Sorting Senators, Collecting Data

Sept. 11, 2014
Warmup

• Get Google Spreadsheet from last class open!
## Index Cards

<table>
<thead>
<tr>
<th>Name</th>
<th>Concentration</th>
<th>Where You’re From</th>
<th>Something Interesting About Yourself</th>
</tr>
</thead>
</table>

![Self Portrait](image.png)
Plan

• Pick liberal senator, Senator X
• Compare others to Senator X to determine liberalism
Problem

• What if senator X isn’t the most liberal?
  – Even those more liberal will be rated as some distance from senator X, and hence appear more conservative!
Slight improvement

• Pick liberal Senator X, and conservative Senator A.
• Compare other senators to both of these
• Now a senator more liberal than X will not only be distant from X, but more distant from A than X is
Analogous problem

• Put the stations on Amtrak’s Northeast Corridor in order
• You’re given only
  – Distances between stations
  – An example station near the NE end
  – An example station near the SW end
## Distance table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B (SW)</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G (NE)</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>180</td>
<td>300</td>
<td>220</td>
<td>120</td>
<td>80</td>
<td>165</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>120</td>
<td>40</td>
<td>380</td>
<td>260</td>
<td>345</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0</td>
<td>80</td>
<td>500</td>
<td>380</td>
<td>465</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>0</td>
<td>420</td>
<td>300</td>
<td>385</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>120</td>
<td>35</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>85</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B (SW)</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G (NE)</td>
<td>H</td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>--------</td>
<td>-----</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>180</td>
<td>300</td>
<td>220</td>
<td>120</td>
<td>80</td>
<td>165</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>120</td>
<td>40</td>
<td>380</td>
<td>260</td>
<td>345</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>80</td>
<td>500</td>
<td>380</td>
<td>465</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>420</td>
<td>300</td>
<td>385</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>120</td>
<td>35</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>85</td>
<td>180</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td>265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B (SW)</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G (NE)</td>
<td>H</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>--------</td>
<td>-----</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>180</td>
<td>300</td>
<td>220</td>
<td>120</td>
<td>80</td>
<td>165</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>120</td>
<td>40</td>
<td>380</td>
<td>260</td>
<td>345</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>80</td>
<td>500</td>
<td>380</td>
<td>465</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>420</td>
<td>300</td>
<td>385</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>120</td>
<td>35</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>85</td>
<td>180</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Because BG distance is 345, BA = 180, and AG=165, A must be between them!

What about D?

Spend a minute locating others in order
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B (SW)</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G (NE)</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (NY)</td>
<td>0</td>
<td>180</td>
<td>300</td>
<td>220</td>
<td>120</td>
<td>80</td>
<td>165</td>
<td>100</td>
</tr>
<tr>
<td>B (Balt)</td>
<td>0</td>
<td>120</td>
<td>40</td>
<td>380</td>
<td>260</td>
<td>345</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>C (Richmond)</td>
<td>0</td>
<td>80</td>
<td>500</td>
<td>380</td>
<td>465</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (DC)</td>
<td>0</td>
<td>420</td>
<td>300</td>
<td>385</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E (BOS)</td>
<td>0</td>
<td>120</td>
<td>35</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F (New Haven)</td>
<td>0</td>
<td>85</td>
<td>180</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G (Prov)</td>
<td></td>
<td></td>
<td>0</td>
<td>265</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H (Phila.)</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion?

• Example suggests that we need not pick the most liberal or most conservative senator to do our ranking

• We can use comparisons to find senators “further out”

• Tonight’s homework will suggest otherwise 😞
Collecting Data

• Last class we showed you XML file structure
• Talked briefly about CSV (“comma separated values”) file structure
• Had you load a CSV file
• Promised further info about loading XML
Getting at the contents of an XML file

Structure:
<roll_call_vote>
  <congress>113</congress>
  <session>2</session>
  <congress_year>2014</congress_year>
  <vote_number>8</vote_number>
  <vote_date>January 14, 2014, 03:22 PM</vote_date>
  <modify_date>January 14, 2014, 04:01 PM</modify_date>
  <vote_question_text>On the Motion to Table S. 1845</vote_question_text>
  <vote_document_text>
A bill to provide for the extension of certain unemployment benefits, and for
other purposes.
</vote_document_text>
  <vote_result_text>Motion to Table Failed (45-55)</vote_result_text>
  <question>On the Motion to Table</question>
  <vote_title>
Motion to Table the Motion to Commit S. 1845 to the Committee on Finance with
Instructions
</vote_title>
  <majority_requirement>1/2</majority_requirement>
  <vote_result>Motion to Table Failed</vote_result>

...
Interpretation: A “roll call vote” contains a “members”, which is itself a container, containing many “member”’s. A “path” to senator Alexander’s first name could be written
document/members/member/first_name
...but this would also be a path to any other senator’s first name
XPath

• Listing “tags” separated by slashes is an instance of an “Xpath”, which is a standard for describing locations of data in an XML file.
• Google’s importXML uses this.
Example of importXML

