LESSON PLAN: EXPLORING SCIENTIFIC ORGANIZATIONS’ METADATA HEADINGS WITH BASIC TEXT MINING IN R

Lesson motivation:

In this lesson, learners will be introduced to the technical skills needed to perform basic text mining methods, notably cleaning data and looking for common words. As a vehicle to teach these skills, learners will be using the R language (in the IDE RStudio), and they will be using those text mining methods to answer these questions: “What metadata headings do scientific organizations use for their various datasets? How do they relate to each other?”

The instructor will use the NASA (National Aeronautics and Space Administration) metadata headings as an example for live coding and teaching. The CERN LHC (L’Organisation européene pour la recherche nucléaire, Large Hadron Collider) data and ArXiv’s metadata headings can also be used as an example for living coding or as homework to practice the skills learned. The packages used here will be <dplyr>, <tidytext> and <jsonlite>.

In the past, mining metadata headers and content has been useful for creating machine learning models the predict new metadata headers and content, such as for Youtube’s automatic categorization system (Algur and Bhat 2016), or for understanding the distribution of scientific research (Skluzacek et al. 2022).

Lesson materials:

Please refer to the attached folder for these files:

1. A lesson flowchart to distribute to learners so that they may understand each step of the learning process. This is in a file named ‘1. flowchart_of_lesson_for_students.’
2. An RStudio script (code) for as an example of how the lesson should be structured. This is in a file named ‘2. rscript_metadata_headings’. Ideally, when live-coding with the students, the instructor will begin with a blank RStudio and walk through this lesson line-by-line.
3. A video walkthrough of the code for instructor’s use only. This walkthrough explains each section of the code and shows instructions how they can use it to teach while live coding. This is in a file named ‘3. video_walkthrough’.

This lesson was designed in RStudio 2022.07.2+576, using <tidytext> 0.4.1.

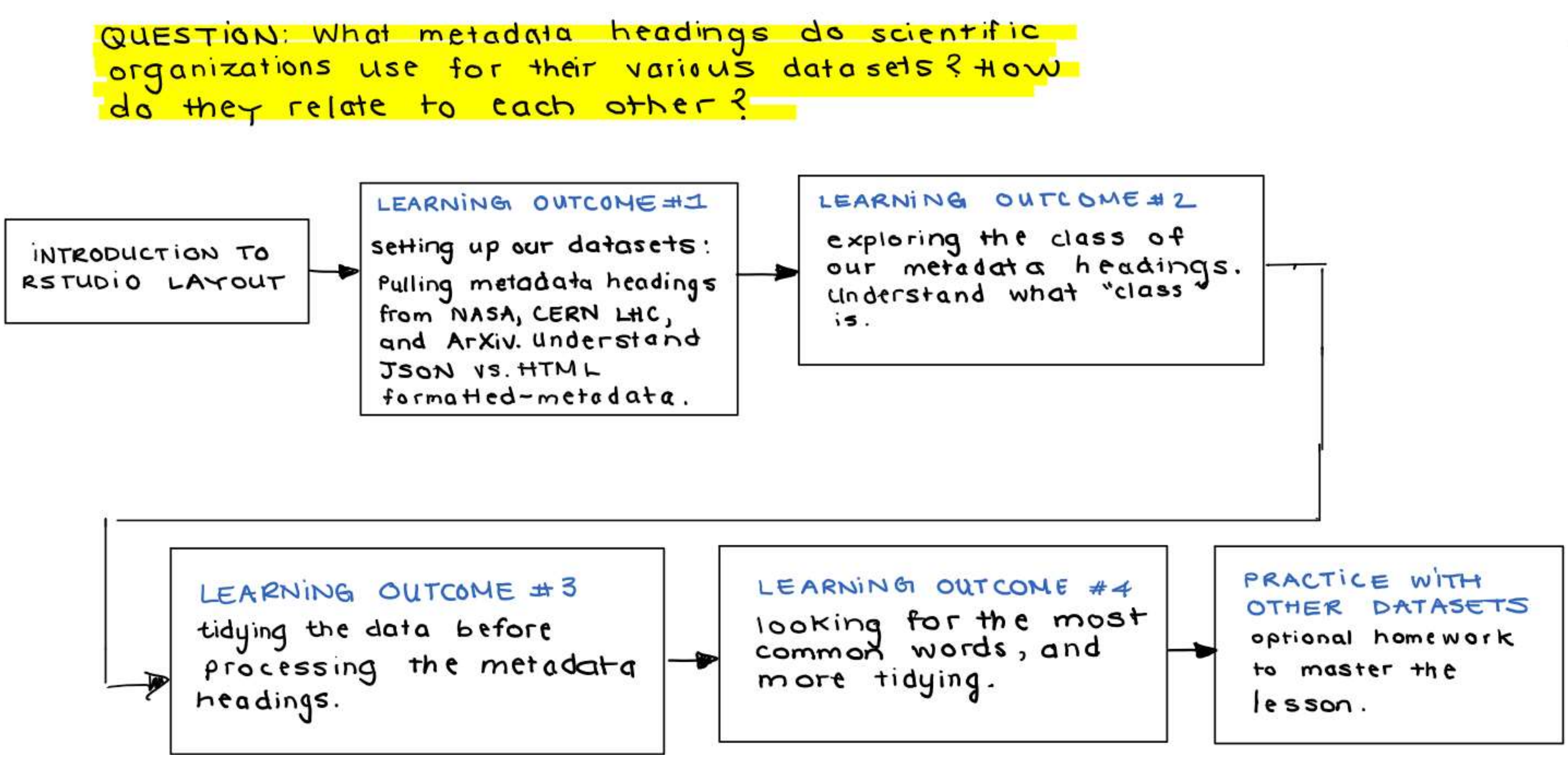
Target learners:

