Stat 628: Data Science Practicum

Module 3 Goals and Data Background

Data Background:

Yelp is an Internet company founded in 2004 to “help people find great local businesses” by providing a platform for users to write reviews of businesses. As of Q4 2021, there are more than 244 million reviews.

Recently, Yelp has released some of its reviews to the public as a part of the “Yelp Dataset Challenge.” Note that only reviews that were recommended at the time of the data collection are included; see link here for how Yelp recommends reviews to its users <https://www.yelp-support.com/Recommended_Reviews>. Also, Yelp utilizes automated software to filter out bad reviews; see link here for more details. <https://www.yelp-support.com/article/What-is-Yelp-s-recommendation-software?l=en_US>.

Project Goals:

The two goals for the project are

1. Provide **useful, analytics** to \*\***business owners**\*\* on Yelp from the datasets and, based on these insights, propose **data-driven action plan/recommendations** to said owners that can improve their ratings on Yelp.
2. Build a web dashboard/widget/web application that **visualizes** your analysis and **makes the analytics easier** to understand for \*\***business owners\*\***

Note that your are providing data analytics to **business owners** who may have only taken basic statistics/quantitative courses a long time ago.

Variable Names and Predictors

We included the following JSON files for you. JSON files are one of the de-facto standards of transmitting data on the Web and this module will help you learn how to manipulate JSON files. The JSON files are

1. **review json**: this contains the reviews. There are 6,990,280 entries of reviews.
2. **business.json**: this contains information about businesses. There are 150,346 entries of businesses.
3. **user.json**: this contains information about users. There are 2,189,457 entries of users.
4. **tip.json**: this contains information about tips written by users on a business. There are 908,915 tips generated by users.

See <https://www.yelp.com/dataset/documentation/main> for details on how the JSONs are structured.

The data is not cleaned nor is it merged. Your task is to manipulate this data in a way that serves the project’s goals.

Also, you are **strongly encouraged** to come up with new, relevant features based on the data we have given you, especially the review text. HOWEVER, you must follow the academic integrity guidelines set below.

Link to Data:

The four JSON files can be downloaded from the following link to Box.

<https://uwmadison.box.com/s/2bpkip5e71nna1peb6plx54wbmlgtro8>

You may be asked to log into UW-Madison’s portal before accessing the dataset. Please allow ample time to download the dataset; the total (compressed) zipfile for download is roughly about 5 Gbs.

Please do not share this link to those who are not in the class.

Rules and Academic Integrity

Each student assumes the responsibilities of an active participant in UW-Madison’s community of scholars in which everyone’s academic work and behavior are held to the highest academic integrity standards. Academic misconduct compromises the integrity of the university. Cheating, fabrication, plagiarism, sabotaging other groups’ work, unauthorized collaboration, and helping others commit these acts are examples of academic misconduct. Specific examples include, but are not limited to,

1. Copying, plagiarizing, stealing, fabricating any of the deliverables, especially the code, from other groups, students outside of the class, or the Internet. In particular, while you may ask other groups for general ideas and questions, you cannot ask for help cleaning the data set, analyzing the dataset, and doing other activities that would be inconsistent with the academic integrity at UW-Madison. If you are unsure, you are always welcome to ask the TA or the professor.
2. Using unauthorized sources, including the original Yelp dataset on Yelp’s website or the original ratings (or summaries of ratings of businesses) which can be derived from Yelp’s website. You are also not allowed to directly copy, steal, plagiarize, paraphrase, or use any analysis that was already conducted on the Yelp data by others (e.g. data science courses online, someone’s blog post or R markdown, Google Cloud’s API platform for sentiment analysis, any pre-written software/code that does sentiment analysis automatically, etc.).

However, you are **strongly encouraged** to browse through Yelp, resources on natural language processing (NLP), sentiment analysis, and other researchers’ analysis of the Yelp data and gather **background information**. You are strongly encouraged to use the information from your background research **to complement** your own analysis and **provide proper attributions**. In short, your analysis of the data must be **original** and **must be your own work**. Or, in industry-lingo, you should not be stealing others’ intellectual property.

If you have any questions about this, please come talk to the TA or the professor.

1. Attempting to gain an unfair advantage by recreating the original Yelp data and using predictors that are not part of the data set. You must only work with the data set you were provided with.

You are strongly encouraged to create your own predictors based on the data set you were given. Again, please come talk to the TA or the professor if you have any questions about this.

1. Sabotaging others’ work by removing, deleting, manipulating their code base or statistical analysis or providing misleading and/or false information about the data, relevant analysis, or any information related to the module.
2. You may not ask someone or some entity to do any part of the analysis on your behalf.

Committing said acts can result in disciplinary action, which includes, but is not limited to failure on the assignment/course, disciplinary probation, or suspension. Substantial or repeated cases of misconduct will be forwarded to the Office of Student Conduct & Community Standards for additional review. For more information, refer to <https://conduct.students.wisc.edu/academic-misconduct/student-resources-for-academic-integrity/>