2020

Using Natural Language Processing to Categorize Fictional Literature in an Unsupervised Manner

Dalton J. Crutchfield
University of Denver

Using Natural Language Processing to Categorize Fictional Literature in an Unsupervised Manner

Abstract

When following a plot in a story, categorization is something that humans do without even thinking; whether this is simple classification like “This is science fiction” or more complex trope recognition like recognizing a Chekhov's gun or a rags to riches storyline, humans group stories with other similar stories. Research has been done to categorize basic plots and acknowledge common story tropes on the literary side, however, there is not a formula or set way to determine these plots in a story line automatically. This paper explores multiple natural language processing techniques in an attempt to automatically compare and cluster a fictional story into categories in an unsupervised manner. The aim is to mimic how a human may look deeper into a plot, find similar concepts like certain words being used, the types of words being used, for example an adventure book may have more verbs, as well as the sentiment of the sentences in order to group books into similar clusters.

Document Type

Thesis

Degree Name

M. S.

Department

Computer Science

First Advisor

Chris GauthierDickey

Second Advisor

Adam Rovner

Third Advisor

Scott Leutenegger

Keywords

Natural language processing, Plot lines, Fiction, Computer science

Subject Categories

Computer Sciences | Other Computer Sciences | Theory and Algorithms

Using Natural Language Processing to Categorize Fictional Literature in an Unsupervised Manner

---

A Thesis

Presented to

the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

Dalton J. Crutchfield

June 2020

Advisor: Chris GauthierDickey

Author: Dalton J Crutchfield
Title: Using Natural Language Processing to Categorize Fictional Literature in an Unsupervised Manner
Advisor: Chris GauthierDickey
Degree Date: June 2020

Abstract

When following a plot in a story, categorization is something that humans do without even thinking; whether this is simple classification like “This is science fiction” or more complex trope recognition like recognizing a Chekhov's gun or a rags to riches storyline, humans group stories with other similar stories. Research has been done to categorize basic plots and acknowledge common story tropes on the literary side, however, there is not a formula or set way to determine these plots in a story line automatically. This paper explores multiple natural language processing techniques in an attempt to automatically compare and cluster a fictional story into categories in an unsupervised manner. The aim is to mimic how a human may look deeper into a plot, find similar concepts like certain words being used, the types of words being used, for example an adventure book may have more verbs, as well as the sentiment of the sentences in order to group books into similar clusters.

Acknowledgements

Without the support of others, completing this degree would have been impossible, and for this I am forever thankful to those who supported me. I would like to thank Dr. Chris GauthierDickey for all of his support, guidance, and advice while working on this thesis; without his commitment to me, despite his always busy schedule, I would have never gotten to where I am today. I would also like to thank my loving Fiancée Lindsay Todd: Without her constant support in both the good and the bad days, I would not be where I am today in either academia or life. Lastly, I would like to thank my family, particularly my parents, for all of the support they have given me from my childhood, through both my Undergraduate program and my Graduate program.

Table of Contents

List of Figures

Text

Introduction:

Storytelling has been a cornerstone of society since the early days of the human race. Cave drawings detailing animals and themes of survival date back to the Chauvet cave in France from 30,000 years (Mendoza, 2015). Storytelling continued to evolve mainly as oral tradition until around 9,000 years ago when the earliest known written stories were transcribed. The first written stories were manually transcribed, whether on paper, stone or clay. Writing began as drawings, but over time changed into prose (Mendoza, 2015). In ancient Greece, where the earliest inscriptions date from 770 to 750 B.C., Scholars suggest that "The Iliad" by Homer is the oldest surviving work in the Greek language that originated from oral tradition (Mendoza, 2015).

Scholars have pinpointed the invention of fiction, defined as a mode of writing in which both author and reader are aware, and know that the other is aware, that the events described cannot be known to have happened, to 12th century England (Ashe, 2018). During this time, there was a general increase in economic prosperity across Europe among the aristocratic elites. This, along with the Roman Catholic Church pushing a new focus on interiority and selfhood, among other things, led to the creation of the first known fictional

works (Ashe, 2018). Around 1155, a Norman clerk called Wace presented a long poem to Eleanor of Aquitaine. This poem was a French translation of Geoffrey of Monmouth's *History of the Kings of Britain* (c. 1136), a long work of Latin prose that purports to recount the history of Britain. In these poems, the most prevalent hero was King Arthur, the conqueror of most of Europe, before treachery forced him to return to a civil war which led to the ultimate downfall of his people (Ashe, 2018). When Wace translated Geoffrey's Latin into French verse, he took the opportunity to elaborate and embellish on the descriptions of Arthur's court. In doing this, he transformed the historical poems into fictional narratives, meaning the characters were known to both the author and the reader as being something that could not have existed in reality (Ashe, 2018).

From here, fiction continued to build and evolve until becoming what we know of today: by the time of the Elizabethan Age, religious inspiration was becoming distinct from scientific fact, truth was something to be proven by observation and experiment, and the aesthetic event was a self-conscious production (Doctorow, 2006). This is where William Shakespeare comes in and changes the way fiction is told from the writing of his plays, which portray fictional storyworlds. Shakespeare wrote about timeless themes such as life and death, youth versus age, love and hate, fate and free will, to name but a few. Not only did Shakespeare change structures of fictional writing, but he also invented around 1700 words which we still use in everyday English today. He often changed nouns into verbs, verbs into adjectives, connecting words together and coming up with wholly original ones

too (Celtic English Academy, 2017). From here, fiction continues to evolve: Mark Twain, a famous author in the 19th Century, said that he never wrote a book that didn't write itself (Doctrow, 2006). Henry James, an American-British author regarded as a key transitional figure between literary realism and literary modernism, in his essay "The Art of Fiction," describes this empowerment as "an immense sensibility … that takes to itself the faintest hints of life … and converts the very pulses of the air into revelations." What the novelist is finally able to do, James says, is "to guess the unseen from the seen." (Doctorow, 2006). Modern fiction continues to build off of the foundation set by key authors throughout history. In fact, according to Cristopher Booker, an English journalist and author, there are only seven basic plots that all stories are built from at their core (Booker, 2019).

Booker, who early on in his book *The Seven Basic Plots: Why we Tell Stories* compares the hit 1970s movie *Jaws* to the 700-1000 A.D. poem *Beowulf*, outlines the basic plots as: Overcoming the Monster, Rags to Riches, The Quest, Voyage and Return, Comedy, Tragedy, and Rebirth. He later adds an eighth plot based on detective novels (Booker, 2019). In his over seven-hundred-page exploration of fiction and tropes, he goes into detail about each of these basic plots and their structure:

Overcoming the Monster plots have five parts: "The Call," where the hero decides to defeat evil and may get gifts to help in this endeavor. Next is "Initial Success," where the hero may win a small battle, but the full power of the monster is not yet revealed. After is "Confrontation," where the hero faces their first serious setback. Next is "The Final Ordeal," where the hero must face a deadly trial. Lastly there is "Escape or Death," where

the hero beats trials and either escapes from or kills the monster. Examples of Overcoming the Monster plotlines are: *Gilgamesh, Beowulf, Frankenstein, The Longest Day, Live and Let Die, Star Wars (1977),* among countless others (Booker, 2019).

Rags to Riches plots also have five parts: “Initial Wickedness at Home," where the protagonist experiences an original unhappy state, and the reader is introduced to the evil figures around them. Next is “Initial Success," where the protagonist is rewarded for a first, limited success, as well as the first encounter with the prince or princess when applicable. After that there is “The Central Crisis," where everything suddenly goes wrong. Next is “The Final Ordeal," where the protagonist emerges from the crisis with character growth. All that is left is the last dark figure standing between the protagonist and the end goal. Lastly there is “Fulfilment” where the protagonist gets the princess/prince, rules the kingdom, etc. Examples of Rags to Riches plotlines are: *Aladdin*, *Cinderella*, *Some Stories of King Artur*, *Jane Eyre*, *Great Expectation*, among countless others (Booker, 2019).

Similarly, The Quest plots have five parts: “The Call," where life has become intolerable for the protagonists and they realize that they need a long journey. Next, there is “The Journey," where the heroes and their companions set out across hostile terrain. After is “Arrival and Frustration," where the party arrives within sight of their goal, but new obstacles arise. Next is “The First Ideals," where the party needs to endure a series (often three) of tests. Lastly there is “The Goal” where the party makes their last thrilling escape, winning the treasure and an assurance of a better life. Examples of this are: *The Odyssey, Moby Dick, Water ship Down,* T*reasure Island*, among others (Booker, 2019).

Voyage and Return plots also have five key parts: “The Anticipation Stage," where the protagonist is in a state which leads to a shattering experience; this is often character descriptors like bored, naive, reckless, etc. Next is “The Initial Fascination” where the protagonist experiences their first exhilarating exploration. After is “The Frustration Stage” where the mood changes to frustration, and oppression. Next is “The Nightmare Stage” where the shadow of the aforementioned feelings becomes dominating. Lastly is “The Trilling Escape” where the protagonist escapes from the shadow or new world back to where they started. The protagonist and reader now get to see how far the protagonist has come, what has been learned, and how the protagonist has grown. Examples of these plots are: *Alice in Wonderland*, *Goldilocks*, *The Time Machine*, *The Wizard of Oz*, *Peter Pan*, *The Lord of The Flies*, among many others (Booker, 2019).

Unlike the aforementioned basic plots, Comedy plots only have three key parts: part one, where we see a little world in which people have passed under a shadow of confusion, uncertainty, and frustration. Next is part two, where the confusion gets worse. Lastly things that were not previously recognized come to light, and perceptions are changed. The Shadow gets dispelled, and the world is transformed for the better. Examples of Comedies are: *A Night in Casablanca*, *Lysistrata*, *The Alchemist*, *The Tempest*, *Pride and Prejudice*, *The Philadelphia Story*, among others (Booker, 2019).

Tragedy plots go back to the five key structure: First is “Anticipation," where the hero is incomplete or unfulfilled. Next is “The Dream Stage," where the protagonist

becomes committed to a course of action. After is “The Frustration Stage," where things begin to go wrong; the hero cannot find rest, and are often compelled towards dark acts. Next is “The Nightmare Stage," where things are slipping out of control. Opposition and fate are closing in. Lastly is “The Destruction Stage," where the hero dies. Examples of this plotline are: *Macbeth*, *Lolita*, *Dr. Jekyll and Mr. Hyde*, *Bonnie and Clyde*, *Romeo and Juliet*, *The Snow Goos*e, among others (Booker, 2019).

The last basic plot described is Rebirth, which also has five key parts: part one where the hero falls under the shadow of the dark power. Next, in part two, things go reasonably well for the hero at first. This does not last long as in part three, the threats builds, and the hero is seen imprisoned in a state of living death. Next, the threat continues until the dark power triumphs. Lastly, the hero experiences redemption, often defeating the dark power. Examples of Rebirth plotlines are: *Sleeping Beauty*, *Snow White*, *A Christmas Carol*, *Crime and Punishment*, among others (Booker, 2019).

Reading this book had become the basis for this paper. While many others, namely Robert McKee, Joseph Campbell, Vladimir Propp, among others, have discussed the idea of plots, Booker’s work was the first that I had read, and thus had been a big inspiration in generating the concept for this project. Whether in agreement with Booker’s organization or not, reading Booker makes it clear that below the surface of all literature there are key similarities that link many works together. Even in a very simplified manner, the way books are sorted into categories like Fantasy, Science Fiction, etc. are a way to categorize works of literature into broad categories using contextual clues. These clues may be clear to a

reader, but there is no formula to decide if a book is a Science Fiction novel or an Overcoming the Monster plotline, or any other categorization. While there will likely never be an objective way to categorize stories into simplified categories, as inherently, moving to broader categories results in the loss of information, but this project intends to explore these categories and what structure points in stories could lead to these categorizations by a computer model.

