**Group Members**

|  |  |  |  |
| --- | --- | --- | --- |
| S Number | Name | Discord/Messenger | Signature |
| S5097965 | Timothy Chew | C:\Users\aytimothy\Documents\facebook-messenger.png aytimothy  C:\Users\aytimothy\Documents\discord.png aytimothy#0919 |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Fandoms Connect

**3804ICT Data Mining**

A data-mining investigation into how popular culture fandoms are interconnected, spread and related.

**Project Proposal & Investigation Report**

S5097965 DRAFT 3804ICT T1/2019

# Preface

Thanks to the internet and the rise of internet message boards to the wider public such as Reddit, 4chan and not Newsgroups, the past 5 years have never been a better time for fans of pop-culture and interest groups to connect with one another. However, despite the internet being interconnected and everything out in the open, there is still the tendency for groups of people to clump up or join the same set of groups (which is still clumping up).

The aim of the investigation is to see what percentage of users belong to which fandoms, and how different fandoms overlap or avoid each other. For example: Do *Digimon* fans also interact with those who also like *Pokémon*, and vice versa[1]?

In order to gather the data, first, uses are aggregated and posts are scraped of Reddit using the Reddit API[2]. Using the raw post data of [parent, user, score, subreddit, content], we can then clean it into a machine-usable state by using text analysis to build our database of data points. Before processing, we can already see content by their authors, communities posted to and upvotes. From there, we can look at the comments to see what is being mentioned where. Keywords, for example such as Pokémon names can be used to signify that a person knows about, well, Pokémon.

In order to work out how likely a person from a fandom is part or knows about another, we can emply methods such as k-Nearest Neighbours, Decision Trees and Support Vector Machines to class users into their fandoms, and, to work out (theatrically) where new users may sit within the multi-dimensions of different fandoms.

# Investigation

First, we need to gather data. Using [praw](https://praw.readthedocs.io/en/latest/), data is collected from Reddit. You can find the code at [aytimothy/3804ict-fandoms-connect](https://github.com/aytimothy/3804ict-fandoms-connect) (may be invisible due to private repository settings).

Next, we can work out how fandoms are connected using the content that have been taught in Week 5. It is possible to find correlation between fandoms by looking at users' posting behaviour. This correlation can be gained using techniques such as k-Nearest Neighbours, or as taught: the Apriori Principle or FP-Growth tables/trees.

# Preliminary Data Analysis

**Note: The data is currently being scraped, and this information does not reflect the whole dataset.  
The information is accurate as of the 6th of August, 9:01 PM. Since this is a continuous mining operation, some numbers may change while this is written.**

* 635,332 comments and 6,989 submissions have been processed from Reddit, after approximately 60 hours of runtime.
* We have processed the entire post/comment history of 82 users.
* There are 340,817 unique users for comments and 3,312 for submissions in the database.
* There are 1,325 unique communities between all comments, and 1,473 between all submissions.
* There are 394 submissions and 31,985 comments created by deleted users.
* There were 3,397 comments and 311 submissions removed by moderators.
* There were 9,826 comments and 302 submissions deleted by their authors.

Submissions and comments are distributed as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Property | Min | Q1 | Median | Q3 | Max | Mean |
| Submission Score | 0 | 3 | 35 | 466 - 467 | 153793 | 3139.2162 |
| Comment Score | -532 | 1 | 4 | 17 | 58234 | 116.9329 |

Minimums are maximums are skewed while the medians still trend in positive numbers as only the outliers generate high/low scores. Most comments/submissions get few or little upvotes.

Finally, a few more interesting information about the content that was scraped:

* The most up-voted submission is a [picture of a cat](https://www.reddit.com/r/aww/comments/ckbolc/this_is_tiger_he_just_turned_31_we_are_told_he_is/) from [/r/aww](https://reddit.com/r/aww) with a score of 153,793.
* A [Rick Roll](https://youtu.be/dQw4w9WgXcQ); Rick Astley - Never Gonna Give You Up was the fifth [highest scoring post](https://www.reddit.com/r/videos/comments/5gafop/rick_astley_never_gonna_give_you_up_sped_up_every/)collected.
* [/u/SovietRussiaBot](https://reddit.com/u/SovietRussiaBot) holds the 1st, 3rd and 5th most downvoted comment. It's a [bot](https://github.com/dneu/SovietRussiaBot) that replies to anything it considers a [Russian Reversal](https://en.wikipedia.org/wiki/Russian_reversal) joke, and also uses praw.
* The [top comment](https://www.reddit.com/r/AskReddit/comments/bu1s5i/what_fact_is_common_knowledge_to_people_who_work/ep6aqhz/) is about elevators when they have a catastrophic failure.

[1] Original Research: Older audiences of both franchises (the 1999 to 2003 demographic who were ages 6 to 18 at the time, or parents) are aware of each other due to the virtual pet and early toy gatcha market. This later spread down the line towards the current younger generation (these people are now in their late teens to adulthood as of 2019).

[2] <https://www.reddit.com/wiki/api>