Introduction to Computation for the Humanities and Social Sciences
Who am I

- PhD Student studying AI, robotics and smart homes
- Call me Stephen
- stephen_brawner@brown.edu
HTA and TAs

HTA: Linda Chang

TAs:

• Palak Goel
• Emma Sloan
• Monica-Ann Mendoza

cs0030tas@lists.cs.brown.edu
Supervising faculty

Shriram Krishnamurthi
sk@cs.brown.edu

John “Spike” Hughes
jfj@cs.brown.edu
What is this course teaches and why you should take it!

• A problem-solving workflow to build defensible arguments backed by relevant data sources and appropriate methods

• How to find and process various data sources, including text, pre-formatted data, and web-based like twitter feeds

• How to solve computational problems using old tools like spreadsheets and some new ones, like the Python programming language

• This course is designed for humanities and social science concentrators. No programming background is expected, and this course won’t use or require advanced math or science.
Problem Solving with Computers

Think and discuss

• Have you ever seen people make claims by presenting graphs or analyses of data?

• Think of projects that involved a dataset and required a computer to analyze the dataset
Lecture topics

Assessing the political leanings of legislators (Google Sheets)
  • Forming computational problems
  • Finding datasets
  • Formulas and pivot tables in spreadsheets
  • Ranking senators according to their voting record similarity

Textual Analysis (Python),
  • Designing and writing programs
  • How data is stored and can be manipulated in a program
  • Iteration through text data
  • Building a concordance of large texts

Data visualization and data collection through Application Program Interfaces (APIs)
  • Plotting data in informative ways
  • Visualizing data on maps
  • Learn how to get data from public APIs
Homework, labs and projects

• Homeworks handed out each class due next class

• Most homeworks have an accompanied in-class lab activity that are crucial to understanding homework

• 3 projects. Each project includes a proposal. The final project also includes a presentation
Grading

Homeworks  50%
Labs (good effort)  10%
First project  10%
Second project  10%
Third project  20%

10% Extra credit available on homework assignments by adding components or analyses
Homework Rubric

• 40 total points per assignment
• 8 points for style (Python)
• 4 points extra credit
Late Policy

• 3 late days for assignments, cannot be used for projects or labs

• If you are sick, provide notification with a note from Health Services before the due date to receive an extension.

• Don’t get behind! Each homework builds on the previous.
What to do if you will miss a class

• Email me for slides

• Try to complete the activity the day before so you can go to TA hours
Course resources

• Website:
  http://cs.brown.edu/courses/csci0030

• Emails
  HTAs (Linda and me), for issues related to course infrastructure
  cs0030headtas@lists.cs.brown.edu

  TAs for individual matters
  cs0030tas@lists.cs.brown.edu

  Email the TA that graded your homework for grading

• Piazza, for questions about homeworks
  • Go to piazza.com, search for csci0030 and sign-up as a student.

• TA Hours, check website

• Meet with me, by appointment
On Political Media Bias
Observations inform research

“In 1981 … I was named a national correspondent, which allowed me to cover bigger, more important stories anywhere in the country … It was in New York that for the first time I started noticing things that made me feel uneasy.

“I noticed that we pointedly identified conservatives as conservatives, for example, but for some crazy reason we didn’t bother to identify liberals as liberals … in the world of the Jenningses and Brokaws and Rathers, conservatives are out of the mainstream and need to be identified. Liberals, on the other hand, are the mainstream and don’t need to be identified.”
Let’s turn this into a research problem

Think, then discuss with your neighbor

• What claim did he make?

• What was his justification for why that claim held true?

• How could we verify his claim?
What is our question?

“In 1981 … I was named a national correspondent, which allowed me to cover bigger, more important stories anywhere in the country … It was in New York that for the first time I started noticing things that made me feel uneasy.