Natural Language Processing, or NLP, is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages (Yse, 2019). Natural Language processing is a quickly growing field in computer science. Fields such as voice-driven assistants like the Amazon Echo, Google Assistant, Siri, etc., financial trading, disease prediction, news site pruning for fake news, among many other fields are constantly being improved by natural language processing. Despite all of the work being done in Natural Language Processing, long form text analysis, especially in the field of fiction has not been a focus.

This project aims to use existing practices in Natural Language Processing in order to gain insight into what underlying structures lead to our categorization and the writing tropes that have been identified. The main practices used are namely unsupervised learning, in which classification is done without the help of manual labeling (Radford, 2020), sentiment analysis, in which words are categorized as having a positive or negative connotation, part of speech tagging, in which words are classified automatically as nouns,

verbs, adjectives etc., lemmatization, where words are changed into their root word, data cleanup using stopwords, known names, etc., among other practices.

In this project, you will find an exploration of literature, sourced for the open source Project Gutenberg, using the aforementioned NLP practices, culminating the eventual clustering and categorization of fictional books from a range of categories like Fantasy, Adventure, Science Fiction, etc. his paper will explore the results and the limitations of current NLP practices in analyzing long form prose.

Related Work

Fictional Book Categorization

Outside of Christopher Booker's analysis shown above, many other researches have discussed the categorization of fiction, generally for the optimization of libraries:

In one of the older sources I came across, dating back to 1977, Pejtersen, and Annelise Mark discuss the results of user-librarian conversations about fiction recorded under everyday library conditions in Danish public libraries in 1973-74, and a further analysis of 134 conversations recorded in 1976. It was found that users' subconscious classification of fiction can be characterized by means of four dominant dimensions: (1) subject matter, including the categories of action and the course of events, psychological development and description, and social relations; (2) type of frame, including time and

geographical/social environment frames; (3) author's intention, including provision of an emotional experience or provision of information; and (4) accessibility, which includes readability and physical characteristics of a book (Pejtersen, and Annelise Mark, 1977). This was an interesting concept, and gave further proof towards the unconscious categorization of fiction, and how fiction may be categorized in an unsupervised manner, but by the nature of being based off conversations, categorization is subjective and there are no concrete categories defined.

In a more recent study published in 2007, Vernitski discusses a classification for fiction as a scholarly discipline, in contrast to the existing genre classifications used for fiction reading. Their paper proposes an intertextuality-oriented classification scheme for fiction. While not fully related to this project, this paper was important to forward the concept of classification of fiction in new ways based on the content of a text, such as an unsupervised algorithm might do.

In 2008, an article by Richard Maker discusses the differences between bookstore and library organization and the importance of “the genre stigma." This article mentions how readers may believe they do not enjoy Fantasy novels, which, in actuality, will push them away from novels that they may enjoy. Again, while this article does not directly relate to this project, it gives merit to the idea that there are more classifications than just by genre, and that the core structure to a fictional work matters just as much as the setting and other clues that would lead to a genre classification.

Natural Language Processing

There are many resources relating to natural language processing and text classification, however not many relate directly to longform fiction classification. To start, the primary resources used in learning the principles of natural language processing were: firstly, the textbook *Natural Language Processing with Python* In which the basic principles of natural language processing are taught using the nltk library in Python. Many of the concepts used, like sentiment analysis, lemmatization, stopword removal, among others were based on the concepts taught in this book. Next, the course "A Code-First Introduction to Natural Language Processing" by the fast.ai team was important in building a foundational understanding of natural language processing. In this course, concepts like tokenization, naive bayes, regex, sentiment analysis, and others are taught with code implementations and video lecture recordings. These sources, among others, were fundamental in building the knowledge needed for this project.

Beyond works used to teach the principles of natural language processing, there were multiple prior experiments that helped to build the foundation for this project as well.

Published works like "Automatic Affect Recognition Using Natural Language Processing Techniques and Manually Built Affect Lexicon." by Cho, Y. H., and K. J. Lee and "Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose" by Dell'Orletta, Felice, Montemagni, Simonetta, and Venturi, Giulia, while dealing with other languages, provided interesting

input into what is needed to analyze a longform text. “Automatic Affect Recognition Using Natural Language Processing Techniques and Manually Built Affect Lexicon.” uses a manually built affect lexicon in order to be able to detect various emotional expressions in a Korean textual document. Like many current natural language processing projects, their project deals with a smaller source, compared to the large texts used in this project, but gave an interesting look into the complexities of emotional analysis, an important piece in how humans would categorize texts. “Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose” was on the surface a more similar project to this project. Long form texts were analyzed and categorized into genres. These genres were much more simplified however, namely literature, journalism, scientific studies, etc. While their project was based in categorizing less complex categories, it was shown that things like the lexicon used, the sentence length, the word length, etc. changed throughout different genres. This was promising, as similar differences may be apparent in fictional genres as well.

Further work in this field has been done like “Personality Profiling of Fictional Characters Using Sense-Level Links between Lexical Resources” by Flekova, Lucie, and Iryna Gurevych. In their project, characters from many fictional books were analyzed and categorized as introvert or extrovert. To do this, things like actions taken, words used, adverbs used, descriptions, etc. were used to generate the group they most fit into. While their project deals with a different issue than this project, it shows how parts of speech, actions, and words used can accurately give a peek into what makes a character, and this may expand into what may make a book a certain genre.

Other general natural language processing works were used as references. “NLP on spoken documents without ASR” by Mark Dredze, Aren Jansen, Glen Coppersmith, and Ken Church breaks audio speeches into one second long phrases that are then used in clustering, a concept that is adapted in this project in the part of speech clustering section. Furthermore, sources like “Unsupervised and Semi-supervised Clustering: A Brief Survey” by Nizar Grira, Michel Crucianu, Nozha Boujemaa showed the merit and use of unsupervised clustering, which was used in this project to classify a large amount of texts by their fundamental pieces.

All in all, while there were no sources that were directly built upon for this project, many existing projects, often which focused on smaller pieces, showed the use of fundamental language used in the classification of fictional works. This project will expand on those fundamental ideas, as well as the aforementioned literature-based text classification research in order to explore and classify long form fictional pieces by their fundamental pieces.

Experimentation

Problem

Here we discuss the initial experimentation performed. Using a set of fictional texts, current Natural Language Processing practices are applied to the texts. The purpose is to

explore the fundamentals and key similarities between texts. The end goal is to use the explored data to categorize the texts, much like how humans would read a work of fiction and group it with similar works, without much thought.

In the rest of this section, we describe the experimental setup, and the multiple Natural Language Processing practices applied, as well as the observations and analysis of the results.

Experimental Design

As previously mentioned, the purpose of this project is to explore the structure of long form fictional literature using Natural language Processing practices. The first step was to acquire the resources in order to have a varied, relatively large set of data to train, test, and analyze.

Setup:

This project uses Project Gutenberg as the primary data source. Project Gutenberg is a library of over 60,000 free eBooks with focus on older works for which U.S. copyright has expired (Gutenberg, 2020). The first step was to manually acquire all book IDs in which a fictional work is included. These IDs were acquired from digging through the website and grabbing IDs as well as the tags given by Project Gutenberg i.e. Fantasy, Historical Fiction, etc. Once the IDs were acquired, the Project Gutenberg Python Library was used

to load in the full text for all gathered IDs. These texts are organized using a pandas dataframe with the text, title, and the project Gutenberg tag. Minor cleanup was needed to remove headers and endnotes added by project Gutenberg, and to include only English books as an effort to reduce noise in the results. For the language detection, the Python library langdetect was used. This library is a port of Nakatani Shuyo's language-detection library for Python, which supports fifty five languages.

Part 1:

After the texts were all gathered, the first step was to analyze random sets of books to see what potential differences in texts may be. This part is separated from any categorization and was included as an exploration of the building blocks of stories using Natural Language Processing, to later be used in analysis of the categorization. This was done in two parts originally: testing how the parts of speech and the sentiment of words may change in the duration of a text.

Both of these sections followed the same structure: First read in the texts. This was explored in the above Setup section. Next, punctuation is removed, but not stopwords. While in preprocessing it is often commonplace to remove stopwords, for these two pieces, stopwords were not removed in order to retain meaning. For example, in sentiment analysis, “This movie is not good” could be changed to “movie good” when removing

stopwords, which has a different sentiment. Lastly, the text is processed, in ways that will be later be expanded on, and a dispersion plot is outputted as a visual representation.

For the part of speech section, the nltk Python library was used. The nltk Python library is a leading platform for building Python programs to work with human language data, providing easy-to-use interfaces to over fifty corpora and lexical resources (nltk, 2020). Using this library, there is a "part of speech" tagger, where an imputed string will result in a list of tuples with the word and the part of speech. The dispersion plot is outputted representing a tick for the most common and useful parts of speech, namely nouns, verbs, adjectives, adverbs, proper nouns, conjunctions, and personal pronouns (I, he, she, etc.) in order of occurrence.

For the sentiment analysis, the nltk Python library was once again used. Using this library, and the included vader sentiment analyzer, each sentence in the text was analyzed. Using the returned floating-point value for the compound score, sentences are determined to be positive, negative or neutral. Sentences were used since context is often important in determining sentiment and limiting to words often will not get a fully accurate result. The dispersion plot is outputted representing a tick for each of the categories: positive, negative, and neutral, in order of occurrence

These plots were outputted for a random selection for each project Gutenberg category, for example mystery fiction, science fiction, etc. These plots were used as a basis for the next part, as well as in the analysis of the overall results, giving important insight

to the data set and how it is structured, as well as the potential differences between categories.

Part 2:

After the original exploration, the next part of this project focuses on the categorization of texts. This was split into two sub-sections, both with similar structure: the clustering of the texts by their words, and the clustering of texts based on the part of speech of the words.

Both of these sections begin with the same structure: first read in the texts. This was explored in the above Setup section. From here, the subsections are handled differently.

For the part of speech section, stopwords are not removed for the same reasons mentioned in the previous part. The text is split on all punctuation. For each entry, a "word" is created from the combined parts of speech. For example, "Bob ate cookies." Would become NNP VB NN for Proper noun, verb then noun, which would result in the "word" NNPVBNN.

For the word clustering, additional preprocessing was necessary in order to remove some noise from the results. The preprocessing used was to first remove punctuation. This

was done by tokenizing the text using the nltk library. Next, stopwords were removed. The purpose of this is to remove common words that would add noise to the results, which would make interesting words not hold as much weight in clustering. The stopwords used were the standard nltk stopword set, expanded slightly with words like “Mr.," “Mrs.," as well as some characters like “…” among others. Another step in cleanup was to remove common names. The purpose again was to remove words that may overshadow interesting words in the results. For example, in *Moby Dick,* Captain Ahab may be mentioned often. Ahab is a word that would not need to be considered in clustering as another book having a character with the same name does not denote that they are similar. To do this, a dictionary of common names was used to remove occurrences from the text.

