Project 3 Report

CSE590-59 Python and Data Analytics

Jonathan Loyd

This project requires three texts representing different media or writing formats to be downloaded and analyzed. The types of text chosen for this project were a novel, a movie script, and a children’s book. The texts are respectively *Fear and Loathing in Las Vegas* by Hunter S. Thompson, *Ex Machina,* which is written and directed by Alex Garland, and *The Giving Tree* by Shel Silverstein. All three of the works are widely different in content. *Fear and Loathing in Las Vegas*, which can be called Fear and Loathing as a short name, is a novel that is known for playing a huge role in the genre of gonzo journalism and has themes of drug use and mania. *Ex Machina* is a movie that is generally easier to follow than Fear and Loathing, but still presents some interesting commentary on artificial intelligence and the dangers it could present at its highest levels. Finally, *The Giving Tree* is a children’s book that describes a tree sacrificing everything to a boy throughout his life and is the shortest and easiest to follow due to its nature as a children’s book. For information regarding the URL that the text was obtained from and its method of retrieval, please refer to the URLS\_for\_files document attached.

The next portion of the project was to compute the word counts of each of the three documents and remove English stop words. After this, the 25 most frequent words and their word counts were graphed on a bar graph for each document, and the 100 most frequent words were displayed on a word cloud. There is one noticeable similarity, where said is in the top three most frequent for *The Giving Tree* and for Fear and Loathing, where both texts describe a lot of dialogue. *Ex Machina* has a lot of dialogue, but it is said outright instead of being described as it is a script for a movie rather than a novel or children’s book, so it does not have said as one of its most frequent words. Otherwise, the graphs are different. Fear and Loathing has “said”, “n’t”, “attorney”, and “like” as frequent words, where he travels with his attorney, and there is a lot of dialogue where “said” and “like” could come up frequently. *Ex Machina* has a lot of dialogue between the characters and often about the other characters, which is why the three main characters (two if you don’t count Ava the AI), “caleb”, “nathan", and “ava” are the most frequently used words. *The Giving Tree* is a shorter text than the other two, so its word frequencies are generally lower, but it constantly mentions simple words regarding the main characters, the “boy”, and the “tree”, and uses said quite a few times for dialogue. Most of the common words are quite simple between the three documents, but it is especially expected out of *The Giving Tree* because its main audience is children, and slightly less so from *Ex Machina* because it was a movie meant to have a large audience. This data can be analyzed more directly from the bar graphs, but all of it can be visualized from the word clouds.

After analysis of word frequencies, the Textatistic library was used to compute the average of the Flesch-Kincaid, Gunning Fog, SMOG, and Dale-Chall scores for each document. Fear and Loathing had the highest scores for Flesch-Kincaid, Gunning Fog, and SMOG and the second highest score for Dale-Chall. *Ex Machina* had the second highest scores for Flesch-Kincaid, Gunning Fog, and SMOG and the highest score for Dale-Chall. *The Giving Tree* had the lowest values of the three documents for all four of the measures of readability. The average readability scores between these four measures of readability were 5.188 for Fear and Loathing, 4.756 for *Ex Machina*, and 2.176 for *The Giving Tree*. The Giving Tree is noticeably the lowest score of the three documents for the average and all four measures of readability used, which can reasonably be expected because the book is meant to be for children. It is often read to children at early ages or read by children who are learning how to read, which would explain its lower score. Fear and Loathing and *Ex Machina* have similar scores, which could be due to the simplicity of their dialogue throughout the media, while still being marginally more complex than a children’s book. *Ex Machina*’s movie script may be artificially higher or lower than expected because it has not been preprocessed to remove page numbers or other oddities within the text file, but the audience for *Ex Machina* and Fear and Loathing are both generally broad, where *Ex Machina* might be slightly larger, and thus more simple, because a goal of *Ex Machina* was to make money, whereas the goal of Fear and Loathing was more so to tell a story. These scores should likely be lower, which would indicate more difference, because there is such a difference in the subject matter, the word used, the intended audience, and the format of the media.

The fourth portion of the project was to use the spaCy library to compute the pairwise similarity between the three documents. The similarity between Fear and Loathing and *Ex Machina* was .9853. The similarity between Fear and Loathing and *The Giving Tree* was .9869. The similarity between *Ex Machina* and *The Giving Tree* was .9853. These scores are extremely similar, where the scores of *Ex Machina* with the other two texts had the same similarity score. These scores are much higher than I would expect if I wasn’t familiar with these scores being high on extremely dissimilar documents. Having a similarity of .9853 is incredibly close to 1.0, which would indicate that they are almost identical to someone without experience with this library, however having read all these texts and knowing how dissimilar they are this still seems incredibly high. The main similarities between these texts are their ease to read, and their dialogue-driven stories, but the subject matter and media are completely different, which would lead me to believe these scores should be much lower.

The final portion of the project utilizes spaCy to find the named entities in each document, and to create two bar plots for each document, where the first plots the 20 most common named entities by name, and the second plots the counts of every named entity type such as PERSON, ORG, or TIME. The Giving Tree is noticeably simpler for both plots, where it only has one entity name of the same type that occurs twice. It is marginally shorter than the other two documents, so this is understandable. Fear and Loathing has “Las Vegas” and “Vegas” as its number 1 and 4 more frequent entity respectively, which makes sense because that is the main setting of the novel. It also has “two” and “one” as the number 2 and 3 most frequent entities, which are used in several different contexts throughout the novel. *Ex Machina* on the other hand has the three main characters “NATHAN”, “AVA”, and “CALEB” as the three most frequent entities in order, with “HOUSE” as the fourth, marginally further down. The house is the main setting like Las Vegas is for Fear and Loathing. The names of the characters in capital letters appear frequently for stage directions and dialogue, which is not the case for either Fear and Loathing or *The Giving Tree*, which explains the difference there. As for the entity types, both Fear and Loathing and Ex Machina reference the ORG (organizations) type entity the most, with CARDINAL (numbers mostly) as the 2nd and 3rd most frequent entity type respectively. For the organizations, Fear and Loathing has mentions of landmarks and brands that may lead to the high ORG label count, while “NATHAN”, “CALEB”, and “AVA” are often artificially labeled as ORG instead of as a person, so there is some amount of inaccuracy there. The CARDINAL frequency in Fear and Loathing might be a product of the amount of description that Hunter S. Thompson uses counting people, time, and the number of drugs on the main characters, while CARDINAL being high for *Ex Machina* is a product of the page numbers on the script not being removed through preprocessing. The PERSON label is also high between Fear and Loathing and Ex Machina, since the two pieces of media require a lot of dialogue and interaction between characters.