Case Study - Sentiment Scoring for Star Wars Series

Due: TBD

Why am I doing this? While you may be early on in your Data Science education, you were brought to this major due to your commitment to problem solving and creating something on your own. The journey may be far ahead of you, but this case study will show you what it is like to combine your knowledge of datasets, analysis, modeling, and gathering key insights to create insightful deliverables from real-world problems.

- Course Learning Objective: Analyze text data using VADER and NRC Emotion Lexicon sentiment scoring.

What am I going to do? For this assignment, you will start out by building the knowledge required to conduct this Case Study. After looking into generational differences and an in-depth look at sentiment analysis through Harry Potter novels, you will then begin making your way through the data and scripts, first cleaning the scripts and then running both VADER and NRC Emotion Lexicon sentiment analysis. You will have the opportunity to experiment with different types of visualization tools in order to derive insights into levels of different emotions found in each Star Wars film. Be sure to thoroughly read the rubric, hook document, articles, and the comments that take you through the script, with the end goal being a holistic understanding of the material.

Tips for success:

- **Don't get too myopic**. There may be bits of code or some ideas that you will not have had much experience with. Don't spend too much time worrying about understanding everything, the experience is what matters!
- **Be creative**. Once you have made your way through the project, if there's anything else that interests you or there may be something you'd like to add, please do! Fully interacting with the study will come with great results.
- **Read everything**. This assignment has a great deal of information behind it, and finding out exactly what you are doing throughout the process will help you gain the most from the study.
- **Take your time**. This assignment requires first taking in a great deal of information before starting your analysis. Skipping or not spending enough time on early steps can lead to issues later on.

How will I know I have Succeeded? You will meet expectations on Case Study–Sentiment Scoring for Star Wars Series when you follow the criteria in the rubric below.

Spec Category	Spec Details

Formatting	● Repository – A GitHub online folder with materials necessary for completion of this case study. ○ Scripts ○ Data ○ Outputs ○ README file
Scripts	● Goal: This folder contains all source code for your project. ● Include all the scripts 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 each 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	● Goal: This folder contains all the data for this project. ● You should include the initial data, and the final data analyzed. o Depending on your details things may shift, sometimes a new data set arises, sometimes you just produce a secondary data set based on the first one o If needed, the code in the SCRIPTS folder should be able to get you from the initial data to the final one. (e.g. you may have a cleaning script) ● If your data fits in github, place all of it here. ● If your data does not fit in GitHub use a single file explaining the process to obtain the dataset. (e.g. you may host it on One Drive, Google Drive, etc.) ● A Data Appendix file as a PDF, which will include text that you type, as well as tables, figures, and other descriptive statistics. ○ This file should be organized in sections, with a section for each dataset analyzed. Each section should begin with a statement of what the unit of observation is--that is, it should explain what kind of object each row of the data file represents
Outputs	● Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. ● Importantly, any information like tables, figures shown in your presentation should be here.

	• Use informative names for your files

README File	• 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.
	• Make an H2 (##) section explaining the contents of the repository
	• Section 1: Software and platform section
	  ○ The type(s) of software you used for the project.
	  ○ The names of any add-on packages that need to be installed with the software.
	  ○ The platform (e.g., Windows, Mac, or Linux) you used.
	• Section 2: A Map of your documentation. 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: Instructions for reproducing your results. In this section, you should give explicit step-by-step instructions to reproduce the Results of your study. These instructions should be written in straightforward plain English, but they must be concise, but detailed and precise enough, to make it possible for an interested user to reproduce your results without much difficulty.

References	• All references should be listed at the end of the document

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled from Streifer & Palmer (2020).