Table of Stopwords

Word	Origin
i	nltk
me	nltk
my	nltk
myself	nltk
we	nltk
our	nltk
ours	nltk
ourselves	nltk
you	nltk
your	nltk
yours	nltk
yourself	nltk

yourselves	nltk
he	nltk
him	nltk
his	nltk
himself	nltk
she	nltk
her	nltk
hers	nltk
herself	nltk
it	nltk
its	nltk
itself	nltk
they	nltk
them	nltk
their	nltk
theirs	nltk
themselves	nltk
what	nltk
which	nltk
who	nltk
whom	nltk
this	nltk
that	nltk
these	nltk
those	nltk
am	nltk
is	nltk
are	nltk
was	nltk
were	nltk

be	nltk
been	nltk
being	nltk
have	nltk
has	nltk
had	nltk
having	nltk
do	nltk
does	nltk
did	nltk
doing	nltk
a	nltk
an	nltk
the	nltk
and	nltk
but	nltk
if	nltk
or	nltk
because	nltk
as	nltk
until	nltk
while	nltk
of	nltk
at	nltk
by	nltk
for	nltk
with	nltk
about	nltk
against	nltk
between	nltk

into	nltk
through	nltk
during	nltk
before	nltk
after	nltk
above	nltk
below	nltk
to	nltk
from	nltk
up	nltk
down	nltk
in	nltk
out	nltk
on	nltk
off	nltk
over	nltk
under	nltk
again	nltk
further	nltk
then	nltk
once	nltk
here	nltk
there	nltk
when	nltk
where	nltk
why	nltk
how	nltk
all	nltk
any	nltk
both	nltk

each	nltk
few	nltk
more	nltk
most	nltk
other	nltk
some	nltk
such	nltk
no	nltk
nor	nltk
not	nltk
only	nltk
own	nltk
same	nltk
so	nltk
than	nltk
too	nltk
very	nltk
s	nltk
t	nltk
can	nltk
will	nltk
just	nltk
don	nltk
should	nltk
now	nltk
man	added custom
men	added custom
ll	added custom
just	added custom
did	added custom

mr	added custom
sir	added custom
thee	added custom
us	added custom

Lastly, the words were lemmatized, which is the process of reducing the different forms of a word to one single form, for example, reducing “builds," “building," or “built” to the base form, referred to as lemma, “build." The purpose of the lemmatization is to remove noise in the clustering and make the clustering find matches more easily. The best practice for accurate lemmatization is to include the part of speech so that words are best changed to be only the root words. For this to be done, the nltk part of speech tagger was used, then these were limited to include only nouns, verbs, adjectives, and adverbs. All others are simplified to nouns. The purpose of this that lemmatization of nouns is the simplest, so there will be fewer unwanted.

After the preprocessing was done, the next steps were similar for both subsections. The books were then split into a training set and a testing set. The set was randomly selected with sixty percent of texts going to training and forty percent going to testing. The random collection was also weighted where each Project Gutenberg category was split into the sets with sixty percent to training and forty percent to testing, so that the clustering was trained evenly on each category. The purpose of splitting evenly is to avoid having major bias from training primarily on a single Project Gutenberg Category. For example, if the randomization chose the training set to be mainly science fiction novels, the clusters may

have groupings that are not useful in training and would lead to results that do not accurately reflect the inherent differences in texts as designed.

Next, the scikit-learn Python library was used for generating clusters in an unsupervised manner. *Scikit-learn* is a Python module for machine learning, which includes many built in resources, including unsupervised clustering of data. Using the previously generated training sets, clusters are generated for what the unsupervised, machine learning algorithm generates as similar texts. These clusters include a list of words, which may be actual words or the part of speech "words" from the sources. Each word in the cluster has a weight applied to it as well. This is representative of a vector in the direction of that cluster, which will later be used to classify data. For the number of clusters, inspiration was taken from Christopher Booker, and eight clusters were generated, representing his 7 basic plots as well as his additional plot for detective stories.

After clusters are generated, the testing set was then classified into which cluster they most fell into. To do this, the text is walked through. For each word pulls the classification in the direction of the cluster it appears in depending on the weighting defined by the training set. Like is like a set of vectors added to each other. Each word will pull the text towards a cluster until the text is fully walked through and the cluster that the text is closest to becomes the classification.

The final step was the output of the data. To fully display all of the results, two sets of pie charts were generated for each sub-section. Firstly, eight pie charts are generated,

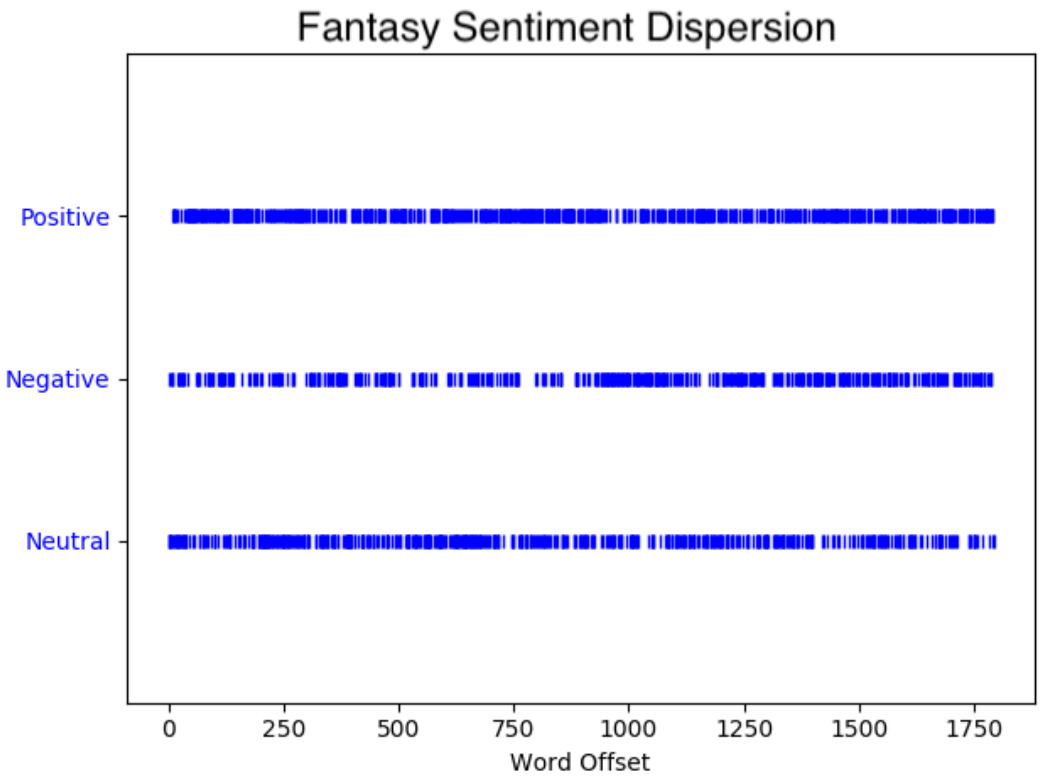
one for each cluster, where the ratios between Project Gutenberg categories for each cluster is displayed. The purpose of these charts was to analyze patterns between the data in each cluster. If each cluster had meaningless data, the pie charts would all look very similar. Next, a set of fourteen pie charts are generated, one for each Project Gutenberg category. The purpose of these charts was to analyze patterns between categories to see which clusters may have dominated each. If a single cluster dominated all categories, we would know that there was too much noise to get meaningful results.

Results

In this section, the results of a run of each part of the project are displayed and analyzed. For readability purposes, only some charts are included in this section, however the full set of all cluster visualizations is available in the later Figures section.

For Part 1, sentiment analysis, fourteen charts were generated, one chart per Project Gutenberg Category. An example output is Chart 1.1, shown below:

1.1 Example of a Dispersion Plot for Sentiment

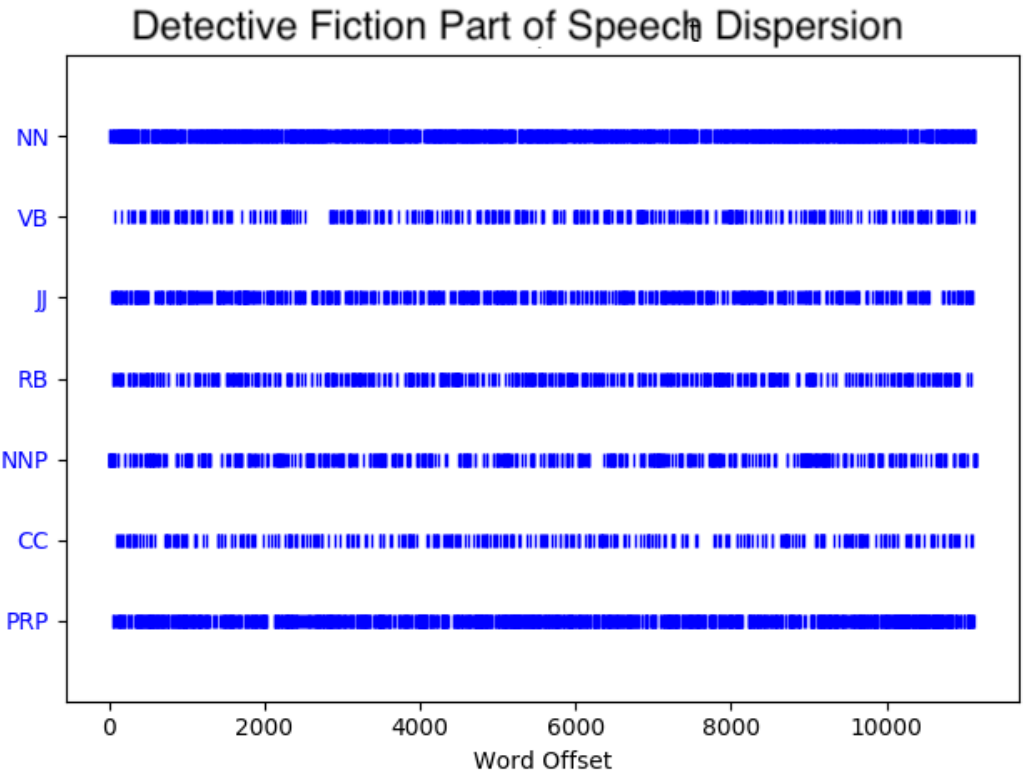


As seen, for a random book categorized as “fantasy” by Project Gutenberg, the words were analyzed as primary positive, with negative words increasing towards the end of the file. This does not give any major insight, but makes logical sense as during the climax of a story, generally part three and four in Christopher Booker’s basic plots, a more intense, dramatic scene could logically be accompanied by an increase in negative words. In comparing the sentiment to the parts described by Booker, along with general knowledge of Fantasy texts, this sample novel would likely fall into a Voyage and Return or Quest plot, where the negative feeling grows with the threat until the end. While this is not

anything quantifiable, it gave a good basis that the words and their underlying meaning could be creating a story for readers without even including any context from setting.

For Part 1, part of speech analysis, similarly, fourteen charts were generated. An example is Chart 1.2, shown below:

1.2 Example of a Dispersion Plot for Parts of Speech



In this plot, the parts of speech are shown using the nltk tags; nouns, verbs, adjectives, adverbs, proper nouns, conjunctions, and personal pronouns are displayed as NN, VB, JJ, RB, NNP, CC, and PNP accordingly. Displayed in this example is a randomly chosen book from the Project Gutenberg category “Detective Fiction." Shown, Nouns, and

Personal pronouns are most prominent in this particular text. This means words like he, she, I, as well as nouns that may refer to setting, objects, or people are used often. Again, while this does not give quantifiable results, it makes logical sense. A Detective novel often would talk about clues, motives etc. This would explain the increased use of personal pronouns, and the decreased use of verbs, as these novels are generally not action focused. Again, these results give a good basis to how the words and their underlying meaning may be creating a story for readers without any context or setting.

