this lab session; create one if you do not already have one. In addition, you will need to create an "application" for the purpose of this lab session, on https://apps.twitter.com/, and request authentication tokens for your application, see https://dev.twitter.com/oauth/overview/application-owner-access-tokens. You will need to note the consumer key, consumer secret, authentication token, and authentication secret.

There are limits on the number of requests to the Twitter API you are allowed to perform (see https://dev.twitter.com/ for details). These are relatively high, but this means you should be careful in how you use the Twitter API and how you test your program.

2 Baseline

A skeleton Eclipse project is available for download from the course Web site. You are free to use another Java development environment; if you do not use Eclipse, you will need to make sure both the config/ directory and all JARs in the lib/ directory are in the class path.

The skeleton project contains the following JAVA source files:

Ethnologue.java This is the main part of the program to implement and the only file you should modify. When submitting your solution to the submission server, you only need to submit this file. *TODO* annotations indicate which parts of the file need to be filled in.

TwitterEthnologueReference.java A passive data structure (i.e., a struct-like Java class) for collecting information about Ethnologue references in Twitter. This class has the following public members:

languageCode the three-letter ISO 639-3 code of the language

tweetUserName the user name of the author of the tweet

tweetContent the textual content of the tweet

tweetId the Twitter id of the tweet

ethnologueURL the URL of the page on the Ethnologue Web site with information about the language; it should use the same protocol to access the Ethnologue Web site (HTTP or HTTPS) that is used in the tweet

ethnologueLanguageName main name of the language, as given by Ethnologue

ethnologuePopulation number of language speakers, as given by Ethnologue; as a simplification, this should be the first number appearing in the "population" field within the Ethnologue Web page

PublicTests.java A collection of public tests that will be run on your implementation.

3 Submission and Evaluation

Submit your solution by Wednesday, July 6th, 11:59pm on the submission server for full credit. Submissions received after this date and before Thursday, July 7th, 8:30am will incur a penalty of 4 points out of 20. Submissions received after this second deadline will not be considered. You only need to submit the `Ethnologue.java` file, not the whole Eclipse project.

Your submission will first be run against the public tests; as they are identical to the ones provided in the PublicTests class, there is no need to use the submission server unless your submission passes these tests locally. A score out of 20 points will be given as an indication (this score is not part of the grade). When confident about your submission, you can release it to test it against the secret (release) tests. The score obtained on the release tests will be used as a grade for this lab session. You are allowed at most 3 release submissions per hour. Evaluation of your submissions will typically take several minutes.

4 Structure of `Ethnologue.java` and What to Implement

Two libraries are included in the skeleton project: Twitter4J to access the Twitter API and TagSoup to parse HTML pages from the Web. The documentation for the former is at `http://twitter4j.org/javadoc/` ; the latter exposes the HTML document as a `Node` object, see `http://docs.oracle.com/javase/7/docs/api/org/w3c/dom/Node.html`.

twitter is a reference to a Twitter4J `Twitter` object used to access Twitter.

reader is a reference to a TagSoup `Parser` object used to parse HTML content from Web pages into a DOM representation.

xpath is a reference to an XPath processor that can be used to retrieve data from a DOM tree.

Ethnologue() is the constructor and initializes the `twitter` and `reader` objects. You should simply add your Twitter credentials in this function, no need to change anything else.

statusToTwitterEthnologueReference should take as input a Twitter4J `Status` object and return an instance of the `TwitterEthnologueReference` class, filled with information from Twitter and Ethnologue. This is the main method to implement.

urlToHTMLNode is an already written utility method that retrieves the content at a given URL and parses it into a DOM representation using TagSoup. You can use this method when you need to retrieve content from the Ethnologue Web site.

get should use Twitter4J to retrieve the tweet that has the corresponding id.

search should use Twitter4J to search for `c` tweets referencing a language on the Ethnologue Web site. You have to experiment with Twitter search function (you can try on the Twitter Web site itself, or via the API) to determine how best to express this search.