COMP 370 Homework 3 – Unix commands & MLP

Assigned Sept 13, 2024
Due Sept 20, 2024 @ 11:59 PM

The goal of this assignment is to use UNIX command line tools to do core data science work.

Task 1: Watch some My Little Pony episodes (totally optional)

In this and the next homework, we're going to be analyzing My Little Pony language. As we've discussed, it's always important to study your source material … particularly when it's very entertaining cartoons! So if you're able, watch a couple episodes!

Task 2: Explore My Little Pony Dataset Properties

We'll be using the dataset available here: https://www.kaggle.com/liury123/my-little-pony-transcript

For the purpose of this study, we'll use only clean_dialog.csv and assume that the dataset is perfect.

Using standard command line tools (e.g., head, more, grep) and csvtool, explore the clean_dialog.csv. Use the command line tools to answer the following questions:

- How big is the dataset?
- What's the structure of the data? (i.e., what are the field and what are values in them)
- How many episodes does it cover?
- During the exploration phase, find at least one aspect of the dataset that is unexpected – meaning that it seems like it could create issues for later analysis.

Task 3: Analyze speaker frequency

Use the grep tool to determine how often each MAIN (Twilight Sparkle, Rarity, Pinkie Pie, Rainbow Dash, and Fluttershy) pony speaks.

Now calculate the percent of lines that each pony has over the entire dataset (including all characters).

Submission Instructions

- explore.md: this file should contain answers to the questions/objectives posed above. Indicate the commands used to obtain the result.
- Line_percentages.csv: this file should contain the percent of all spoken lines comprised by each pony's speech acts.
    - It should have fields: pony_name, total_line_count, percent_all_lines