Activity 1-1

Jan. 31, 2012

Task 0: Get the Files We Need

1. Visit the "Senate URL" link on the course website. Note the three pieces of information from the url (what are they again?). Right-click on the XML link to download the XML file and save it **to your Desktop**. The option will be `Save Link As...` or `Save Target As...` or something like that (this depends on your web browser). Give it a name that includes the url information and make sure it has '.xml' at the end.

2. Download the JAR file from the course website. Move it **to your Desktop**. This is a program written in a programming language called Java. We're going to run this program.

3. Open Excel. What version of Excel are you using? You can often find the year by pressing 'F1' or going to `File...Help`.

Task 1: Convert an XML File to a CSV File

Do the sections that pertain to your operating system.

1A: Open a Terminal and Navigate to the Desktop (Windows)

1. From the `Start` button (lower left), in `Search Programs and Files` type 'cmd' (without the single quotes) and click on the first item (it will be either `cmd` or `cmd.exe`). This should open up a window that has some text ending in your name followed by a dollar sign. It looks something like this:

   ```
   C:\Users\aritz>
   ```

   This is called a *command prompt* - you can use this to type various commands to the computer that can open, move, and edit files, as well as run programs (among other things).

2. You are in your *home directory*. Type 'dir' (for 'directories') and hit Enter. What do you see?

3. 'Desktop' should be a directory in this list. We want to move to that directory. You can do that by using the 'cd' (for 'change directory') command. Type

   ```
   cd Desktop
   ```

   and hit Enter. The line should now look something like

   ```
   C:\Users\aritz\Desktop>
   ```

   You are now in the 'Desktop' directory, which contains all the files and folders you see on your actual Desktop. By typing 'dir', you should see the XML file and the JAR file.

1A: Open a Terminal and Navigate to the Desktop (Mac)

1. In Finder, go to `Applications/Utilities` and open the application called `Terminal` (if you're on a lab machine, the path is `Cluster/Applications/Utilities`). This will open up a window that has some text ending in your name followed by a dollar sign. It looks something like this:

   ```
   Macintosh-102:~Jonah$
   ```

   This is called a *command prompt* - you can use this to type various commands to the computer that can open, move, and edit files, as well as run programs (among other things).

2. You are in your *home directory*. Type 'ls' (for 'list') and hit Enter. What do you see?

3. 'Desktop' should be a in this list. We want to move to that directory. You can do that by using the 'cd' (for 'change directory') command. Type

   ```
   cd Desktop/
   ```

   and hit Enter. The line should now look something like

   ```
   Macintosh-102:Desktop Jonah$
   ```

   You are now in the 'Desktop' directory, which contains all the files and folders you see on your actual Desktop. By typing 'ls', you should see the XML file and the JAR file.

1B: Convert an XML File into a CSV File (Mac & Windows)

1. From the 'Desktop' directory, you can now run the converter. At the command prompt, type:

   ```
   java -jar xml2csv-conv.jar
   ```

   Here's an explanation of what you just did: `java` tells the computer to run a program written in Java, `-jar` tells the computer it is a JAR file (as opposed to some other format), and `xml2csv-conv.jar` is the name of the program you ran.

   When you ran the above command, you got some kind of error along with a bunch of instructions on how to run the program. This is all useful information, but for now we just want to somehow give this program an XML file and have it produce a CSV file.

   If you did not get instructions on how to run the program, perhaps you do not have the Java programming language installed on your computer. If you want to run this program on your computer at a later date, meet with a TA and they will set you up. For now, do the rest of this task with your neighbor.

2. Now let's run the program so it gives us a CSV file. We do this by giving a set of *arguments* to the program. Suppose your XML file is called myfile.xml and you want a CSV file called myfile.csv. At the command prompt, you would type:

   ```
   java -jar xml2csv-conv.jar myfile.xml myfile.csv
   ```

   This means that you are giving the program myfile.xml and the program will create myfile.csv.

3. Type the command above with your saved XML file. Open the resulting CSV file and look at it. What do you think 'CSV' stands for?

4. Why do you think we needed to put the JAR file and the SML file in the same directory (here, we put both on the Desktop)?

5. But shoot, it's not perfect. Compare the last row in the CSV file with the last member in the XML file. **This is a BUG - a problem in the converter!**. The staff will fix this by modifying the code, but how can *we* fix it?

6. Fix the file and double check the last row of the CSV file is the same as the last member in the XML file.

Task 2: Import a XML/CSV File into Excel

Mac Instructions

1. In Excel, click on `Import` and select `CSV file`. Navigate to the CSV file you made (the one with the correct number of rows!). Select `Delimited`, then click `Next`. Since this is a CSV file, we only want `Comma` to be selected for delimiters; click `Next`. Then click `Finish`. When asked where to put the data, make sure the upper left cell of your worksheet is highlighted, and click `OK`.