For Part Two, word clustering, after the texts were read in and preprocessed, the words were clustered. Below, a table of the top 10 words in each cluster is included.

Example Phrases from Word Clusters

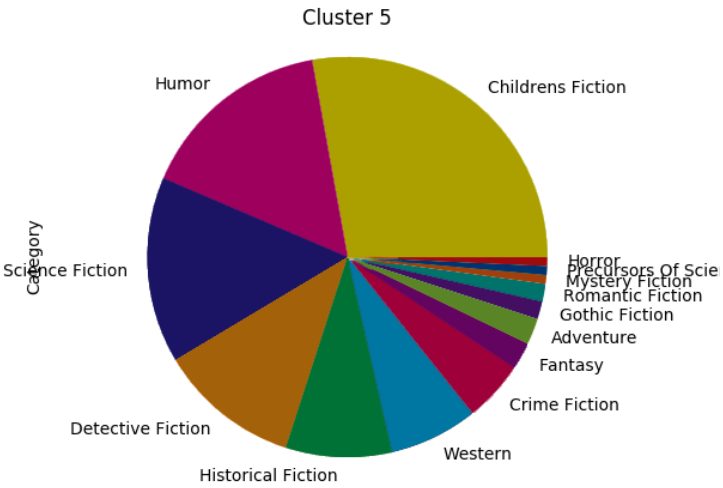
Cluster 0:	**Cluster 1:**	**Cluster 2:**	**Cluster 3:**	**Cluster 4:**	**Cluster 5:**	**Cluster 6:**	**Cluster 7:**
pat	one	say	lifeboat	say	say	one	say
marl	say	one	boat	one	one	would	one
entity	would	upon	one	would	would	upon	would
cetus	could	would	sea	could	could	could	could
could	upon	thou	ship	like	well	like	like
symbol	time	could	slagg	back	time	said	time
nothingness	two	well	vessel	know	little	time	back
cogito	well	shall	say	time	know	come	know
exist	little	time	crew	get	like	back	see

space	great	little	wreck	little	see	great	two

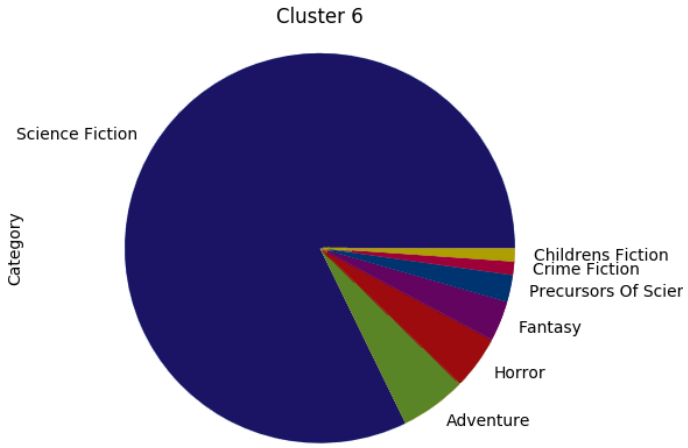
While the top 10 words do not give a full picture of each category, some basic words dominate the most common words for many categories. This example gives insight into what some of the categories might look like. For example, Cluster 3 seems to include many nautical words. On first analysis, many books in the adventure Project Gutenberg category, or children's fiction Project Gutenberg category may be categorized into this cluster, as many adventure stories include sailing across the seas. Also, some science fiction stories with ships may be classified into this category as well. Other insightful categories from other runs have been categories that have themes like time, space, or science.

After the original generation of the clusters, the testing set was classified using the clusters.

2.6 Cluster

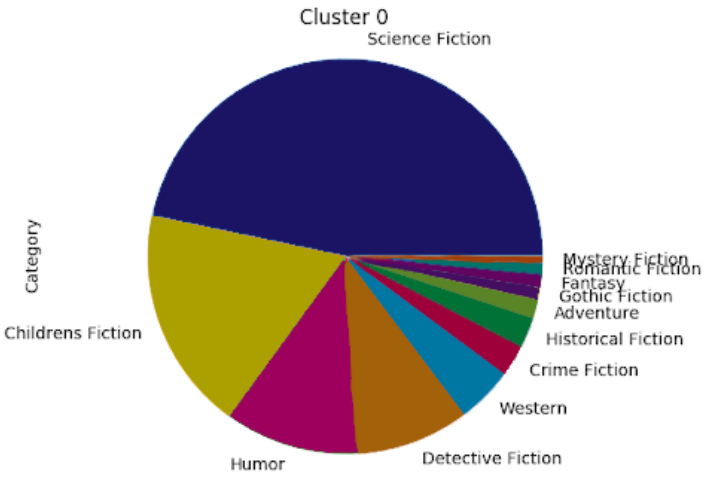


2.7 Cluster 6



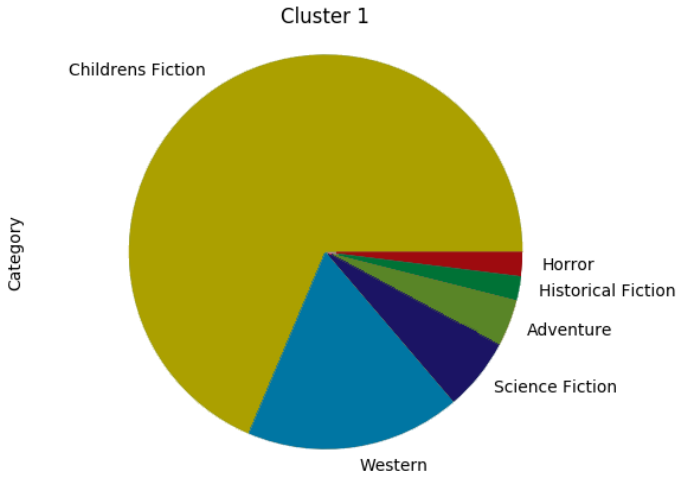
Cluster 5 and Cluster 6 are shown above. As seen, Cluster 6 has a large proportion of Science fiction stories. This is likely partially because in Project Gutenberg, the largest category of fiction included is science fiction, but this cluster implies that, beyond this, there are words that are pulling a larger amount of science fiction stories compared to others. In contrast, Cluster 5 has a large amount of humor novels as well as children fiction compared to science fiction, despite the humor and fiction categories having a much smaller set compared to science fiction. This implies that the words in Cluster 5 are much more lighthearted compared to Cluster 6 with high intensity categories like adventure, and horror joining science fiction. While there is no way to quantify Christopher Booker's basic plots, and the purpose of this project is not to find matches to his groupings, but instead to find underlying similarities in longform fiction, this dichotomy resembles the difference between the Comedy plot, and a more intense plot like Tragedy. In this example, Cluster 5 would represent Comedy, including often lighthearted or positive categories like Comedy and children's fiction. As seen in Part One, sentiment is an important underlying representation of stories, and Cluster 5 seems to represent a more positive sentiment.

2.1 Cluster 0



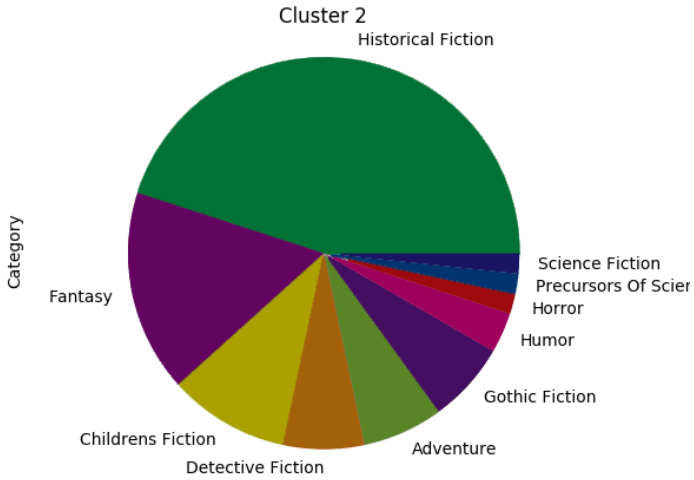
Cluster 0, shown above, unlike the previously described clusters, does not display much information. The ratios within this cluster greatly resemble the ratios between the number of books in each section, implying that this category is equally categorizing all Project Gutenberg Categories. This is likely due to a large number of common words used in all fiction.

2.2 Cluster 1



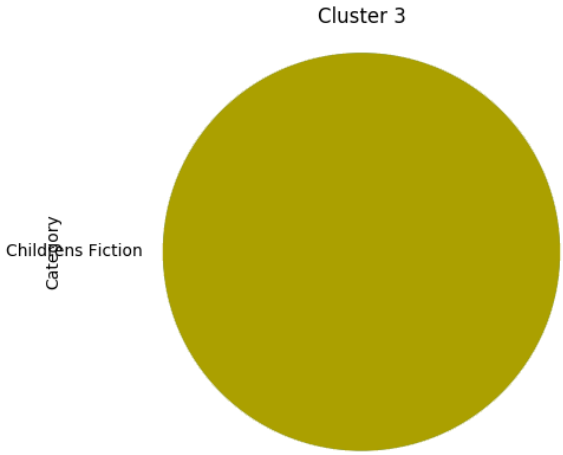
Cluster One, shown above, has a large ratio of Children's Fiction and Western novels. This implies that this cluster may have many words that invoke a lighthearted adventure tone, possibly including sections involving characters that may be cowboys. This connection makes intuitive sense, as cowboys may be common in both Western novels, as well as children's fiction.

2.3 Cluster 2



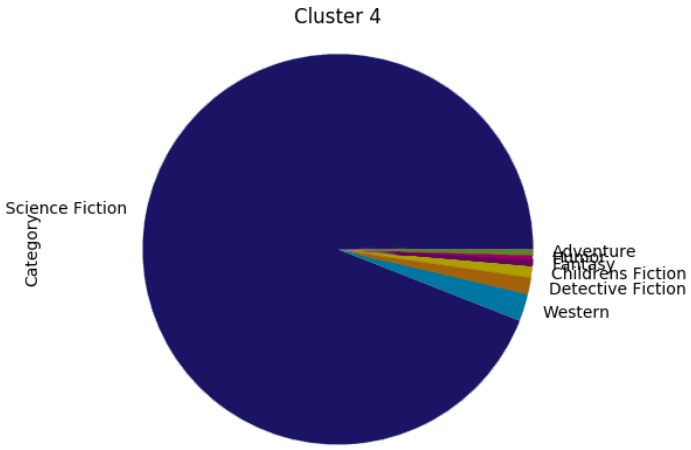
Cluster 2, shown above, has a large amount of Historical Fiction stories. This is notable since, in comparison to others, this category is small. This implies that there may be a large amount of words that imply time periods and settings. This would also explain the amount of fantasy stories, which often are set in historical times as well. children's fiction, and detective fiction novels may also have historical settings, making further intuitive sense.

2.4 Cluster 3



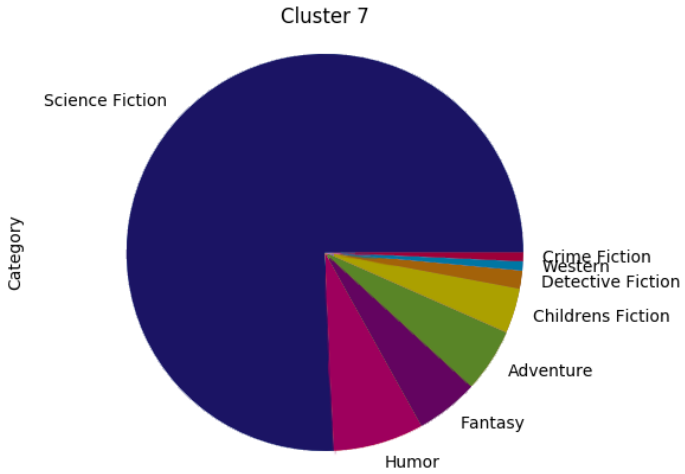
Cluster 3, which was discussed in the above section as including many nautical words, only includes children's fiction novels from the testing set. This makes intuitive sense, as many children's fiction novels could include boats, often pirates or other adventurers.

