Distributed Computing Project 1 Logbook

Group Members:

- Josh
- Sarah
- Fehmi
- Raymond
- Paul

Main Communication:

- Google Space

Github: https://github.com/FNeffati/DCP1

Communication Expectations:

- Willing to check Gchat min 2x a day, more if actively working on a topic.
- Start daily standup thread, put what you've accomplished
- Group meetings as needed in library

Availability:
Fehmi Neffat: 8am - 6pm Monday-Friday.
Joshua Ingram: Flexible, but not available on Sundays.
Sarah Nash: Flexible, less available on weekends.
Paul Hwang: Flexible, but usually busy Thursday, Sunday, Monday nights
Raymond Blaha: Flexible, less so Sunday / Monday. Any other day available from 10:30AM - 9PM

High-level plan:

- Understand AO3 website structure. How to collect data on fanfictions.
- Star wars all media types fandom (if we want to compare fandoms, compare with star trek)
-

Group structure:

Logs

Monday, April 3, 2023

Group Meeting Agenda

- Progress Updates
    - Sarah
        - No updates
    - Fehmi
        - Has worked on the works_tags.csv file by having the id correspond to big list of tags. This id will correspond to metrics tags.
            - Does Paul's metric need to be updated?
    - Paul
        - Log transformation with hits, kudos, etc.
        - We will just use hits for our analysis
    - Raymond
        - More EDA on words.
            - Least common words, number of words, shortest and longest sentences
            -
    - Josh
        - Will create Github Readme and organize folders for final submission
        - Look at post frequencies by author
- Final analysis of author performance
- Create presentation
    - Outline
    - Slides
    - Practice
- Moving forward
    - Josh will send required data formats to group members. These need to be uploaded by Wednesday.
    - Josh will combine data into finalized format and perform an initial analysis on the data.
    - Group members will meet next Monday to finalize analysis, create presentation, and practice presentation. Create a readme file for group project and organize GitHub.
- Notes

Thursday, March 30, 2023

Group Meeting Agenda

- Progress Updates
    - Sarah

        - Sarah still working on the final dataset. Make sure the code that we're writing for the mini dataset will work with the full dataset (write this in a distributed way).
    - Fehmi
        - Worked on the scraper a bit. Will have some analysis done on popular tags.
    - Paul
        - Created a few plots using R. Analysis of metrics. Kudos and hits. How should we penalize the metric?
        - Distributions of metrics are interesting. May be power-laws or log-normals.
    - Raymond
        - Created two python files on the server. One was EDA for word lengths, most common words, number of words. Attempted some NLP sentiment analysis on the sample data. Classification using logistic regression model… NLP not performing well.
    - Josh
        - Post frequencies: number post/month and total number of posts.
- Moving Forward
    - Raymond does not want to present.
    - Meeting Monday to prepare presentation, perform analysis.
    - Sara has contributed enough with web scraper, so she does not need to do the analysis of the character pairings.
- Notes
    -

Monday, March 27, 2023

Project Presentation:
- April 5th
- 15-20 minute presentation
- Provide repository with readme

Meeting with Matt:
- Talk about processes at a general audience level in the presentation, but do not focus on this. Make the presentation about the insights from the data.
- Do not need to have every person present if it flows better that way. Have everyone available for Q&A.

Group Meeting Agenda:
- Progress Updates
    - (Sarah and Fehmi) Web scraper: collected data on almost all 60,000 star wars fandom stories, some did not work due to returning a 404 error. Keep working with the mini dataset for now until Sarah collects it all. Final result will be one folder with several .csv.

    - (Paul) Popularity metric: observe distribution of kudos, bookmarks, hits first.
- Moving Forward
    - Meeting setup for Thursday and Monday
    - Make progress by Thursday's meeting. On Thursday we will discuss issues, framework for presentation. Over the weekend tie up loose ends. Monday prepare and practice presentation
- Notes
    -

Monday, March 13, 2023

Group Meeting Agenda:

- Progress Updates:
    - Sarah and Fehmi - Web scraping
    - Paul - Response from Matt about metrics
    - Server folder
    - Questions
- Moving forward:
    - Assignment due this Thursday - status update, code, etc.
    - Individual project goals for the week
    - Work over spring break?
    - Open discussion
    - Plan tentative meeting
- Notes
    - Sarah and Fehmi
        - Sarah will finish web scraping code edits Fehmi made.
        - Will run the web scraper for a few hours to get a sample dataset of about 100-200 stories worth of data and will put on server. This data will be used to develop analysis code.
        - Will run the web scraper over all of break. Will run analysis code on full dataset week after break
    - Paul
        - Focus mostly on hits for metrics, but consider kudos as well. Don’t need to use bookmarks
        - Going to play around with a metric to see what works best. May use bookmarks.
        -

Wednesday, March 8, 2023

[Sarah]

- Worked on scraping data, preliminary results at https://docs.google.com/spreadsheets/d/13suBaYcZW7p4BL_ozwf-dlS9pS3YwV6H9n9sUfuQDGg/edit
- Gathering metrics like these will be fast, larger data points (like full text of the work, tags, characters) will take a little longer to do
- Want to split “tags” attribute into smaller categories (relationships, characters, etc)
- Look into grouping csv files in a directory, sparkifying the scraping.
- Asked Matt to give us a common directory on the server for file sharing.

Monday, March 6, 2023

Meeting Notes

-

Wednesday, March 1, 2023

Notes

- Use hits, kudos, bookmarks as indicators for successful fics
- Does length affect popularity?
- Score/metric for authors’ popularity
- Exploration on:
    - Popular characters
    - Popular pairings
    - Popular tags
    - Date of posting

Links

Star wars fanfic
https://archiveofourown.org/tags/Star%20Wars%20-%20All%20Media%20Types/works

Mining fanfics on ao3
https://medium.com/nerd-for-tech/mining-fanfics-on-ao3-part-1-data-collection-eac8b5d7a7fa