G52APR Coursework 2011 – Part I, Data Collection

Colin Higgins

Introduction
For the G52APR coursework you are to develop a system for downloading and querying association football (soccer) results and statistics from a variety of leagues and seasons. An introduction was provided in lectures. To aid with the design, implementation and testing of the system the problem is broken into three linked parts. The three sections are:

- Data Collection (DC)
- Data Storage (DS)
- Data Querying (DQ)

As with all large pieces of software, one of the most important tasks is to define the interfaces that the different pieces use to interact. These have been provided for you, but you should take some time to think about how these are going to work. Regarding the coding, you will be able to change earlier sections of your program as you develop the later sections, if you so desire. The interface between the test frame and the DC section is relatively simple. However, the DS and DQ interfaces are a little more complex.

These interfaces should allow you to develop the separate parts of the coursework in sequence and independently. Once you have understood what methods the interfaces use you might want to create minimal implementations for the first section then expand this into the full version. Repeat for each section in turn. In effect, we are using the Strategy pattern here. We could vary how your server stores its data just by creating a new implementation of the interface (and we actually do this for the DC section).

The rest of this document expands on what is expected for the first component of the coursework, including what that its functionality should be, its deliverables and the hand-in date for those deliverables.

Data Collection (40%)
Due: 21\textsuperscript{st} November, 16:00

This section involves requesting a web page that will contain the football results for a given league and given season. The sub-system, which may consist of one or more classes depending on how you design it, will make this request and then parse the HTML page to extract the actual data. It will use the DS interface to store this data. You will be given a simple implementation of the DS which simply
writes out a copy of the data you send it. The resources section at the end of this document indicates suitable web sites and gives some hints on how to approach the problem.

Key:

| given | To do | done |

Requirements

Design and build a Data Collection sub-system that will fetch from the web, extract and store football scores. You should save the league, season, match date, home team, away team, home score and away score (do not worry about half time scores). Make the system as robust as possible, particularly bearing in mind that the internet may not be a completely reliable transmission medium. For instance what happens if the network connection, or the web server is not working. That is your system should fail *gracefully*.

Design and write a test frame that will allow you to show that the DC sub-system works.

You may use any libraries you see fit, provided you reference and acknowledge them both in code comments and in the report.

Testing

You should write a class that contains a main method to exercise the DC sub-system. If you feel further testing is required then do so and explain in the deliverables document below. The more complete and thorough the testing, the higher the marks that will be obtained.

Deliverables

You should deliver (electronically via cw submit – details below) the following:
• Source code for your data collection sub-system and test frame.
• A short (approx 500 word) report explaining how your system works, what choices you have made and why. You may wish to do this in the form of a programmer's manual with javadoc comments. The report should be in plain text.

Bonus marks can be obtained if the second site can also be queried for later seasons results, but beware this is messy and difficult.

Please submit a zip file of your source code and your report to cw submit:

427) G51APR cw1: Data Collection CODE (cah)
428) G51APR cw1: REPORT (cah)

Feedback
You will be given written feedback on the code and report

Resources

Main Site:
http://www.footballresults.org/

Bonus site:
http://www.soccerstats.com/

Java hint:
System.setProperty("proxyHost", "wwwcache-20.cs.nott.ac.uk");
System.setProperty("proxyPort", "3128");

For parsing HTML see:
org.htmlparser.visitors.NodeVisitor;
Or
org.htmlparser.parserapplications.StringExtractor;

Or use your own library or code.