Stat 133, Fall '05
Homework 5: XML
Due 5 p.m. Monday, 21 Nov

The XML Document Object Model

The main goal of this home work assignment is to read data from an XML format into R to create a data frame. The data are located on the web at

`www.stat.berkeley.edu/users/nolan/stat133/data/counties.gml`

These data contain the X and Y coordinates of the county centers for the counties in the US. These data are in GML (geography markup language) format.

The R object must have the following format:

- a list of elements, one per state, with names corresponding to the state abbreviations, e.g. CA for California.
- Each state element is a data frame consisting of two columns named X and Y, and one row for each county, with the row name being the county name in lower case with all blanks and punctuation removed.
- The values for X and Y for a specific county are the x and y coordinate values in the X and Y tags within the respective county tag.

Provide your code in a plain text file and a plot of the (x,y) coordinates. Mail this to s133@stat.berkeley.edu and drop off a printed copy in class or to the department by 5 pm on Monday.

To complete the assignment you will need to use the R package called XML. Functions that you may find useful are **xmlName**, **xmlSize**, **xmlAttrs**, **xmlValue**, **xmlChildren**, **xmlApply**, **xmlSApply**. There are two acceptable ways to do this.

1. Read the data into R then call a function that operates on the resulting XML-Document object to pull out the information and create the data frame.

2. Write handler functions for the xmlTreeParse function where the handler functions are called as the tree parser encounters the tag names. In this case, there is no need to create an XMLDocument as the data frame is constructed as the nodes are processed.

The second method is worth a possible additional 5 points (recall that homework is graded on a 20 point scale).