Understanding Science Fiction using Machine Learning and Natural Language Processing

Thomas Wight (thomas.wight@enmu.edu)

Faculty Advisor: Eduardo Ceh-Varela (eduardo.ceh@enmu.edu)

Department of Mathematical Science; Eastern New Mexico University, Portales, NM

Introduction

Although most people have a general idea of what science fiction (sci-fi) is, a precise definition of the genre isn't easy to give.

In this project, we apply machine learning (ML) and natural language processing (NLP) techniques to analyze text from sci-fi novels and short stories, by creating a topic model of the Project Gutenberg sci-fi corpus.

Background

Science Fiction

Science fiction is a kind of fantasy literature involving fictional scientific discoveries or advancements. But there is some debate around what makes this genre unique, and how it differs from other imaginative and fantastic literature [1, 2].

Natural Language Processing

NLP is a branch of artificial intelligence which employs computational techniques for the purpose of learning, understanding, and producing human language content [5].

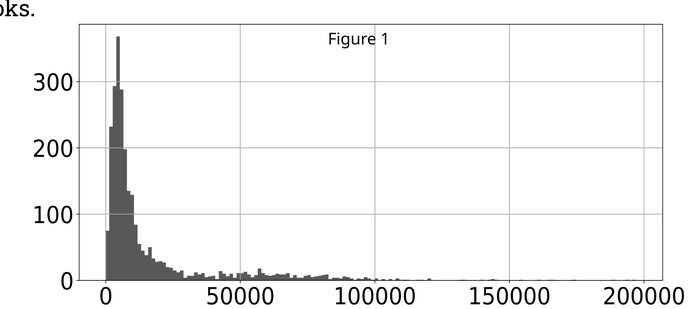
Topic Modeling

Topic modeling is a form of NLP for extracting topics from a corpus of text. Topics are groups of documents that share some common semantic elements (themes). Topic modeling allows us to analyze large volumes of text that we don't have the resources to analyze manually [6]. These topics may be used to better understand the composition of the sci-fi genre.

Dataset

The Gutenberg Library [htttps://www.gutenberg.com] is a collection of free ebooks. We use the 2700+ books that have the subject tag 'Science fiction'. Unlike a pre-formatted corpus, these books are inconsistently formatted, contain some metadata, and are various lengths. Figure 1 shows the distribution of books based on their length. Most of the books are less than 20,000 words, but there is a long tail of especially large books.

Figure 1



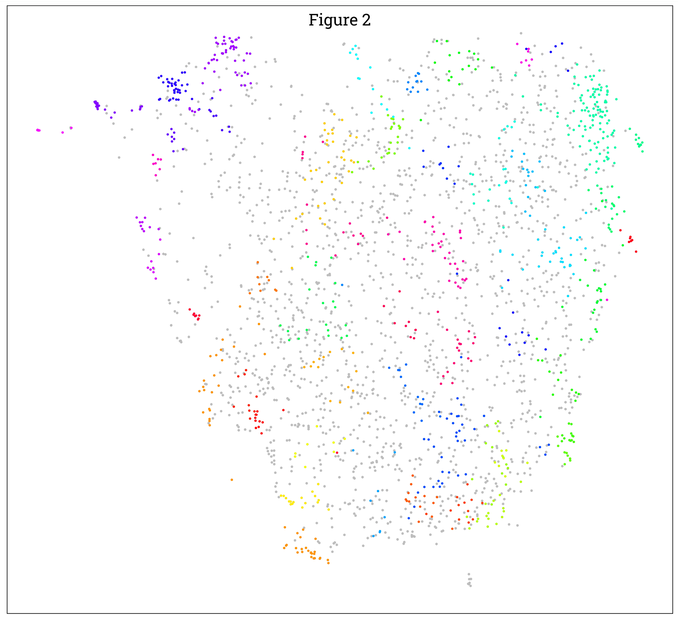
Methodology

1. We use Project Gutenberg to acquire the **corpus**. We don't remove stop words, because the generating embeddings with the transformer model architecture will use stop words for context.
2. We use a pre-trained sentence embedding model [7] to generate **embeddings** for each document. These embeddings are a vector containing 384 features per document. More than just a representation of term frequency, these embeddings should encode semantic information about each document.
3. We use UMAP [8] to **reduce** the embeddings vector's dimensionality to seven features per document. Note that this technique encodes information in the combinations of features, so less information is lost in this process.
4. We use HDBSCAN [9] to **cluster** the embeddings with similar features. These clusters represent distinct groups of documents that should have sentences with similar meanings. Figure 2 represents the clusters as different colors in two feature dimensions.
5. We use c-TF-IDF [3] to approximate the most important words for each **topic**. We limit the words appearing in the results to those with a document frequency (across the entire corpus) of at least .55, to filter out names that appear in novels related to a topic.

$$c-TF-IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j}$$

Results

Figure 2



We represent the c-TF-IDF scores as word clouds for their given topic.

Conclusions

ML and NLP techniques can help researchers extract hidden information from large corpus. Thanks to the pretrained transformer models, applying these techniques doesn't require computationally expensive training.

In our research, the topics found give us a better understanding of the composition of the Science Fiction genre.

References

[1] Parrinder, P. (2013). Science fiction. Routledge.

[2] Roberts, Adam. The history of science fiction. London: Palgrave Macmillan, 2016.

[3] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. ISO 690

[4] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.

[5] NLP [Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266. Chicago]

[6] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

[7] Reimers, N., & Gurevych, I. (2019, August 27). Sentence-bert: Sentence embeddings using Siamese Bert-Networks. arXiv.org. Retrieved March 30, 2022, from https://arxiv.org/abs/1908.10084

[8] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

[9] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2(11), 205.

This research was sponsored by the NM Alliance for Minority Participation (NMAMP) - Undergraduate Research Scholarship (URS)