

HW4 - XGrep

This is a small homework to experiment with the Java's XML parsing features. HW4 is due midnight ending Thu June 7th. We will build an application that is somewhat like the unix tool "grep", but parses an XML file instead of a text file and takes advantage of XML's structure. In the simplest case, grep takes a "target" string and a file, and prints out the lines of the file that contain the target. XGrep takes a target string and an xml file, and prints out any nodes that contain that string..

- A node matches if the characters enclosed by the node include the target string, or if a value in an attribute of the node includes the string. Matching is not case sensitive and is just a plain substring search (not regular expressions).
- If a node matches as above, the names of the parent nodes should be printed on one line, each in brackets < ... >, from the root node down to the match node.
- The next lines should print the attribute=value bindings of the match node, if any, one per line.
- Finally, the characters enclosed by the node should be printed on their own lines followed by a blank line. We won't worry about putting in line breaks, so the lines may run long or short. Remove the whitespace from the beginning and end of the characters before printing.

So the output will look like...

```
<a><b><c>           // sequence of nodes root..matched
pi=3.14            // bindings from the matched node
name=Sally
This time for sure! // characters
```

So with the quotes.xml file...

```
<?xml version="1.0" encoding="UTF-8"?>
<quotes description="Our XML quotes">
  <movies>
    <quote title="Romy and Michele's High School Reunion">
      This dress exacerbates the genetic betrayal that is my legacy.
    </quote>
    <quote title="Aliens">
      In case you haven't been paying attention to current events,
      we just got our asses kicked, pal!
    </quote>
    <quote title="Aliens">
      We'd better get back, 'cause it'll be dark soon, and they mostly
      come at night... mostly.
    </quote>
  </movies>
  <novels>
    <quote title="The Shipping News" author="Annie Proulx">
      Used to say there were four women in every man's heart. The Maid
      in the Meadow, the Demon Lover, the Stouthearted Women, and the
      Tall and Quiet Woman. I don't know what it means. I don't know
      where he got it.
    </quote>
    <quote title="The Shipping News" author="Annie Proulx">
      You all know we are only passing by. We only walk over these stones
      a few times, our boats float a little while and then they have to sink.
      The water is a dark flower and a fisherman is a bee in the heart of her.
    </quote>
    <quote title="Tale of Two Cities" author="Charles Dickens">
      It was the best of times, it was the worst of times.
    </quote>
    <quote title="The Heart of the Matter" author="Graham Greene" year="1948">
      They had been corrupted by money, and he had been corrupted by sentiment.
      Sentiment was the more dangerous, because you couldn't name its price.
      A man open to bribes was to be relied upon below a certain figure, but
      the sentiment might uncoil in the heart at a name, a photograph, even
      a smell remembered.
    </quote>
  </novels>
  <plays>
    <quote title="Julius Ceasar" author="William Shakespeare">
      But those that understood him smiled at one another and shook their heads;
      but, for mine own part, it was Greek to me.
    </quote>
    <quote title="Hamlet" author="William Shakespeare" year="1601">
      'Tis now the very witching time of night,
      When churchyards yawn and hell itself breathes out
      Contagion to this world.
    </quote>
  </plays>
</quotes>
```

With the target "alien" the output is...

```
%> java XGrep alien quotes.xml
<quotes><movies><quote>
title=Aliens
In case you haven't been paying attention to current events,
we just got our asses kicked, pal!

<quotes><movies><quote>
title=Aliens
We'd better get back, 'cause it'll be dark soon, and they mostly
come at night... mostly.
```

Or with the target "night"...

```
%> java XGrep night quotes.xml
<quotes><movies><quote>
title=Aliens
We'd better get back, 'cause it'll be dark soon, and they mostly
come at night... mostly.

<quotes><plays><quote>
title=Hamlet
author=William Shakespeare
year=1601
'Tis now the very witching time of night,
When churchyards yawn and hell itself breathes out
Contagion to this world.
```

Logistics

We will use the Sun Jaxp library. Set your CLASSPATH to include /usr/class/cs108/jar/xml.jar. The API docs for Jaxp are at <http://java.sun.com/xml/jaxp-1.0.1/docs/api/>. See the HandlerBase class.

Start with the XMLReader.java lecture example, and adapt it to make XGrep. Use the SAX strategy. Use a Vector or an ArrayList to keep track of the root..match node names.