Anime Text Analytics

University of Illinois at Urbana-Champaign | *CS 410 (Fall 2020)* | *Course Project*

Mihir Yerande, Karan Bokil

The following is a progress report by *Team Nani (何 !?)* summarizing the current state of the *Anime Text Analytics* project. The report includes...

1. Completed Tasks
2. Pending Tasks
3. Challenges

Completed Tasks

Web Scraping

- Built web-scraper using *scrapy* library in Python
- Tweaked scraper for politeness
- Ran scraper to obtain data for ~4.7K animé TV shows from `myanimelist.net`
    - URL
    - Title
    - Description (for LDA)
    - Genres (as given by `myanimelist`)
- Stored scraped JSON data in *scraped.jl*

Latent Dirichlet Allocation (LDA)

- Looked into *gensim* to run LDA in Python
- Cleaned text before running LDA, and wrote to *lda_input.jl*
- Initially ran LDA on subsets of scraped data to try it out
- Confirmed that interesting topics (“genres”) can be discovered

Website & Database

- Whiteboarded initial concept for website
- Flask environment setup:
    - Basic routes for index and new entries (GET/POST)
    - Boilerplate for models, controllers, and views
- Registered Azure Cosmos DB
- Added Materialize CSS framework for UI

Pending Tasks

Latent Dirichlet Allocation (LDA)

- Further prune/clean text to remove non-discriminative words
- Determine optimal number of topics for LDA (using coherence/perplexity?)
- Configure LDA params (starting state, etc) to achieve most interesting outcome
- Look into other potentially interesting features:
    - e.g. Similarity between shows, using generated topics

Website & Database

- Migrate scraped data to Azure Cosmos DB
- Connect Azure Cosmos to Azure Cognitive Search
- Create UI for various types of pages:
    - List of all shows
    - List of all genres
    - Single show
    - Single genre
- Deploy site to Azure VM
- Consider any additional ideas for the website, if time permits

Challenges

Latent Dirichlet Allocation (LDA)

- Need to examine text input and rerun LDA to find non-discriminative words
- Need to research established methods of finding optimal number of topics
- Need to re-run LDA using different random states to achieve interesting outcome
- Need to look further into gensim for other helpful LDA features

Website & Database

- Need to research Azure Cosmos DB integration with Azure Cognitive Search
- Need to determine whether to make pages reactive or generate hard links
- Need to look into graph visualization libraries for genre-based show similarity