“I noticed that we pointedly identified conservatives as conservatives, for example, but for some crazy reason we didn’t bother to identify liberals as liberals … in the world of the Jenningses and Brokaws and Rathers, conservatives are out of the mainstream and need to be identified. Liberals, on the other hand, are the mainstream and don’t need to be identified.”
What is our question?

“In 1981 … I was named a national correspondent, which allowed me to cover bigger, more important stories anywhere in the country … It was in New York that for the first time I started noticing things that made me feel uneasy.

“I noticed that we pointedly identified conservatives as conservatives, for example, but for some crazy reason we didn’t bother to identify liberals as liberals … in the world of the Jenningses and Brokaws and Rathers, conservatives are out of the mainstream and need to be identified. Liberals, on the other hand, are the mainstream and don’t need to be identified.”
“On the bias”
Geoffrey Nunberg, *Fresh Air*

“Bernard Goldberg is hardly the first person to claim that the media have a liberal bias, and his *Bias* is far from the best-written or best-argued book to try to make that point. Even so, it has climbed to the top of the New York Times bestseller list, maybe because Goldberg is himself a CBS insider with lots of tell-all tidbits to offer about the likes of Dan Rather and Bob Schieffer.

“For the most part, Goldberg's book is a farrago of anecdotes, hearsay, and unsupported generalizations. But at one point he strays into territory that can actually be put to a test. That’s when he claims that the media "pointedly identify conservative politicians as conservatives," but rarely use the word "liberal" to describe liberals.”
Problem Solving Outline

1. Define our question with precise, defined terms that can be proven true or false.

2. Select a relevant data source

3. Design a feasible method that would definitively answer your question

4. Evaluate and analyze the results

5. Communicate your results
Clarifying our question

Think, then discuss with your neighbor

• What is media bias?

• How might bias manifest in different ways in the media?

• To whom/what are they biased?

• What types of media are we interested in analyzing?
Types of Bias

• Omission: Using only arguments from one side
• Source selection: Include more sources or more authoritative sources for one side over the other
• Story selection: Regularly including stories that agree or reinforce the arguments of one side
• Placement: Using the benefit of the perceived importance of position to highlight certain stories
• Labeling: (two types)
  • Using labels to categorize sources or individuals as extreme
  • Labeling people on one side of the argument with labels and not the other
• Spin: Story provides only one interpretation of the events
What is our question?

• Instead of “Is the media biased toward liberals”

• “Are liberal politicians identified as liberals more or less often than conservatives in major newspapers?”
What data sources should we use?

• TV video segments are very hard to analyze, but TV transcripts are significantly easier

• Text from newspaper articles is easy, but newspaper layout and organization is harder

• This course will teach you how to analyze pre-formatted data, text data sources and websites
What data sources should we use?

Think, then discuss with neighbor

• What would be relevant, accessible data sources to solve this problem?

• How do we measure labeling bias in this dataset?

• How do we avoid introducing our own bias by the selection of data sources?
Nunberg

30 Major Newspapers

• The New York Times
• The LA Times
• The Washington Post
• The Boston Globe
• The Miami Herald
• The San Francisco Chronicle
Measuring label bias

Our question
“Are liberal politicians identified as liberals more often than conservatives in major newspapers?”

Nunberg’s Method

• Choose a representative set of politicians from each side
• Find mentions of politicians in the dataset
• Look at seven words before and after the politician’s name
• Compute fraction of times these words around their name contain the politician’s appropriate label
Nunberg’s chosen politicians to follow

**Liberal Politicians**
- Sen. Barbara Boxer
- Sen. Paul Wellstone
- Sen. Tom Harkin
- Sen. Ted Kennedy
- Rep. Barney Frank

**Conservative Politicians**
- Sen. Jesse Helms
- Sen. Tom DeLay
- Sen. John Ashcroft
- Sen. Dick Armey
- Rep. Trent Lott
Looking at words around politician’s name

• There will inevitably be false positives

• These words may include the label even if it doesn’t label the person. Consider:

