Systems & Technologies: Python

DS 64502

UNIVERSITY OF NOTRE DAME | DATA SCIENCE

Frederick
Fall 2018

SYSTEMS & TECHNOLOGIES: PYTHON – **FINAL PROJECT**

Overview

Using the Yelp business dataset, your job is to pitch a new business to investors. You can start any type of business in any location you'd like. This is an individual project.

There are three parts to this final project:

Part 1: Exploratory Data Analysis (EDA)
Part 2: "Start A Business" Presentation
Part 3: Predict the Star Ratings

During the last week of the course (Week 7), you will present your 3-slide PowerPoint pitch to a small group of your peers (5 minutes). You will also peer review your (ipynb) code with that group (5 minutes).

The Data

"The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps." (From yelp)

Dataset description:
https://www.yelp.com/dataset/challenge (dataset in json format)

Usage restrictions:
Yelp Dataset Terms of Use (PDF)
Note: You must read and agree to the Dataset License (Terms of Use) in order to download the dataset.

Part 1: Exploratory Data Analysis (EDA)

Using the Yelp business dataset, perform exploratory data analysis and generate a new business idea. Based on your analysis, generate at least two data visualizations, using matplotlib, to help you communicate in your pitch.

- Store your code in a file named **eda.ipynb**

Part 2: “Start A Business” Presentation

Using the Yelp business dataset, develop a three-slide presentation pitch (5 minutes) for a new business. You will need to justify your business proposal, using the results of your exploratory data analysis, with three PowerPoint slides.

1. Describe your proposed business.
2. Provide your data-driven explanation of why this business will get high ratings.

Your presentation must include **at least two data visualizations** which you have generated using matplotlib.

If the work of others is quoted, be sure to cite your sources either quoted in the body of the presentation or as a possible 4th slide.

- Create your presentation as a PowerPoint file, which you will convert to PDF for submission as **pitch.pdf** *

Part 3: Predict The Stars For These Businesses

The dataset **yelp_business_official_test_empty.csv** contains 8 new businesses that are not found in dataset provided by Yelp. These 8 businesses are missing their star rating. Using what we’ve learned in this course, you are to “predict” the stars for these 8 new businesses.

For example, if a new business was a fast food restaurant in Austin, TX, you might “predict” the star rating by assigning the average rating for similar restaurants in Austin, TX.

- Store the code that generates the predictions in **star_prediction.ipynb**

Additional rules for predicting the star ratings:

- You can use ONLY what we have learned in this course.
- You can use ONLY the data in the dataset.

Evaluation & Grading

Note that this is an individual project. Group work will not be accepted.

Project Component	Percent of project grade
Presentation (5-minute new business presentation pitch)	25%

Exploratory Data Analysis (EDA)	25%
Star Ratings Predictions	25%
Code/Script Quality and Correctness	25%
Total	100%

Some Notes on Code Quality:

- Is the code repeatable?
- Are you using vectorization where appropriate?
- Did you use functions, where appropriate?
- Did you properly document your functions?
- Did you properly comment your code?
- Do your variable names make sense?

Project Grading Rubric

	Excellent	Good	Satisfactory	Below Satisfactory
Presentation	Effective storytelling techniques are employed in the presentation of information. There is a clear and comprehensive understanding of the work that was carried out. Presentation demonstrates interesting and nuanced insights into the dataset. Slidedeck is very well-organized and very effectively conveys information and looks professional.	Some storytelling techniques are employed when presenting information. There is a clear understanding of the work that was carried out. Presentation demonstrates a some insights into the dataset. Slidedeck is well-organized and effectively conveys information.	Effective storytelling techniques are not employed in the presentation of the information. There is some understanding of the work that was carried out, but few insights into the dataset. There is a slidedeck but it is not well-organized and/or does not effectively convey information.	Little or no evidence of storytelling techniques employed in the presentation. No clear rationale as to why or how work was carried out. There are no insights into the datasets. There is no slidedeck.

EDA (iPython Notebook)	Provides a complete description of the analysis process, good insights into the analysis including what worked and what didn't, and an overall accurate and comprehensive understanding of the analysis. Appropriate tone for intended audience. Few errors in spelling or grammar.	There is a semi-complete description of the analysis process, decent insights into the analysis including what worked and didn't work, and an accurate understanding of the analysis. Appropriate tone for intended audience. Few errors in spelling or grammar.	Parts of the description of the analysis process are missing. Lacking in number and quality of insights into the analysis including Written in a tone appropriate for the audience. Some grammatical and spelling errors throughout.	Little to no description of the analysis process. An inaccurate understanding of the analysis. Extensive grammatical and spelling errors throughout. Inappropriate tone for intended audience.
Star Ratings Prediction (iPython Notebook)	Provides a complete description of the prediction process, good insights into the analysis including what worked and what didn't, and an overall accurate and comprehensive understanding of the method used to predict. Appropriate tone for intended audience. Few errors in spelling or grammar.	There is a semi-complete description of the prediction process, decent insights into the analysis including what worked and didn't work, and an accurate understanding of the analysis. Appropriate tone for intended audience. Few errors in spelling or grammar.	Parts of the description of the prediction process are missing. Lacking in number and quality of insights into the analysis including Written in a tone appropriate for the audience. Some grammatical and spelling errors throughout.	Little to no description of the prediction process. An inaccurate understanding of the analysis. Extensive grammatical and spelling errors throughout. Inappropriate tone for intended audience.
Code/Script Quality and Correctness	Excellent quality code. Good usage of spaces and comments. Can be rerun and produce the same results reported. Good use of functions and vectorization.	Good quality code. Adequate usages of spaces and comments. Code can be rerun and produce the same results reported. Adequate usage of functions and vectorization.	Poor quality code. Little to no usage of spaces and comments. Can be rerun, but does not produce the same results reported.	No code or the code does not compile due to errors.

This page intentionally left blank.