2.5 Cluster 4



Cluster 4, shown above, has a very large proportion of science fiction. While this is likely partially due to the larger ratio of science fiction stories overall, it also implies a further similarity between texts in this section making them categorize as similar. Likely, this includes common science fiction words, possibly relating to space, time travel, aliens, etc.

2.8 Cluster 7

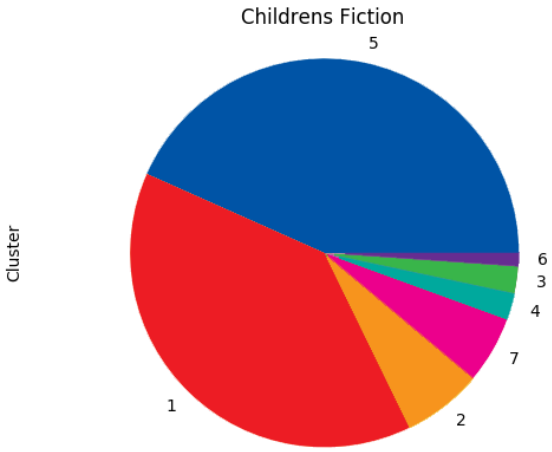


Lastly. Cluster 7, shown above, while somewhat following the ratios of the stories included in the data set, also has a notably large portion of humor stories, which is a relatively small category.

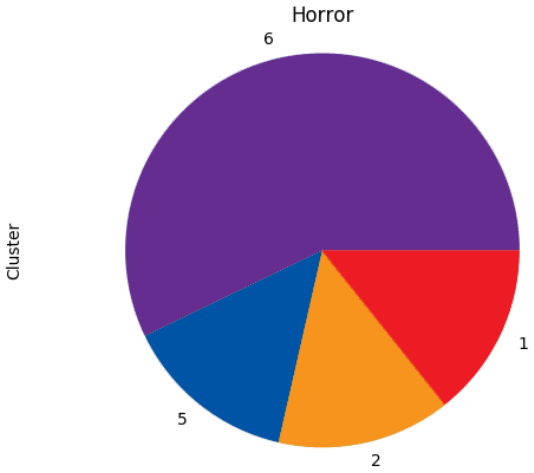
Overall, while the results are not definitively finding similar literary structures in the clusters, there are similarities to known categorization, like the aforementioned example.

Next, figure 4.2 and 4.8 are shown and analyzed below.

4.2 Children's Fiction



4.8 Horror

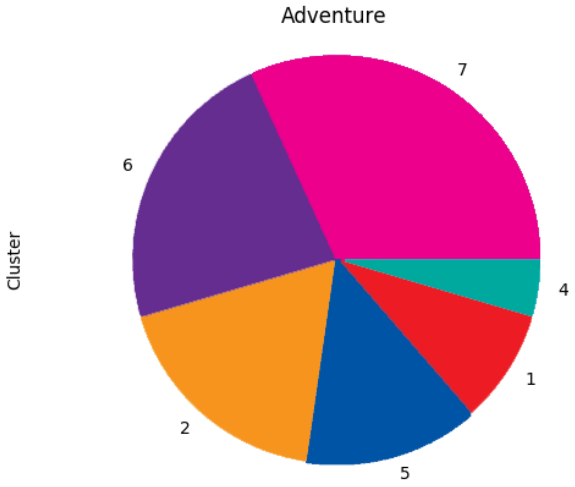


Charts for children's fiction and horror are shown above. As seen, there is little crossover between these genres. This makes intuitive sense: children's fiction often handles more upbeat themes, whether that may be more adventure based, or comedy based, by the

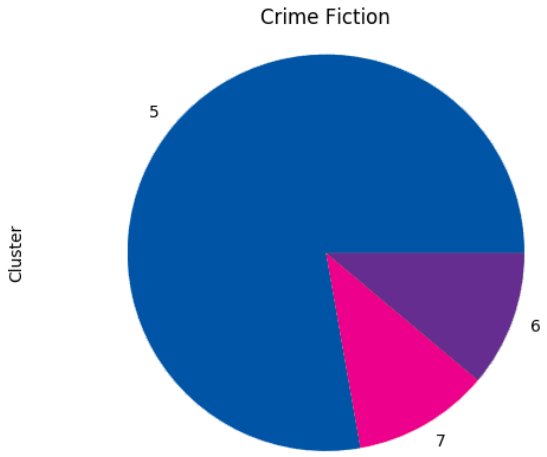
nature of being “for children," these books will be lighter content. In contrast, horror books are often on the opposite side of the spectrum: this may be gore, violence, monsters, etc. Thus, it is important that there is not much crossover between these.

The remaining Project Gutenberg Categories are shown next. There are some interesting notes, like how Crime fiction and Detective fiction both are primary Clustered into Cluster 5. Furthermore, similarities between categories like Children’s Fiction and Western, among others, stand out.