Which of these examples will be correctly categorized by the method?
• “‘And now it’s gone too far,’ said John Ziegler, a conservative radio host”
• “When David [Brooks] complains that ‘conservative opinion-meisters began to value politics over everything else,’”
• Mr. Obama, according to the system, rates as being slightly more conservative than Jimmy Carter
## Nunberg’s results

<table>
<thead>
<tr>
<th>Liberal Legislators</th>
<th>Total instances in newspapers database</th>
<th>Pct within 7 words of relevant label</th>
<th>Total instances in &quot;liberal&quot; papers</th>
<th>Pct. within 7 words of label in &quot;liberal&quot; papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul Wellstone</td>
<td>2939</td>
<td>10.9%</td>
<td>578</td>
<td>8.48%</td>
</tr>
<tr>
<td>Barney Frank</td>
<td>8501</td>
<td>4.7%</td>
<td>1439</td>
<td>3.89%</td>
</tr>
<tr>
<td>Tom Harkin</td>
<td>10,147</td>
<td>3.7%</td>
<td>1784</td>
<td>2.02%</td>
</tr>
<tr>
<td>Ted Kennedy</td>
<td>17,197</td>
<td>3.0%</td>
<td>2444</td>
<td>2.74%</td>
</tr>
<tr>
<td>Barbara Boxer</td>
<td>8977</td>
<td>2.0%</td>
<td>3093</td>
<td>1.78%</td>
</tr>
<tr>
<td><strong>Avg. pct. for liberals, all papers</strong></td>
<td>4.8%</td>
<td></td>
<td><strong>Avg pctl in &quot;liberal&quot; papers</strong></td>
<td>3.78%</td>
</tr>
</tbody>
</table>

## Conservative Legislators

<table>
<thead>
<tr>
<th>Conservative Legislators</th>
<th>Total instances in newspapers database</th>
<th>Pct within 7 words of relevant label</th>
<th>Total instances in &quot;liberal&quot; papers</th>
<th>Pct. within 7 words of label in &quot;liberal&quot; papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jesse Helms</td>
<td>19,874</td>
<td>9.1%</td>
<td>4718</td>
<td>6.02%</td>
</tr>
<tr>
<td>Tom DeLay</td>
<td>6351</td>
<td>3.6%</td>
<td>1859</td>
<td>2.90%</td>
</tr>
<tr>
<td>John Ashcroft</td>
<td>10,187</td>
<td>2.1%</td>
<td>1157</td>
<td>3.03%</td>
</tr>
<tr>
<td>Dick Armey</td>
<td>9222</td>
<td>2.1%</td>
<td>1460</td>
<td>1.44%</td>
</tr>
<tr>
<td>Trent Lott</td>
<td>18,048</td>
<td>1.4%</td>
<td>4976</td>
<td>1.05%</td>
</tr>
<tr>
<td><strong>Avg. pct. for conservatives, all papers</strong></td>
<td>3.6%</td>
<td></td>
<td><strong>Avg pctl in &quot;liberal&quot; papers</strong></td>
<td>2.89%</td>
</tr>
</tbody>
</table>

http://people.ischool.berkeley.edu/~nunberg/table.html
Using Google Spreadsheets
Getting started

• If you don’t already have one, create a google account.

• We highly recommended Google Chrome

• Install Google Drive web app (Chrome only). This allows a few features that aren’t available otherwise
Quick Demo
Cell Value Types

Numerical Types
- Right Justified
- Numbers
- Percents (1.5%)
- Currencies ($3.00)
- Scientific Notation (1.01E+03)
- Date (01/26/2016)
- Time (11:05:00 AM)

Text
- Left Justified
- “:-)"
- Hello
- 10A90
- Numbers can be converted into text and vice-versa
Formulas

• Formulas allow you to compute a value from references to other cells

• Formulas can as simple as a single reference or can be complex operations like downloading data from the web

• Formulas automatically recompute when the cells they reference are updated

• Copying and pasting a cell’s formula updates the formulas references
References