=importXML("http://www.senate.gov/legislative/LIS/roll_call_votes/vote1132/vote_113_2_00008.xml", "/members/member/first_name")

• The URL for the XML file:
  http://www.senate.gov/...008.xml

• The Xpath search string:
  "//members/member/first_name"
Meaning of XPath String

- "//members/member/first_name"
- "//members/member/first_name"
  - Means “any path at all can go here"
  - Full path would be
    "//roll_call_vote/members/member/first_name"
  - Alternative short form that works for this doc:
    "//first_name"
- Different form: "//roll_call_vote/*//*/first_name"
- Any “first_name” that’s a great-grandchild of the roll_call_vote. (“*” means “replace with any one item”)
- Many fancier forms available...if you need them.
Use Bernie Sanders’ votes to compare how liberal other senators are.

So far:

Define Problem

Find Data

Write a set of instructions

Solution

XML Format

Make a HUGE spreadsheet table

Computer (spreadsheet)
Use Bernie Sanders’ votes to compare how liberal other senators are. The solution involves:

1. **Define Problem**: Use Bernie Sanders’ votes to compare how liberal other senators are.
2. **Find Data**: Gather data on Senators’ votes.
3. **Write a set of instructions**: Create instructions for processing the data.
4. **Solution**: Implement the instructions to analyze the data.
   - **XML Format**: Use XML for data exchange.
   - **CSV Format**: Use CSV for readability.
   - **Computer (spreadsheet)**: Use a spreadsheet software for data manipulation.

*Vote on bills only!*

---

CSCI 0931 - Intro. to Comp. for the Humanities and Social Sciences
This is going to be a lab day

• Ask for help/clarification *at any point*. 
Soon you’ll have a big (but not so big) table of votes

XML Format → TAs → CSV Format → you → Make HUGE Table

- Import the data we want into spreadsheet
- Format the table to get what we want
So far, we’ve done this...

XML Format → TAs → CSV Format → you → XML Format

XML Format → Make Small Table

CSV Format → you → Make HUGE Table

Import the data we want into spreadsheet

Format the table to get what we want
A note on names

- We call files by their file *extension*: an XML file ends with `\.xml\`, a CSV file ends with `\.csv\`, etc.
Why?
We’re learning how to gather data off the web, then format into something we can work with.
**Ctrl** is your friend

**Bottommost Cell:** Ctrl and ↓

**Topmost Cell:** Ctrl and ↑

**Rightmost Cell:** Ctrl and →

**Leftmost Cell:** Ctrl and ←

Pressing **Ctrl** selects each cell you click

Shift is your friend too: Pressing **Shift** selects all cells between clicks.

Pressing **Shift** and using arrow keys selects blocks
Activity 1-1

• Proceed from wherever you got to during the last class
• If you finish early, fill out your 3x5 card
• Hint for task 3 (formatting data)

Tip: Press ‘Ctrl’ and an arrow to go ALL THE WAY to the beginning/end of a row/column.

Tip: To get back to original sort order: Sort by both ‘session’ and then by ‘vote_number’
Look at your spreadsheet

1. Open the spreadsheet.
2. You should have 3,501 rows (Task 3.7)
3. You should have columns through E.
Task 3.8

• We want a unique identifier for the vote of each bill in this congress.
  – Which two columns together make a unique key?
Task 3.9

• Add another column to the table by entering a `vote_id` column in cell F1.
• Write a formula to output `session:vote_number` values for this row.
• Use fill down or copy/paste, if necessary, to apply this formula to all the other rows.
Task 3.10

• Add a numerical_vote column in cell G1.
• Write a formula to output:
  – 0 if the senator did not vote
  – 1 if the senator voted Nay
  – 2 if the senator voted Yea
  – Error otherwise
Decision Trees

\[
\text{vote\_cast} = \begin{cases} 
\text{“Not Voting”} & \text{if the senator did not vote} \\
\text{“Nay”} & \text{if the senator voted Nay} \\
\text{“Yea”} & \text{if the senator voted Yea} \\
\text{Error} & \text{otherwise}
\end{cases}
\]

0 if the senator did not vote
1 if the senator voted Nay
2 if the senator voted Yea
Error otherwise
Task 4 (Make a Pivot Table)
What is a pivot table good for?
We did these things!

XML Format → CSV Format → Make HUGE Excel Table

- Import the data we want into spreadsheet
- Format the table to get what we want
Now what?

Compare senators to Bernie
Ranking Senators against Bernie Sanders by comparing Votes
See anything odd in the pivot table?

• **Use** Home...Conditional Formatting to color all blank cells **red**
See anything odd in the pivot table?

- **Use** Home...Conditional Formatting to color all blank cells **red**

- What happened to Schatz and Ensign?
See anything odd in the pivot table?

• **Use** Home...Conditional Formatting to color all blank cells **red**

• Name **csv_output** RawData

• Name **Sheet2** PivotTable

• Make column widths nice:
  - Home-> Cells-> Format->AutoFit Column Width
Activity 1-2

• Do Task 1
Homework

• Will be ready a little later this evening