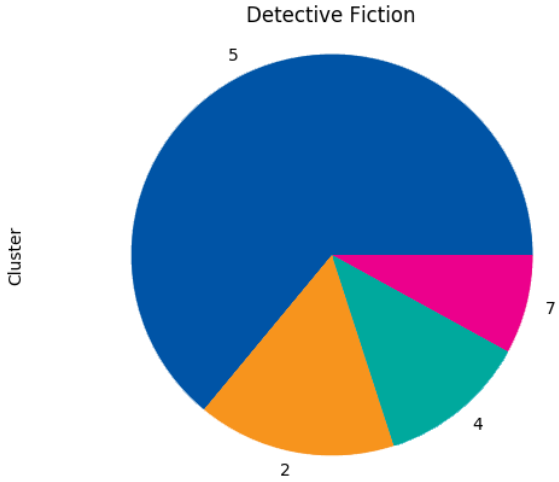
4.1 Adventure



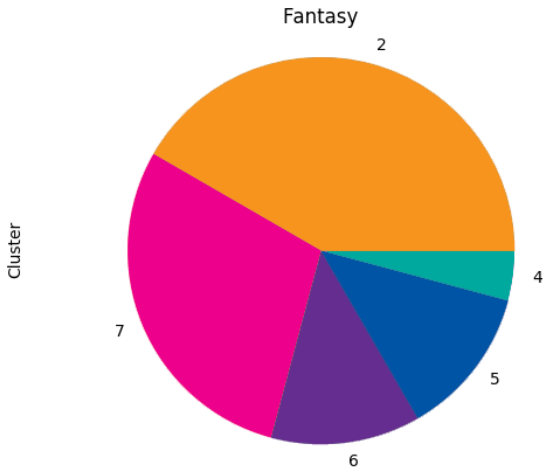
4.3 Crime Fiction



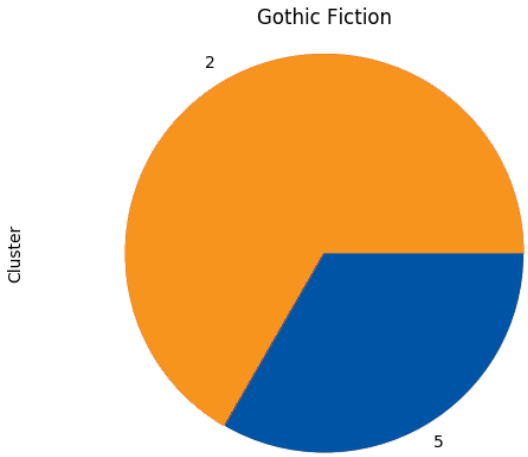
4.4 Detective Fiction



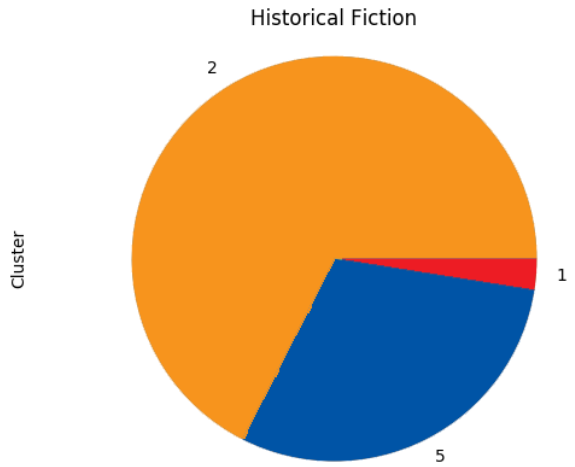
4.5 Fantasy



4.6 Gothic Fiction

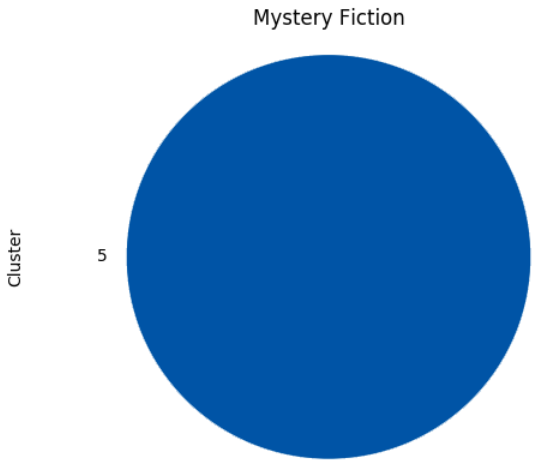


4.7 Historical Fiction

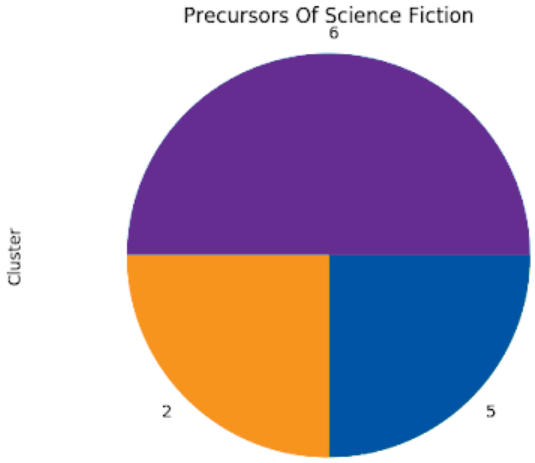


4.9 Humor

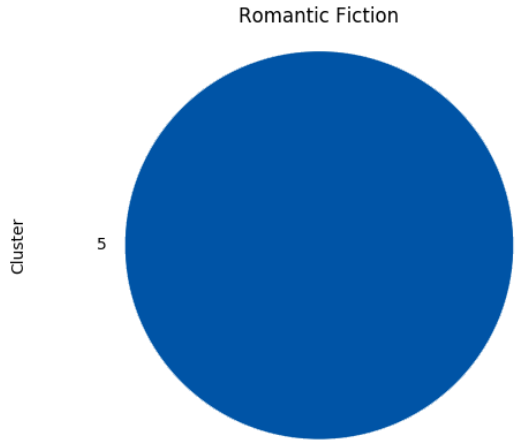
4.10 Mystery Fiction



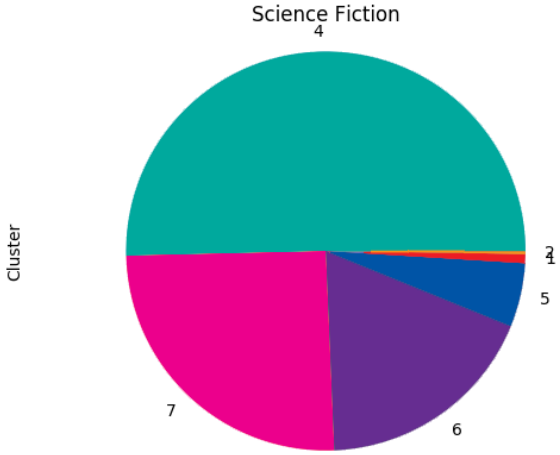
4.11 Precursors of Science Fiction



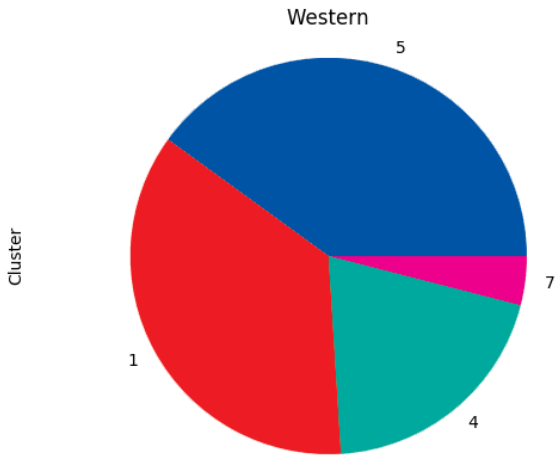
4.12 Romantic Fiction



4.13 Science Fiction



4.14 Western



There are other important notes that come from these plots: While there are seemingly some clusters that catch a plethora of Project Gutenberg Categories, likely due to noise in the data, it is important to see differences between these plots showing that the categorization is not happening arbitrarily. If the categories were being chosen blindly, each Cluster Graph and Category graph would look very similar, representing nothing more than the larger number of texts in some categories compared to others.

Overall, despite some noise in the graphs, others represent narratives that make intuitive sense, matching not only knowledge of genre classification, but matching Booker's categorizations as well. This is then continued to cluster based on parts of speech, mirroring Part One analyzing both sentiment and parts of speech.

For Part Two, part of speech clustering, after the texts were read in and preprocessed, the part of speech "words" were clustered. Below, a table of the top 10 "words" in each cluster is included.

Example Phrases from Part of Speech Clusters

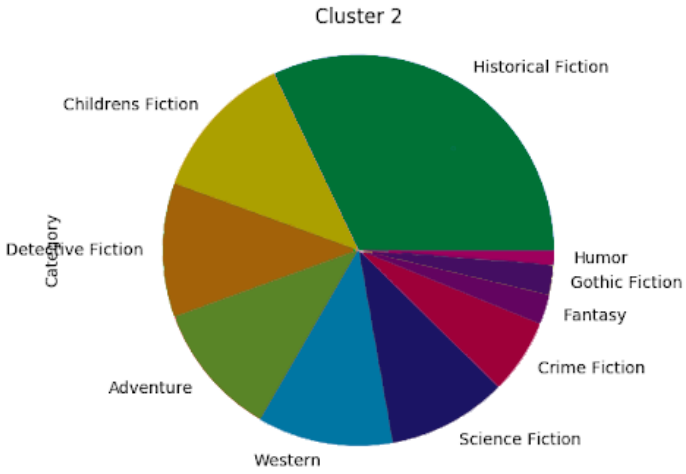
Cluster 0:	**Cluster 1:**	**Cluster 2:**	**Cluster 3:**	**Cluster 4:**	**Cluster 5:**	**Cluster 6:**	**Cluster 7:**
nn	nn	nn	nnp	nn	nn	nn	nnp
nnp	vbdnnp	nnpvbdnnp	nn	nnp	prp	nnp	nn
nnpvbd	nnp	rb	nnpvbd	rb	nnp	prp	dtnnvbdrbvb nvbntovbton npnnp
prp	rb	nnp	prpvbd	prp	rb	rb	innnpvbddtjj nninwdtprpm dvbrpdtnn
nns	prp	prp	rb	nns	nns	prpvbd	prprbvbddtjj nnrbincdnn
nnpnnpvbd	nns	nns	prp	prpvbd	prpvbd	nns	nninjjsinrbrbi nnnpinprpvb pinnnp
dtnnvbd	prpvbd	jjnn	nns	jjnn	jjnn	jjnn	nnsrpdtnnscc vbdrpinprp

indtnn	jjnn	nnprpvbd	jjnn	inprp	jj	nnpvbd	nnpmdvbinn npvbdvbn
nnpvbdinprp	uh	nnvbdnnp	uh	nnpnnp	inprp	vbdnnp	prpmdvbtovb prpvbdvbn
rb	vbddtnn	nnpnnp	nnpnnp	vbdnnp	dtnn	jj	nnnnjjvbdrbc djjnn

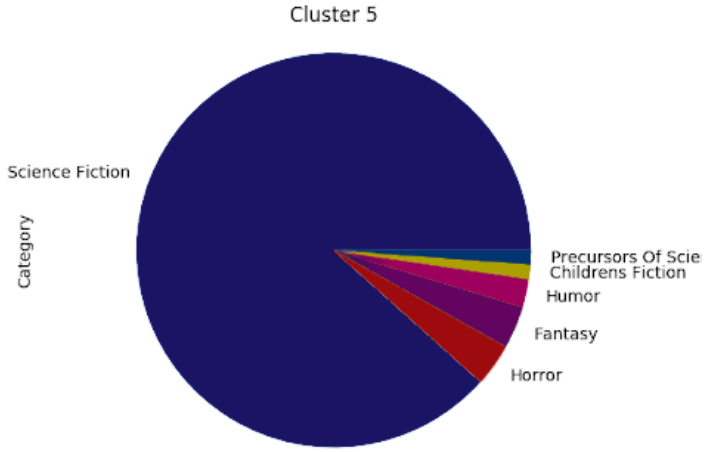
While these results are harder to make sense of an immediate glance, on investigation some interesting similarities are seen. For example, Cluster 1 has vbdnnp fairly high. While this does not clearly show what texts may have phrases like this, it does show that phrases beginning with verbs are common in this Cluster. This implies more action-based novels may be found here. Another note is Cluster 7, which seems to be filled with longer phrases. On original analysis, this implies a category that is mainly catching noise from long phrases between punctuation, this category may also catch texts that are wordier. For example, older texts seem to have much longer sentences than some modern prose.

After the original generation of the clusters, the testing set was classified using the clusters. As an example. 3.3, 3.6 will be shown below.

3.3 Cluster 2



3.6 Cluster 5

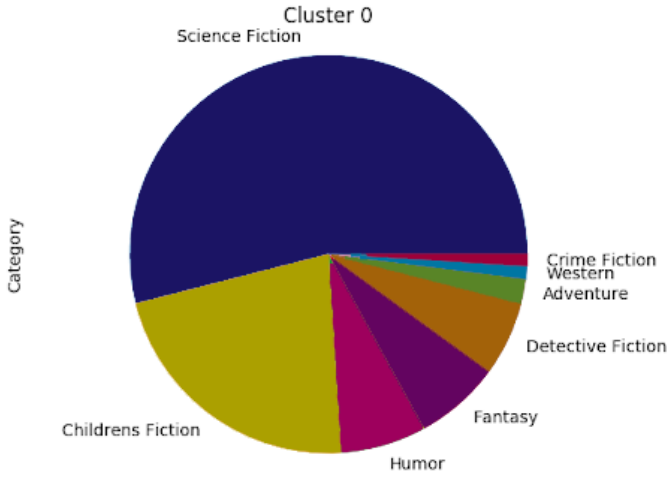


Shown above are Clusters 2 and 5. Interestingly, Cluster 2 has a very large proportion of historical fiction, which stood out since that is a fairly small category ratio

wise. Intuitively, categories like Historical Fiction, Detective Fiction, Crime Fiction, and others, may have fewer verbs than others, as discussed with Detective Fiction in Part One. This is likely the case in this category: many, less action-based novels may be caught here.

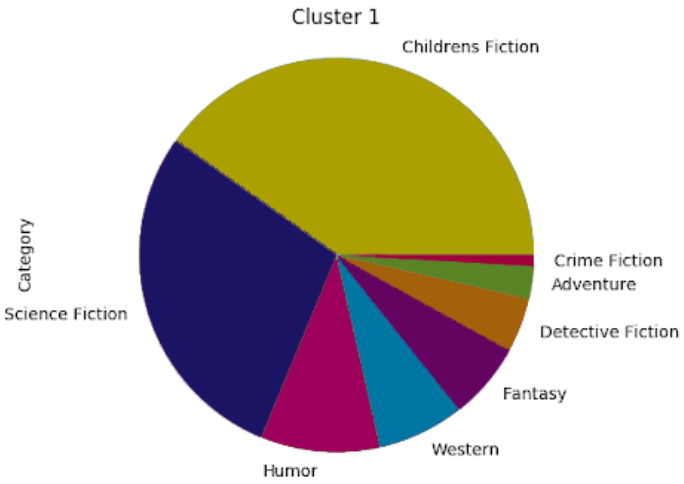
Cluster 5 looks very similar to Cluster 6 in the word analysis section, with Science fiction and Horror taking top spots, dominated by Science Fiction. While this is not as intuitive as the word categorization for these categories, it is an interesting parallel between both categorizations, implying subtle similarities between many stories.

3.1 Cluster 0



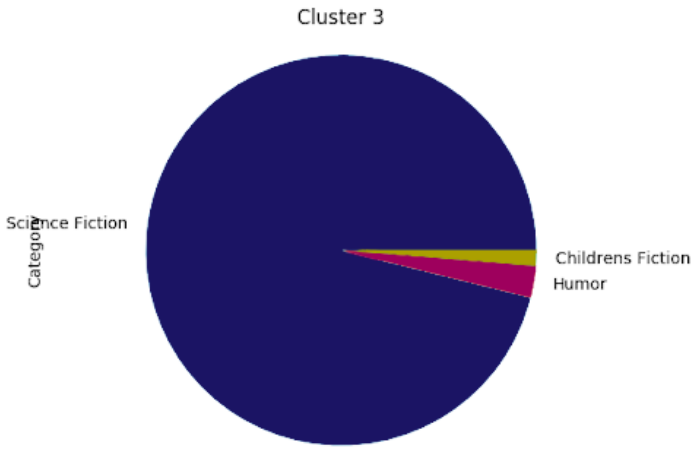
Cluster 0, shown above, similar to Cluster 0 in the Word Categorization section, greatly matches the ratios of each Project Gutenberg Category in the data set. Again, this does not portray any meaningful results and is likely the result of common phases used in every work of fiction.

