Anime Text Analytics

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

Mihir Yerande, Karan Bokil

Team Members

Member #	NetID	First Name	Middle Initial	Last Name	Captain?
1	*yerande2*	Mihir	-	Yerande	☑
2	*karanb2*	Karan	S	Bokil	

Project Topic Description

For our project, we have chosen to pursue a free topic: *Anime Text Analytics*. The project aims to produce text-based analytics providing insight related to Japanese animé. Our plan is to scrape text data, in the form of descriptions and reviews of anime shows, from the website `myanimelist.net`. This is a popular website with a lot of user-based anime data. Using this text data, we will perform topic mining and analysis in order to discover a set of topics (genres) similar to the niche genres which have been produced by Netflix. Additionally, these analytics may further be used to produce an anime recommendation system and UI for browsing based on the discovered genres.

This would be an interesting direction for a project because there are many anime subgenres which users would be interested to browse within. For example, there are 5 broad genres of anime (*Shōnen*, *Shōjo*, *Seinen*, *Josei*, *Kodomo*) which correspond to different viewer demographics based on age and gender. In addition, there are the usual content-related genres that you would expect to see elsewhere (e.g. *Romance* or *Action*). However, there are also some well-known subgenres, specific to anime, such as *Mecha* or *Vampire*. Our project would aim to discover such subgenres and perhaps go further by discovering more granular subgenres or intersections thereof. The results of such an analysis could further provide a means of browsing similar anime based on the calculated genre profiles.

Our approach would involve scraping the text data using *scrapy* and then cleaning and structuring the data with *pandas*. We would then explore existing frameworks to perform topic modeling on the scraped data (perhaps using Latent Dirichlet Allocation). The resultant topics (genres) could then be explored via a website built upon *Flask* and *Materialize*. The expected outcome would be a website allowing for exploration of different animes in terms of the discovered genres. Furthermore, we may use the results of the discovery to produce add-ons such as a simple recommendation system, a graph visualization for browsing, or a show description generator.

To evaluate the topics generated by our system, we will manually check the discovered genres to intuitively understand which genres have been discovered. Then we will examine some examples of shows to determine whether their respective genre coverage makes sense. We might try to incorporate simple measures such as precision or the F-measure by simply sampling the genre outputs and stating whether or not the genre coverage for various animes seems correct. It may also be possible to cross-reference the existing recommendations on myanimelist to evaluate a recommendation system, since the site aggregates a lot of user data in the form of watch suggestions.

We would like some advice as to how we should evaluate the output of this project. Evaluation is difficult here since we are aiming to discover genres that haven't necessarily been described yet, which means there is no simple baseline for comparison.

Programming Languages

- Python
- JavaScript
- HTML
- CSS

Workload

Here is a rough breakdown of the expected time expenditure for the project:

● Data Gathering:	
○ Scraping	5 hrs
○ Cleansing/Structuring	2 hrs
● Topic Modelling:	
○ Background Research	5 hrs
○ Implementation	5 hrs
○ Configuration/Exploration	5 hrs
○ Output Evaluation	4 hrs
● Website:	
○ Backend Dev (Connecting DB, Routing, etc)	6 hrs
○ Frontend Dev (Topic UI, Search, Graph)	6 hrs
○ DevOps (Provisioning, Deploying live site)	5 hrs
● Further ideas	[Open-ended]
Total	43+ hrs