- Cell references are [COLUMN][ROW]

- You can reference cells in other sheets by adding NameOfSheet! at the beginning

- Use a $ to lock the column, row or both when copy and pasting or doing fills
References and Naming Cells

- Cell references [COLUMN][ROW]
- You can reference cells in other sheets by adding Sheet#! at the beginning
- Use a $ to lock the column, row or both when copy and pasting or doing fills
Notes

• Use notes to briefly explain reasoning behind formulas and context of the cell
Comments

- Use comments for detailed discussions with collaborators, usually with regards to correctness or style.
Fills

- Gives an alternative to copy and paste for many cells
Functions

- Functions are named computations you can do over a single cell or a range of cells
- Include a name, inputs and an output
- Consult documentation for specific information about each function
- Google Sheet’s documentation list of functions

https://support.google.com/docs/table/25273?hl=en
Function Documentation

- Can find from Google’s functions list
- Find via google by searching for “google sheets FUNCTION”
- Includes information about the inputs and the expected outputs
- Notes can add some clarity

**TEXT**

Converts a number into text according to a specified format.

**Sample Usage**

```
TEXT(1.23, "$0.00")
TEXT(A2, "#.##")
TEXT(24, "#.0")
TEXT(DATE(1969, 7, 20), "yyyy-MM")
```

**Syntax**

```
TEXT(number, format)
```

- **number** - The number, date, or time to format.
- **format** - The pattern by which to format the number, enclosed in quotation marks.

- 0 forces display of zeros if a number has fewer digits than the format specifies. For example, `TEXT(12.3, "000.00")` produces 012.30. Numbers which have more digits to the right of the decimal point than the pattern are rounded to the indicated number of places. For example, `TEXT(12.305, "00.00")` results in 12.31.

- # is similar to 0 but does not force the display of zeros on either side of the decimal point. For example, `TEXT(12.3, "###.##")` produces 12.3.

**Notes**

- The format argument to TEXT cannot contain an asterisk (*).
- TEXT does not support the ? pattern in Google Sheets.
Ranges

• A reference to more than one cell at the same time

• Format is Cell1:Cell2

• Sheets will detect whether you mean by column, row or both

• $ Works similarly as the cell reference, but you should include it on both cells

MAX(A1:A5) = 5
MAX(A1:C1) = 5
MAX(A1:C5) = 9
Named Ranges

• Use Named Ranges to provide clarity in formulas

• Instead of D1:D5, here we can call it a meaningful name

• Name cannot contain spaces or other non-alphabetic characters (!, @, #, etc)
Conditional Functions

- Conditional functions test truth values (TRUE/FALSE, 1/0)

- IF takes in three arguments, a logical expression, a value returned if the logical expression is TRUE and value returned if FALSE

- Logical expression will evaluate to either TRUE or FALSE
NOT, AND, and/or OR

- **AND** returns TRUE if all of its inputs are TRUE.
- **OR** returns TRUE if any of its inputs are TRUE.
- **NOT** inverts the input value.
Pivot Tables

- Provide a fast method to interactively explore the data
- Choose the row or column labels
- Choose a summarization method (max, min, sum, etc)
- Can optionally choose a filter value
### Pivot Tables

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Row 1</strong></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Row 2</strong></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Row 3</strong></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Row 4</strong></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Row 5</strong></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
String concatenation

- Compose text strings using a &

- Values that aren’t text strings don’t need to be explicitly converted to text, unless you need them specifically formatted.
Style guidelines

• Name your sheets

• Use notes to describe what you think the formulas are doing

• Use NamedRanges to communicate to yourself the components of a formula

• In general, data types should only vary across a row or across a column but not both.
  • Exception if it is a column or row header
  • Exception if its a column of static constants

• Create new sheets liberally
  • Use new sheets to prevent significant scrolling
  • Use new sheets to group similar computations

• When writing long text, merge cells as needed, and wrap text