Textbook Analysis Case Study Rubric

DS 4002 – Fall 2024 – Allison Kerper
Due: December 9
Submission Format: upload link to GitHub repo

Why am I doing this?
This case study will allow you to leverage your data science knowledge by using sentiment analysis techniques to identify how language varies in different history textbooks. Despite being displayed as fact, history textbooks often contain biases that can impact a students' education. You will use this sentiment analysis to see how textbooks taught in different settings portray historical events differently.

What am I going to do?
The GitHub repository for this case study can be found at https://github.com/abkerper/DS4002_CS3/. You will obtain a collection of textbooks, separated by one of two categories. You can either choose to compare textbooks by grade level, such as those taught in colleges and high schools. Alternatively, you can compare textbooks by the region in which they are taught. You should aim to acquire a nearly equal number of textbooks for your two points of comparison. You will then select one historical event to focus your analysis on. You should construct a null hypothesis with a baseline to measure whether your difference in sentiment scores is statistically significant.

From the pdfs of each textbook, collect the passage or page that describes this event. You will construct a CSV file that contains each textbook name, ISBN, author(s), publisher, grade level/location (depending on your comparison), and passage. With this file, you will use python and a sentiment analysis (ex. VADER package) to calculate the compound sentiment score of each passage. This will be added as a new column to your dataset. You can also explore other variables to analyze, such as the word count of each passage. You will the calculate the average sentiment score of your two samples and determine the difference between them. Compare this number to your null hypothesis to report your results.

Your final deliverables should include:

- The dataset you constructed with information on each textbook
- A data dictionary
- A visual displaying the data points of your sentiment analysis
- Well documented, commented source code
- A GitHub repository containing all materials used, including your final results

Tips for Success:

- Try your best to obtain a large dataset, as this will be the most difficult part of your project. There are a variety of resources online to download pdfs. You should also ensure that with each textbook, you can clearly identify which sample it belongs to in your comparison.
- Make sure you select a significant historical event that will be discussed in each textbook you acquire.
- Try to investigate other variables in your dataset. See if there exists any other correlations with the textbooks' sentiment scores.

How will I know I succeeded?
You will meet expectations on this case study when you successfully follow and complete the criteria in the rubric below.

Spec Category	Spec Details
Formatting	• One GitHub repository (submitted via link on Canvas) • Create a new GitHub repository for this assignment that contains: o A README.md file (which auto displays) o A LICENSE.md file (use MIT as default) o A SCRIPTS folder o A DATA folder (with the CSV file you created) o AN OUTPUT folder
README.md	• Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings. • Use markdown headers to divide content. • Section 1: Outline This does not need to be very detailed. You should describe your process acquiring your data, creating your dataset, and performing your analysis. • Section 2: Documentation map In this section, you should provide an outline or tree illustrating the hierarchy of folders and subfolders contained in your Project Folder, and listing the files stored in each folder or subfolder. • Section 3: References At the bottom, include a section citing any references used in your project. Use IEEE citation style.
LICENSE.md	• Goal: This file explains to a visitor the terms under which they may use and cite your repository. • Select an appropriate license from the GitHub options list on repository creation (usually, the MIT license is appropriate)

SCRIPTS folder	• Goal: This folder contains all the source code for your project. • Include all the scripts you used. Try to name each script according to the order it needs to be executed to reproduce the results. • All script files should include header comments at the beginning of a script to provide information that anyone working with or executing the script should be aware of. Throughout all your scripts, you should include copious comments explaining what each command or sequence of commands accomplishes and what the purpose is.
DATA folder	• Goal: This folder contains all of the data for this project. • Your data will likely not all fit in GitHub (all of the textbook pdfs). In that case, you can simply include the CSV file you created o If organized correctly, this file will have all necessary information to reference the textbook for each selected passage
OUTPUT folder	• Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. • This should include something that outlines the results of your analysis, stating whether you found a statistically significant difference in your comparison. • Use informative names for your files.

Acknowledgements: Special thanks to Alex Gates and Ravza Aykan for providing models for this rubric structure.