3.2 Cluster 1



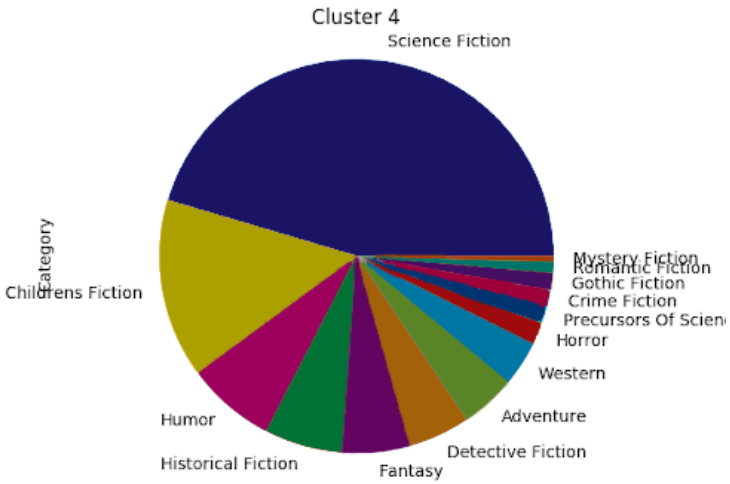
Cluster 1, shown above, is similar to Cluster 0 in the sense that is greatly matches the ratios of each Project Gutenberg Category in the data set, however, it is notable that Children's Fiction has a larger proportion, implying a possible increase of action phrases in this Cluster, which would likely be common in Children's Fiction.

3.4 Cluster 3



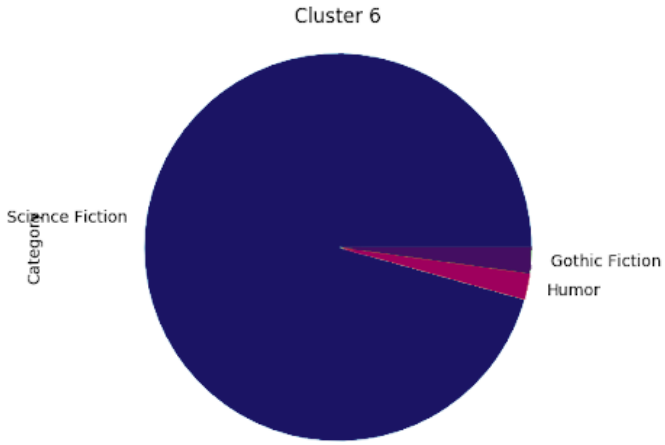
Cluster 3 shown above, similar to Cluster 4 in the above Word Categorization section, has a very large proportion of Science Fiction. While this is likely partially due to the larger ratio of Science fiction stories overall, it also implies a further similarity between texts in this section making them be categorized as similar.

3.5 Cluster 4



Cluster 4, shown above, again greatly matches the ratios of each Project Gutenberg Category in the data set. Again, this does not portray any meaningful results and is likely the result of common phases used in every work of fiction.

3.7 Cluster 6



Cluster 6, shown above, is very similar to Cluster 3 above, having a very large proportion of Science Fiction. Again, while this is likely partially due to the larger ratio of Science fiction stories overall, it also implies a further similarity between texts in this section making them be categorized as similar.

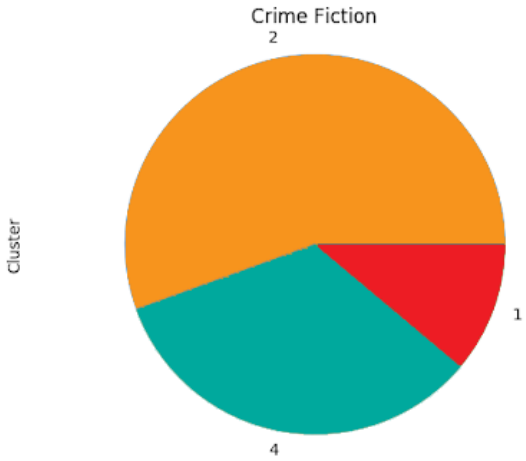
3.8 Cluster 7

No matches to testing set in this run. This is likely due to the nature of this Cluster, which is many large phrases, as discussed previously.

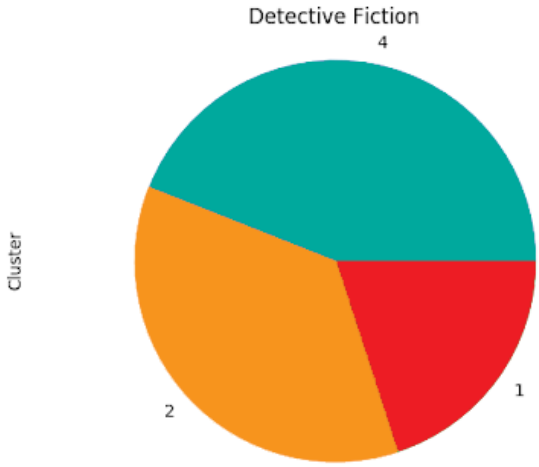
Overall, despite some flaws which are discussed later, some interesting pictures are being painted by these graphs, paralleling the clustering done in the word section.

Next, a similarity between two Gutenberg categories is shown below in figures 5.3 and 5.4

5.3 Crime Fiction



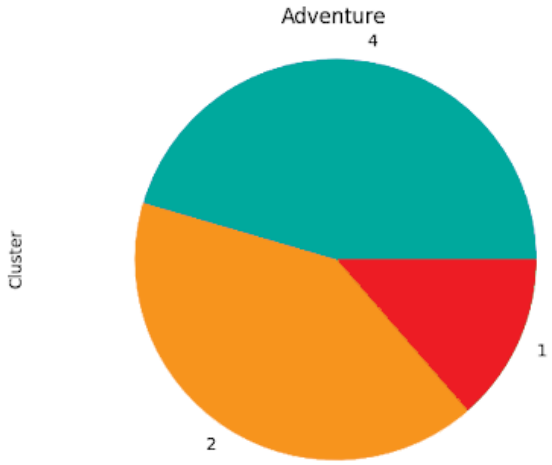
5.4 Detective Fiction



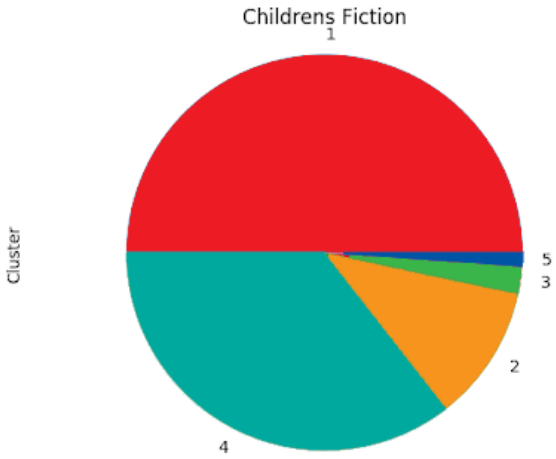
The charts for Crime Fiction and Detective Fiction are shown above. It is clear that these two charts are very similar, which intuitively make sense. While these categories are not identical by any means, the nature of Detective Fiction dealing with the solving of crimes, as well as Crime Fiction dealing with the commitment, and sometimes solving of crimes leads to these categories likely being similar.

Shown below are the remaining Project Gutenberg Category charts. There are some interesting notes, like both Mystery and Romantic fiction only being categorized as Cluster 4, as well as similarities like Westerns and Adventure stories being incredibly similar, among others.

