Midsemester Project

CS 181-01/DA 210-01
Fall 2021
Instructor: Nathaniel Kell

Overview and Guidelines

For your midterm project, you will compile three to four data sets from the web, and then attempt use these data sets - along with the tools and techniques we've developed in the course - to answer a central question of your choosing. You may choose to work on the project with a partner or individually.

The topic, central question, and data sets you choose are entirely up to you. However, the methods/tools you use to do your analysis must satisfy the following requirements:

- **Requirement 1:** Your data sets should be tabular data, likely .csv files similar to the examples we've seen in class.
- **Requirement 2:** At some point when processing your data, you must use regular expressions in some meaningful/non-trivial way.
- **Requirement 3:** One of your data sets should be read-in to construct a DoL, and another data set must be read-in to construct a LoL. These data structures can then be used to construct pandas data frames. Note all remaining tables can be directly converted into data frames (if desired or necessary).
- **Requirement 4:** You must use pandas and data frames in some meaningful way.

As a toy example, your central question could be something like: *Which characters from the Avengers movie series are most popular?* Your data sets could then perhaps be: (i) a table showing the total revenue from Avengers merchandise, broken down by character, (ii) a table containing tweets from the night the movie *Avenger: Endgame* premiered, and (iii) a table with the number of followers each actor/actress from Avengers has on Twitter.

Then for the project, you could perhaps do the following: Using table (ii), build your own frequency table by parsing tweets with regular expressions, counting the tweets that involve each character in the movie (satisfying **Requirement 2**). You could then read-in tables (ii) and (iii) as a DoL and LoL (respectively), and then use these data structures to construct pandas data frames (satisfying **Requirement 3**). Then, using the pandas functionality we've seen in class, manipulate and process the data frames to output basic statistics and graphs that demonstrate the popularity of each character according to your data (satisfying **Requirement 4**).

Note: You are not required to use any sophisticated statistics or analysis in order to answer your central question. (Although if you have experience with such techniques, you

are welcome to do so). It is also fine if your findings are somewhat inclusive. For example, if the Twitter data seems to indicate Iron Man is the most popular character, but The Hulk is the most popular character according to merchandise revenue, this is a fine conclusion to make for your central question. (Just be sure to show/explain in detail how your results support this claim).

Time Line and Assessment Criteria

The time line and rubric for the project are as follows:

- **Progress Check-in Due Friday October 22 at 5:00pm [5 points]:** For your progress check-in, you should submit a one to two paragraph summary of your plans and progress so-far on the project. In the summary, you should (at a minimum) describe both the central question(s) you will explore and data sets you intend to use. Also, if you're working with a partner, you should indicate who you're working with. All group members should submit a progress check-in individually.

  Shortly after you submit your report, I will give you feedback about the complexity of your chosen project and whether I think you should to make it simpler or more complex.

- **Final Submission Due Wednesday November 3 at 11:59pm [60 points]:** Your final submission should be a .ipynb Notebook file that is formatted like a lab report, where your code, tables, graphs, etc. correspond to experiments/finding, and then such outputs should be interleaved with Markdown cells that collectively specify your central question(s), explain how your code works and is taking steps to answer this question, and discuss/interpret your results.

  Your project will be holistically evaluated based on the following criteria:

  1. **Correctness:** Is your code correct? Did you follow the guidelines in the project guidelines given above? (Requirements 1 through 4.)
  2. **Novelty:** Are the results and questions you explore interesting? Did you put in an effort to reveal something non-trivial about your data sets?
  3. **Style and Organization:** Is your file organized and well-formatted? Is your code annotated with docstrings and comments?
  4. **Writing Clarity:** Are your explanations (in Markdown) clear and well-written?

  Just to emphasize: For this assignment (as well as the final project), the above criteria imply you are will evaluated on both *how* you do your work and on *how well* you communicate your results. A project that just hacks together some monolithic or ill-structured code to "get an answer," or does a poor job expressing how the output of the code addresses your central question, will not receive a good grade.

Partner Submissions: Finally, if you work with a partner, *you will submit separate lab reports.* The code in the two reports can be identical, but the writing and exposition should be your own.