Super Heroes and Natural Language Processing (NLP)

Daniel Maiorano, Justin Moczadlo, Grace Seaton, Aaron St. John, Reilly Williams

I. INTRODUCTION

Have you ever thought to yourself who might be the strongest super hero? Perhaps you don't know many super heroes, but you wonder what might be the strongest super power to posses. Well the objective of this project is to dive deep into a database of super hero information and apply natural language processing.

The main objective of this project is to use the text columns provided in the dataset to classify the creator of the super hero. If time permits we could also look at things such as the "most powerful" super hero and more.

The motivation for this project came from our nerdy tendencies and interest in all things related to super heroes. We also didn't know an entire database of super hero data existed. This will be a great opportunity to explore that kind of data and see super heroes we haven't even heard of.

II. DATA DESCRIPTION

The data being used is offered on Kaggle and consists of information on history and powers of over 1400 super heroes from Marvel, DC, and more.

The dataset has exactly 1450 records with a total of 81 different columns. However, 50 of the columns are dummy variables for various powers and will be of no use to our project. Thus, after removing the 50 dummy coded columns we're left with 'name', 'real_name', 'full_name', 'overall_score', 'history_text', 'powers_text', 'intelligence_score', 'strength_score', 'speed_score', 'durability_score', 'power_score', 'combat_score', 'superpowers', 'alter_egos', 'aliases', 'place_of_birth', 'first_appearance', 'creator', 'alignment', 'occupation', 'base', 'teams', 'relatives', 'gender', 'type_race', 'height', 'weight', 'eye_color', 'hair_color', 'skin_color', and 'img'.'

Our primary columns of interest are 'full_name', 'history_text', and 'powers_text'. The 'full_name' column simply contains the full name of the super hero. The 'history_text' column contains a string describing the background and history of the super hero. Finally, the 'powers_text' column contains a string describing the powers of the super hero.

III. TEAM MEMBER RESPONSIBILITIES

There are 5 team members in our group and we are going to split up the various responsibilities amongst the 5 members. There are multiple steps that go into text analysis.

- Daniel Maiorano: Daniel will be responsible for doing the first round of cleaning on the data. This will include finding useful columns, cleaning NA's from the data, as well as creating the final dataset that we will use in our text analysis.
- Justin Moczadlo: Justin will be responsible for some preliminary text processing. This will include making the words lower case, removing stopwords, as well as any other steps that are needed.
- Grace Seaton: Grace will come up with some rudimentary text visualizations. These are subject to change but may include a word cloud, some text rank bar graphs, as well as any other visualizations that she deems useful.
- Aaron St. John: On top of being the owner of the GitHub repository, Aaron will also be responsible for testing various models. Some of these may include rather simple models like logistic regression, or perhaps more complex models like artificial neural networks. This might be too much for a single person so other group members are likely going to help him.
- Reilly Williams: Reilly will be responsible for the final leg of the project. This will include our model evaluations and comparisons, as well as compiling our findings. Reilly will be the one to create the final PowerPoint presentation for the class.

IV. TIMELINE OF MILESTONES

Here is a general timeline for when each of the team member responsibilities:

- Week 1: clean the data
- Week 2: text cleaning and stop word removal; initial text analysis
- Week 3: create word visualizations
- Week 4-5: testing different models on the data
- Week 6: model evaluation, compilation of findings, and building presentation

V. EXPECTED OUTCOME

The expected outcome of the project is that it predicts a creator based off of new data it has not yet seen. The project will be considered a success if it predicts the new creator with reasonable accuracy based off of the data we feed it. We are intentionally being vague with the term "reasonable accuracy" because we do not know if the model will struggle or not. Until we dive into the data we do not know if an accuracy of 90 percent is feasible, or if an accuracy of 30 percent is feasible.