2. Tip: If Excel defaults to a print-view, then click `View...Normal` to see the table all at once.

Windows Instructions

1. In Excel, click the `Data` tab. Go to `Get External Data...From Other Sources...From XML Data Import`. Use this to open the XML file that you just downloaded. If you get a warning, click `OK`. When asked where to put the data, make sure the upper left cell of your worksheet is highlighted, and click `OK` again.

Questions (Mac & Windows)

1. Look at the Excel table. How many rows are there?

2. Examine all the columns. How are these related to the XML file itself?

3. Which columns contain useful information?

4. Suppose I have the XML file for Vote #2. How would you add that information to this table?

Task 3: Format the Data

1. Instead of having you load in all the votes from the 109th congress, which takes about 20 minutes, we've done that for you.

2. Right-click and **save** the `congress109_allvotes.xlsx` file on the course website onto the Desktop (saving it allows you to edit it, rather than viewing it immediately). If it got saved as a `.zip` file, change the extension to `.xlsx`.[1] Open `congress109_allvotes.xlsx` in Excel (this may take a few minutes).

3. Get familiar with the data. How many rows are there? How many votes are in the first session? How many votes in the second session?

4. Spend some time deleting columns from the data (unfortunately, you have to do this manually). We'll need the following columns:

   session

   vote_number

   vote_question_text

   member_full

   vote_cast

5. Save the much-reduced spreadsheet with a different, descriptive name (maybe use the word 'trimmed'). Save this spreadsheet often!

6. Now figure out how to delete all the rows of the spreadsheet that correspond to votes that are *not* on the passage of a bill. Hint: this should only take a few minutes; it does *not* involve deleting the non-bill-vote rows one by one.

7. The resulting spreadsheet should be a lot smaller - how many rows are there now?

8. We want a unique identifier for the vote of each bill in this congress. Which two columns *together* produce a unique identifier for each vote?

9. Add another column to the table by entering a `vote_id` column in cell `F1`. Excel automatically infers that this is a new part of the table. In

[1] This problem arises when you use Internet Explorer; if you use Firefox, it should not occur.

row 2 of the `voteID` column, write a formula to combine the "session" for this row with the "vote_number" for this row, placing a colon between them. Use `fill down` or copy/paste, if necessary, to apply this formula to all the other rows.[2] The table should look like this:

	A	B	C	D	E	F
1	session	vote_number	vote_question_text	member_full	vote_cast	vote_id
2	1	9	On Passage...	Akaka (D-HI)	Nay	1:9
3	1	9	On Passage...	Alexander (R-TN)	Yea	1:9
4	...	...	...	...	...	...

10. Add another column to the table, just to the right of `vote_id`, and title it `numerical_vote`. In the second row of that column, enter a formula that has a `0` if the senator did not vote, a `1` if s/he votes "Nay," and a `2` if s/he voted "Yea." If s/he has some other vote than these three, your cell should contain the word "ERROR". Verify that no cell contains the word "ERROR." The table should now look like this:

	A	B	C	D	E	F	G
1	ses...	vote_n...	vote_q...	member_full	vote_cast	vote_id	numerical_vote
2	1	9	On...	Akaka ...	Nay	1:9	1
3	1	9	On...	Alexander ...	Yea	1:9	2
4	...	...	...	...	...	...	

Task 4: Summarize the Data with a Pivot Table

1. For Windows, Click on the `Insert` tab, and insert a pivot table, using the suggested `Table 1` (that's what Excel automatically named our vote data). For Mac, Click on `Data...PivotTable Report...` and select the entire table. Clic `OK`.

2. Let the rows be `member_full`, the columns be `vote_id`, and the value in the cells be the `numerical_vote` that you just created.

[2] Excel may do this automatically because you're working on a table; whether it does so or not depends on how it's been set up.

3. Change the value in the cells to be the maximum rather than the sum of the votes. For Windows, left-click on the top-left corner of the pivot table and select `Value Field Settings`; choose `Max` from the list that appears, and click `OK`. For Mac, click `PivotTable...Field Settings`, select `Max`, and click `OK`.

4. Delete the `Grand Total` row and column. For Windows, select `Option ...PivotTable ...Options ...Option`,select the `Totals & Filters` tab, and uncheck the "show grand totals" checkbox. For Mac, select `PivotTable...Table Options` and uncheck the "grand totals" checkboxes.

5. When you're done, save the spreadsheet and email it to yourself or put it on a flashdrive. We will use this next class.