Comic Characters Data Mining

In-depth Analysis of DC & Marvel Characters

Huy Tran
u1228479
u1228479@umail.utah.edu

John Duffy
u0783287
u0783287@utah.umail.edu

Ethan Olpin
u1018382
u1018382@umail.utah.edu

Jeff Gay
u1098246
u1098246@umail.utah.edu

1. DATASETS

For our proposed project we plan to use the FiveThirtyEight Comic Characters Dataset. The data originates from the Marvel and DC Wikias and specifies the attributes of a wide array of comic book characters. These attributes include things such as their name, identity, moral alignment, gender, and aspects of their appearance (eye and hair color). The dataset also includes page URLs for each character's Wikia entry which we can utilize to expand our considered dataset. We believe that when this data set is subjected to common data mining techniques, it will reveal insights into the trends in the appearance, identity, and writing of comic book characters.

2. RESOURCES

As we are still in the initializing process and finding more data to work with, these are the potential resources that could be of great use. More datasets as well as resources will likely be included as we work through the project.

1. Datasets

https://www.kaggle.com/danoozy44/comic-characters?select=dc-wikia-data.csv

https://www.kaggle.com/dannielr/marvel-superheroes

2. DC Comics Database

https://dc.fandom.com/wiki/DC_Comics_Database

3. Marvel Database

https://marvel.fandom.com/wiki/Marvel_Database

3. DATA PROCESSING

We plan to keep the dataset in .csv files. The data will be in table type, each row is a unique character with corresponding columns being the attributes of the character. For example, Batman (Bruce Wayne) is an item in our dataset, with attributes like Secret/Public identity, Good/Bad, Sex, Number of Appearances.

For reading, processing, and analyzing data we anticipate using Python with standard data science libraries such as Pandas and Numpy.

As we explore more datasets that are useful for the project, there will be more data cleaning and processing in order to combine the datasets and keep them in portable sizes.

4. PROJECT EXPECTATIONS

We expect to gain insights into how certain trends will affect the creations of new characters.

The trends that we will examine are historic events and consumer demographics. We hope to derive how a character's popularity will bring that character into other media such as television and movies.

Some example data mining techniques includes:

1. Perform regression on the number of appearances of each character to predict their next appearance on bigger screens (TV, Theater, etc.)
2. Learning different attributes of a comic character based on the data to build a model on predicting the likely attributes or characteristics of future character creations.
3. Apply clustering algorithms on characters to determine if a comic fan enjoys a particular character, he/she might as well be interested in another character in the group.
4. Apply graph analysis to a set of characters in the same network (universes, cities, regions, etc.) to reveal how a character can be influential to the story and to other characters.

5. PROJECT EVALUATION

As we explore the data using different data mining techniques, the goal of this project is to successfully reveal hidden patterns within these comic characters. We can train and test the models built with these insights for high accuracy.

Some evaluation plans based on the outcome of the data exploration:

1. Cross-validation testing on our models to see if we can learn from a set of characters from years ago to predict how popular these characters would become in the early 21th century.
2. Evaluate a model that takes a few characters of favorite from a person and generate comic characters that they might like. Survey with a few comic fans and see with just a few characters input, would we be able to tell what they like.