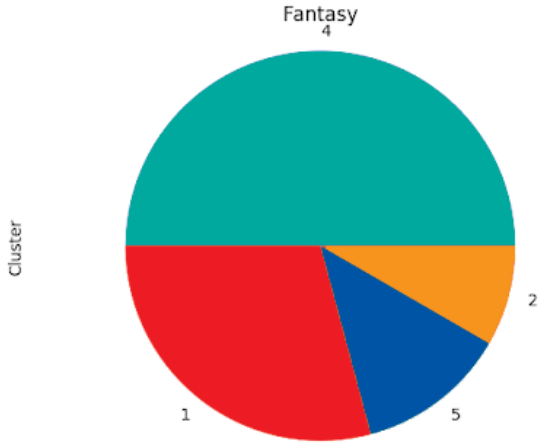
5.1 Adventure



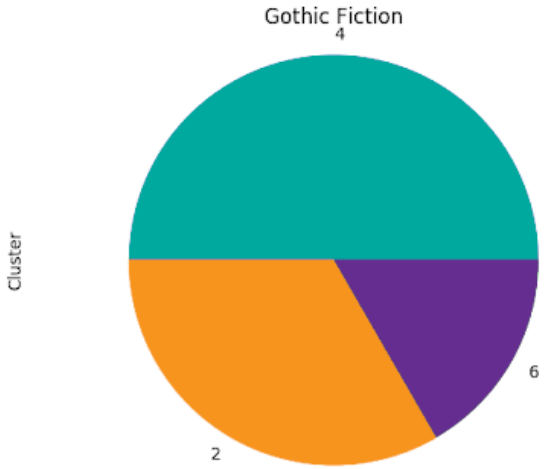
5.2 Children's Fiction



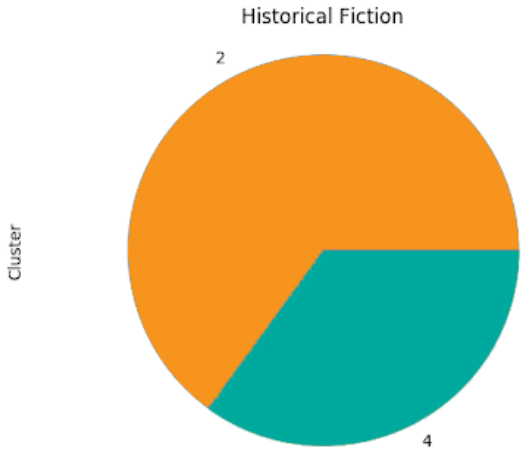
5.5 Fantasy



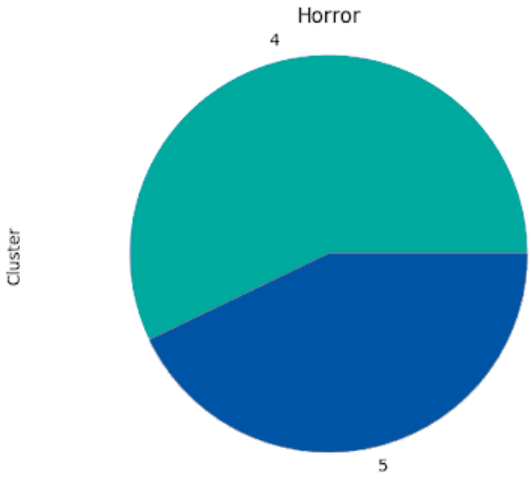
5.6 Gothic Fiction



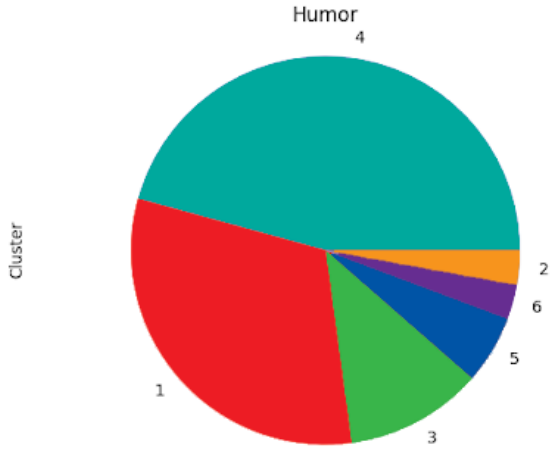
5.7 Historical Fiction



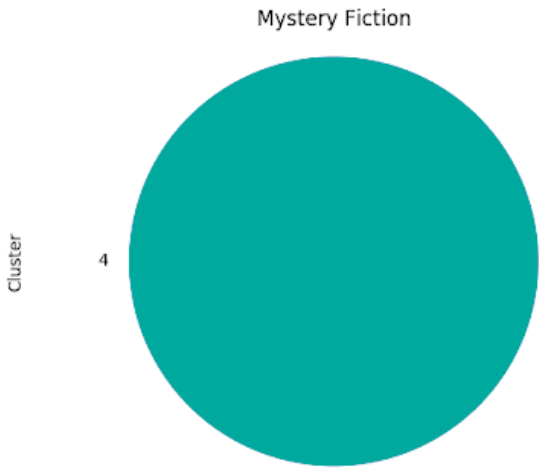
5.8 Horror



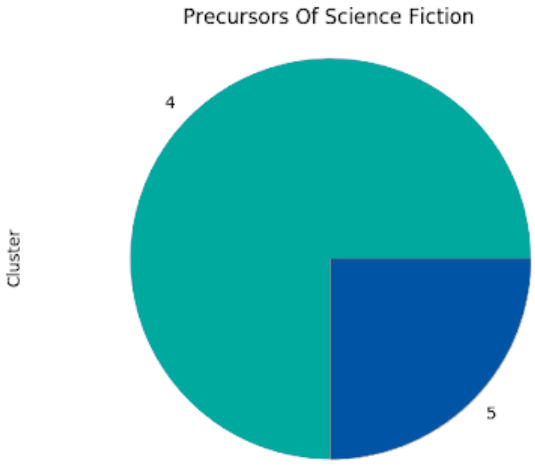
5.9 Humor



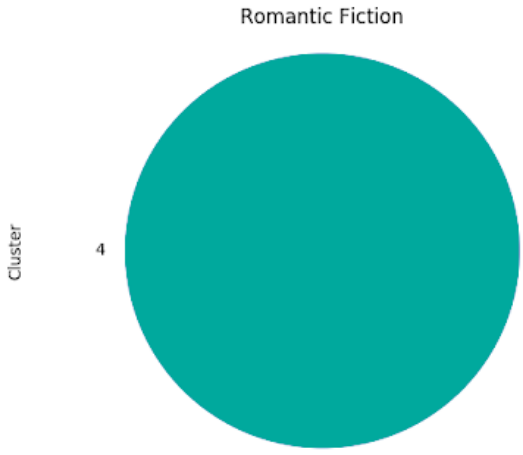
5.10 Mystery Fiction



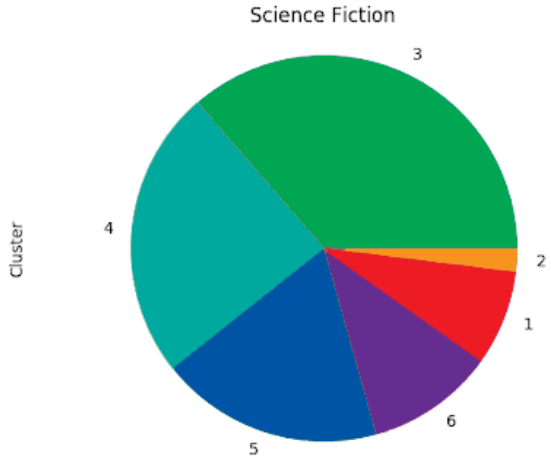
5.11 Precursors of Science Fiction



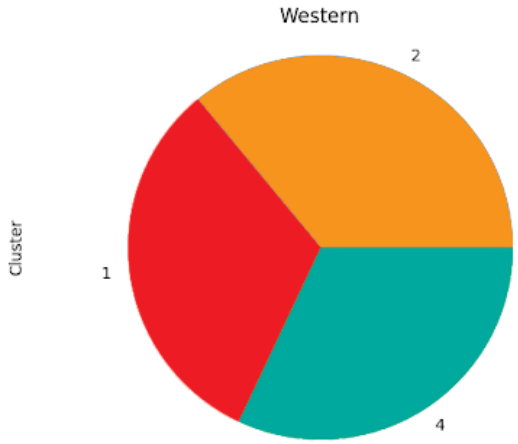
5.12 Romantic Fiction



5.13 Science Fiction



5.14 Western



Unfortunately, the results for this subsection are not as clear as the word clustering sub section. While for some sections, like the above one, the similarities are intuitive, the similarities could also be a sign that the clustering is catching a large amount of texts regardless of context. This is likely due to the noise caused by common sentence structures used regardless of text theme.

Despite the noise and thus uninteresting results of some clusters, there are other results that are intuitive and interestingly mirror previous results. While this could be coincidental results, it is a basis towards understanding the unseen similarities between texts, and an interesting exploration into what leads to fictional text categorization.

Conclusion

Overall, this project used Natural Language Processing practices applied to long-form fictional texts with the purpose of exploration into the building blocks of literature, both implicit and explicit.

In Part One, an original exploration into sentiment and the parts of speech in a text was done using dispersion plots on random selections of novels. For the parts of speech, it was shown that some categories of novels may have different parts of speech used as others. For example, shown in this paper was an example of a Detective Fiction novel in which there were a larger proportion of nouns than verbs, which not only make intuitive sense, but also showed potential that using parts of speech may be useful in categorizing novels. For sentiment, it was shown that the sentiment of a novel may change throughout a novel, for example, a fantasy novel in which the climax of the story included more negative words. Additionally, some categories may generally be more negative than others, for example, overall a Comedy or Children's Fiction Novel may be more positive overall. Again, this made intuitive sense and showed potential that sentiment could be used to categorize novels, and furthermore could be an interesting shadow of what is happening in a story as a whole.

These results were then built upon in Part two, where texts were clustered and categorized using either parts of speech or simply the cleaned-up words. The sentiment analysis in part one was built upon in the word clustering section. It was shown that often,

clusters seemed to not only rely on common words, like ships and other nautical words, but also on a general sentiment. It was shown that some categories were filled with generally more positive or lighthearted categories like Comedy and Science Fiction, while other categories had more negative categories like Science Fiction or horror. The part of speech section in Part One was directly built upon in Part Two. These results were much less clear, but potential was shown, not only in some categories that have less action being grouped together, or parallels to the word categorization section.

Overall, while the clustering and categorization did not have definitive results, parallels to both the categories given by Project Gutenberg, as well as the categories described by Christopher Booker were shown in the results. The results here support the hypothesis that there are underlying structures to fiction that naturally break stories into categories, but future work is necessary to fine tune these results and more clearly understand what the generated categories may describe.

Future Work

The largest barriers to this project were a lack of data and limitations in the data given. In the future, I would like to improve the results by both increasing the data size to include more works of fiction. Including other fictional works as well as newer pieces of work would greatly improve the accuracy of the results. By increasing the number of books available, the training set would be able to more interestingly cluster data and be able to move past some of the noise currently captured in the results. Training and testing on data

including movie scripts, short stories, modern novels, etc. would give a much more well-rounded data set and thus may give more interesting results, though it may potentially create an unbridgeable genre gap, that would need to be accounted for. Furthermore, having a set of books that were formatted more consistently would further improve the training set. This consistent formatting would also help with the issue of word meanings and sentiments changing throughout time. A potential improvement to this could be only sources from a specific era, rather than all eras.

Another continuation of this project would be to categorize the data in a supervised manner rather than the unsupervised manner used in this project. With more resources, being able to manually categorize all works of fiction the training of the data could include the categorization. This would allow for insights into what structurally allows us to categorize fiction in this way. While this is a much different project altogether, it is a related topic that would be an interesting continuation. This would help with vague categories like children's fiction, adventure, etc. which may be a wide variety of story types, as well as categories like detective fiction, crime fiction, and mystery fiction, which could be very similar in some cases.

A further continuation from that, would be the inclusion of a wider variety of tropes. Some examples may be plot devices like Chekov's gun, McGuffins, etc. that are commonly seen in fiction in novels, movies, TV shows, and more. Being able to identify tropes to this specificity would lead to a multitude of improvements in both the fields of literature and

natural language processing and would lead to many opportunities to be used and built upon.

An eventual, larger scale continuation building from the ability to identify a wide variety of tropes would be a tool for writers being able to make more interesting, less predictable works. Another potential use would be using the trained data to generate plots using the tropes. This would be interesting in generating stories for novels, television, or even procedurally generated plots for video games.

All in all, while this project had its own barriers, it is a stepping off point to be built upon and as interesting literary exploration. Further tuning of data, and training among other improvements would lead to a useful tool for writers of all varieties. I fully intend to build off of this project in order to flesh out the project to further explore the complexities of fictional literature and the common ground between works.

Bibliography

Ashe, Laura. “The Invention of Fiction.” *History Today*, 2018, www.historytoday.com/miscellanies/invention-fiction.

Booker, Christopher. *SEVEN BASIC PLOTS: Why We Tell Stories*. BLOOMSBURY CONTINUUM, 2019.

Celtic English Academy. “Why Is Shakespeare Still Important Today?” *Celtic English Academy*, 27 Feb. 2017, www.celticenglish.co.uk/blog/why-is-shakespeare-still-important-today/.

Cho, Y. H., and K. J. Lee. “Automatic Affect Recognition Using Natural Language Processing Techniques and Manually Built Affect Lexicon.” *IEICE Transactions on Information and Systems*, E89-D, no. 12, 2006, pp. 2964–2971., doi:10.1093/ietisy/e89-d.12.2964.

Dell’Orletta, Felice, Montemagni, Simonetta, and Venturi, Giulia. “Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose", 2013

Doctorow, E. L. “Notes on the History of Fiction.” *The Atlantic*, Atlantic Media Company, 1 Aug. 2006, www.theatlantic.com/magazine/archive/2006/08/notes-on-the-history-of-fiction/305033/.

Dredze, Mark, Jansen, Aren, Coppersmith, Glen, and Church, Ken “NLP on spoken documents without ASR." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (*EMNLP ’10*). Association for Computational Linguistics, USA, 460–470. 2010

Flekova, Lucie, and Iryna Gurevych. “Personality Profiling of Fictional Characters Using Sense-Level Links between Lexical Resources.” *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, doi:10.18653/v1/d15-1208.

Nizar Grira, Michel Crucianu, Nozha Boujemaa “Unsupervised and Semi-supervised Clustering: A Brief Survey." 2005.

Maker, Richard. "Reader centered classification of adult fiction in public libraries." *Australasian Public Libraries and Information Services*, vol. 21, no. 4, 2008, p. 168+. *Gale Academic OneFile*

Mendoza, Melissa. “The Evolution of Storytelling.” *Reporter*, 2015, reporter.rit.edu/tech/evolution-storytelling.

Nltk. April 2020. www.nltk.org

“New Fast.ai Course: A Code-First Introduction to Natural Language Processing.” *New Fast.ai Course: A Code-First Introduction to Natural Language Processing ·*, www.fast.ai/2019/07/08/fastai-nlp/.

Pejtersen, and Annelise Mark. “Design of a Classification Scheme for Fiction Based on an Analysis of Actual User-Librarian Communication, and Use of the Scheme for Control of Librarian's Search Strategies.” *ERIC*, 31 July 1977, eric.ed.gov/?id=ED231370.

Project Gutenberg. April. 2020. www.gutenberg.org

Radford, Alec. “Improving Language Understanding with Unsupervised Learning.” OpenAI, OpenAI, 2 Mar. 2020, openai.com/blog/language-unsupervised/.

Scikit Learn. April 2020. https://scikit-learn.org/stable/modules/clustering.html

Vernitski, A. (2007). Developing an intertextuality-oriented fiction classification. *Journal of Librarianship and Information Science*, *39*(1), 41–52. https://doi.org/10.1177/0961000607074814

Yse, Diego Lopez. “Your Guide to Natural Language Processing (NLP).” Medium, Towards Data Science, 30 Apr. 2019, towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.