STAT 287 Final Project Check-In 1

Julia Zimmerman, Philip Nguyen

October 21, 2020

1

1.1

Major activities/milestones planned for this week (in the timeline that you had going into the start of the week).

1. Setting up Github, etc., to begin work – 2 hours (deliverable: being able to push and pull from Github)
2. Downloading or otherwise preparing a data set for use – 2 hours (deliverable: having a reasonable looking file)
3. Parsing the data to find structural information – this could be arbitrarily hard, so generous time should be allotted – 1 week (deliverable: data frames, vectors, or CSV)
4. *Check progress to see if project idea should be abandoned*

1.2

Major activities/milestones accomplished this week (note: this may or may not actually be what was stated in 1a – that's okay, as long as you're making progress and learning!).
Progress compared to original timeline:

1. Setting up Github, etc., to begin work – done (Github repo)
2. Downloading or otherwise preparing a data set for use – done; we downloaded a lot of the site, came up with a strategy for verifying that we have all the information from the site that we care about, changed that plan as explored the site, came up with a new strategy, and are now satisfied we have everything in the repo that we care about analyzing.
3. Parsing the data to find structural information – partly done; we have made dictionaries from all the different index pages on the site and lots of lists of which indices and tropes occur where; we're working on the individual trope pages still. We broke this item down into two parts: (1) make lists of tropes from the index pages; (2) extract what tropes are linked to within the articles. (1) ended up changing quite a bit as we explored the structure of the website and understood more about it - the overarching aim didn't change, but the details of how we were doing it did.
4. *Check progress to see if project idea should be abandoned* – done (no reason to think it should)

Progress as tracked on GitHub via closed issues:
We've been using GitHub as our main workspace, tracking our progress via issues and a shared repository. We also have a shared CoLab notebook for daily-ish updates, and have been using slack and zoom as needed. We have both been working on most of the items, but since the assignment instructions say to make it clear who has done what, we listed the main person who resolved the issues below.

1. Create repo (Julia)
2. Get contents of website (Julia)

3. Figure out what is in the repo - strategy for verifying if we have everything we care about from the website (Phil)
4. Start making data structures from the site data (Julia)

   Supporting tasks (a) and (b) of making data structures reflecting what lives where on the site (Phil)
5. Practical questions of formatting (Phil)
6. Bug-fix an encoding issue (Julia)
7. Bug-fix for idiosyncratic html structure (Julia)
8. Verify we have everything in the repo we care about (Phil)

   Supporting task about understanding website structure (Julia)

2

2.1

Open challenges and questions (including what – if anything – are the challenges that Ethan or I can help provide feedback or pointers on?)
We're not currently stuck on anything, but I think mid next-week - or by the next project check-in at the latest - we're likely to want to get feedback on how we resolve the different interpretations we could make of the structure.

2.2

Major changes to research plan (if any, based on what you've learned or accomplished thus far, and the unexpected challenges you've faced this week) None so far (fingers' crossed) - we changed our strategy for getting a "master list" of tropes several times as we learned more about the site, but that wasn't exactly unexpected.

3 Revised anticipated timeline

1. Setting up Github, etc., to begin work – 2 hours (deliverable: being able to push and pull from Github)
2. Downloading or otherwise preparing a data set for use – 2 hours (deliverable: having a reasonable looking file)
3. Parsing the data to find structural information – this could be arbitrarily hard, so generous time should be allotted – 1 week (deliverable: data frames, vectors, or CSV)
4. *Check progress to see if project idea should be abandoned*
5. Organizing and cleaning the information I've extracted – 1 week (deliverable: vectors, matrices, or other refined structure)
6. *Check progress to see if project idea should be abandoned; last chance*
7. Exploratory data analysis and visualization – 1 week (deliverable: interpretable images)
8. Writing up results – 3 days (deliverable: report in at least outline form)
9. Creating "birth certificate" summary of project [Eli20] – 1 day (deliverable: neat summary)
10. Audit for transparency, ethics, etc. – 2 days (deliverable: action items or approval)
11. Realize something actually makes no sense or wasn't doing what you thought – 1 week (deliverable: a fixed version of whatever needs fixing)
12. Refining visualizations – 3 days
13. Refining write-up – 3 days (deliverable: a polished report)
14. Making presentation – 2 days (deliverable: a presentation)

15. Donating money to (1) a social justice cause and (2) an environmental cause[1] – 1 hour (deliverable: receipts)

Total estimate: ≈ 6 weeks and 1 half day ≈ 6 weeks (I got the 6 weeks total from looking at the course schedule on Blackboard).

[1] We have the resources and opportunity to complete this project due to luck and privileges granted by society, some of which come at the cost of other people and the environment. Therefore some remuneration is within the scope of this project.