Spring 2020

Text Mining Contemporary Popular Fiction: Natural Language Processing-Derived Themes Across Over 1,000 New York Times Bestsellers and Genre Fiction Novels

Morgan Lundy

TEXT MINING CONTEMPORARY POPULAR FICTION: NATURAL LANGUAGE PROCESSING-DERIVED THEMES ACROSS OVER 1,000 *NEW YORK TIMES* BESTSELLERS AND GENRE FICTION NOVELS

by

Morgan Lundy

Bachelor of Arts
University of South Carolina, 2016

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Arts in

English

College of Arts and Sciences

University of South Carolina

2020

Accepted by:

Michael Gavin, Director of Thesis

Amir Karami, Reader

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ACKNOWLEDGEMENTS

This study was made possible by a generous research fellowship provided by the University of South Carolina's Center for Digital Humanities, and the invaluable technical and theoretical advice of my advisers in both English and Library and Information Science. Special thanks also to my family, friends and healthcare providers for their support throughout this research process, and to the study's qualitative second coder, Chris Opie.

ABSTRACT

This study endeavors to apply computational methods to a large dataset of popular fictional material, to see what topics emerge when viewed across genre lines and from a new, “machine” perspective. The dataset consists of 1,136 popular and commercially successful novels published between 2005 and 2016, including *New York Times* bestsellers and “genre fiction,” including science fiction, young adult, romance and mystery novels. Methods are discussed, including dataset preparation, LDA topic modeling and topic number optimization, qualitative topic interpretation, data analysis and visualization. The experiment was conducted in two parts, with the "document" or unit of analysis as each full novel, and then all of the sentences of every novel (over 9 million). 23 topics at the novel level and 66 at the sentence level were qualitatively interpreted, compared across genres and visualized. This study argues that computational tools can be generatively used to vastly broaden the scope of literary analysis, but results must still be interpreted through qualitative means. The novel may be quantitatively analyzed at both the level of the entire novel and the level of the sentence but analyzing at the level of the sentence offers more granular and interesting results. Topic modeling here identifies latent, ubiquitous topics that a human researcher may ignore or miss, re-centers research focus on the human body, its functions and the embodied nature of fiction, and was able to identify novel conventions such as linearity, characterization and settings and to distill many socially relevant topics including violence, surveillance and human institutions and activities. While topic modeling here reinforced some topical

expectations based on genre conventions and tropes, topics also appeared unexpectedly in other genres: helping re-imagine the popular fiction landscape outside of genre-based siloes. Statistical analysis of a fictional dataset offers a new, birds-eye view of the contemporary popular fictional landscape, but also has many limitations, many of which are discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 What is Topic Modeling, and What is it Doing in an English Thesis? