Getting Familiar with Preprocessing

EDP 618 Week 11

Dr. Abhik Roy

Preprocessing

Spreadsheets

- Enter variables in columns
- Limit yourself to one value (piece of information) per cell
- Try to avoid using multiple sheets if possible
- Do not create a table within a sheet
- Avoid merging cells
- For dates, follow ISO 8601 formatting **YYYYMMDD** or ISO 8601 extended formatting **YYYYMMDDhhmmss** (e.g. Nov 7, 2022 becomes 20221107)
- Always save as a `.csv` file!

Anonymizing Data

What is good and bad about these anonymiziation attempts?

So my first workplace was *X* which was about *X* minutes from my home in *X*. My best colleagues from day one were *X*, *X* and *X* and in fact, I am still very good friends with *X* to this day. *X* lives in the some parish still with her husband *X* and their *X X*.

So my first workplace was [name] which was about 20 minutes from my home in Norwich. My best colleagues from day one were Anna, Julie, and Louise and in fact, I am still very good friends with Julie to this day. She lives in the some parish still with her husband Owen and their son Ryan.

Tips

- Switch direct identifiers (name, dote, place) or generalize them
  - **Anna** to **[Beth]** or **[Friend 1]**
  - **Julie** to **[Nicole]** or **[Friend 2]**
- Remote or generalize indirect identifiers, which con ID when combined
  - **born in 1986** to **born in [the 1980s]**
  - **a white data librarian** to **white [academic] librarian**
  - **in Morgantown** to **in [a Northern Region]**, **[West Virginia]**, or **[Appalachia]**
- Consider removing sensitive and/or
  - **my teacher is the worst** to **[comment redacted]**
  - **I hid all of my lottery winnings under my bed** to **[location redacted]**

Taguette

- Free and open source
- Custom toggling options
- Lower barrier to entry
- Data and tags stay your own or can be shared
- Mostly easy to import and export
- Does not currently support audio, video, image, or spreadsheet files

Taguette uses Calibre to convert your documents to HTML for tagging. It can work with

- Microsoft Word files (`.docx`)
- LibreOffice files (`.odt`)
- PDFs (`.pdf`)
- Plain text (`.txt`)
- Rich text (`.rtf`)
- HTML (`.html`)

Task (Part I)

1. Open up Taguette (preferably using the server but a local copy will do)
2. Find a partner and settle on one of the open text data sets
3. Select **Create a new project**, name it whatever you want, and add your name
4. Add codes to the text
5. After finishing, go to **Project info > Export project** and send that file to your partner
6. Load your partner's project by selecting **Home > Import a project file**

Before You Go!

1. Open RStudio
2. In the Console, run `install.packages("remotes")`
3. Then run `remotes::install_github("ropenscilabs/qcoder")`
4. Take a look at the examples on the `qcoder` documentation page

That's It!

Any questions?