This lesson has been developed for colleagues to teach other colleages (i.e. fellow librarians) with little to no experience coding. Some familiarity with basic coding terms, like ‘package,’ ‘variables,’ and ‘class’ will be handy, but this lesson plan also provides definitions for the instructor to share with the learners.

Lesson length: 45 minutes – 1 hour of live coding.

Lesson structure:

See the next page for the flowchart of the lesson.



QUESTION: What metadata headings do scientific organizations use for their various datasets? How do they relate to each other?

INTRODUCTION TO RSTUDIO LAYOUT

LEARNING OUTCOME #1

setting up our datasets: Pulling metadata headings from NASA, CERN LHC, and ArXiv. Understand JSON vs. HTML formatted-metadata.

LEARNING OUTCOME #2

exploring the class of our metadata headings. Understand what "class is.

LEARNING OUTCOME #3

tidying the data before processing the metadata headings.

LEARNING OUTCOME #4

looking for the most common words, and more tidying.

PRACTICE WITH OTHER DATASETS

optional homework to master the lesson.

Figure 1. How the lesson should flow. This image is also available in the lesson materials for distribution to students.

Lesson content:

In the lesson materials, please refer to the provided fully-commented R code for explanations of each lesson goal. The code naturally provides the commands that you should be using to live code for your learners. It is recommended that you encourage learners to write their own code from scratch while watching you code live.

Common errors, questions, and anticipated challenges

1. *Learner's code does not run; what general debugging tips are there?*
    - "Warning" messages caution users that something is going on, but does not stop the script from running. Oftentimes the warnings provide valuable information about something that will become an issue. Learners should be taught to always understand why a warning is appearing and how to discern if they need to take follow-up actions.
    - "Error" messages stop the code from running. Like warning messages, learners should be taught to always understand why an error is appearing and how to discern what follow-up actions they must take.
    - "Typos" are the most common types of mistakes that your learners will make. They should be taught to carefully compare the lesson code and their own code, and to ensure that spacing, spelling, and grammar matches.

2. *The website URLs provided in this lesson plan/attached code does not link to the datasets for metadata at NASA, CERN, or ArXiv anymore. What else can I use to teach this lesson?*

    You (the instructor) can find similar datasets for any large scientific organization with a mandate to share their data publicly. Oftentimes there will be an open guide on how to access this information, or the organization's information professional(s) can help you. The keywords for these datasets shared in this lesson are "metadata headers," "metadata as JSON," and "metadata schema."

    CERN's mandate can be viewed here ("CERN Open Science Policy" 2022) and NASA's mandate can be viewed here ("Strategy for Data Management and Computing for Groundbreaking Science 2019-2024" 2019).

3. *Do we have to use the <tidytext> package for text mining? What are our other options to get to the same goals that we have here – i.e. word count and tidying data?*

    There are other packages for text mining, and this lesson can be adjusted to fit those packages' commands. For example, <quanteda> and <tm> is popular in the digital humanities (Feinerer 2023; Benoit 2022). However, not all packages can work smoothly with each other, and <tidytext> works well with several other packages in the <tidyverse>; so it is recommended that learners use <tidytext> first.
4. *How can I assess the learning outcomes of the learners?*

    By the end of the lesson, learners should be able to apply these same commands, using this same package, to other JSON-formatted metadata from any source.

References

Algur, Siddu P., and Prashant Bhat. 2016. "Web Video Mining: Metadata Predictive Analysis Using Classification Techniques." *Information Technology and Computer Science* 2: 69–77. https://doi.org/10.5815/ijitcs.2016.02.09.

Benoit, Kenneth. 2022. "Quantitative Analysis of Textual Data [R Package Quanteda Version 3.2.4]." Comprehensive R Archive Network (CRAN). December 8, 2022. https://CRAN.R-project.org/package=quanteda.

"CERN Open Science Policy." 2022. October 1, 2022. https://cds.cern.ch/record/2835057/files/CERN-OPEN-2022-013.pdf.

Feinerer, Ingo. 2023. "CRAN - Package Tm." February 5, 2023. https://cran.r-project.org/web/packages/tm/index.html.

Skluzacek, Tyler J., Matthew Chen, Erica Hsu, Kyle Chard, and Ian Foster. 2022. "Models and Metrics for Mining Meaningful Metadata." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13350 LNCS: 417–30. https://doi.org/10.1007/978-3-031-08751-6_30/COVER.

"Strategy for Data Management and Computing for Groundbreaking Science 2019-2024." 2019. 2019. https://science.nasa.gov/science-red/s3fs-public/atoms/files/SDMWG%20Strategy_Final.pdf.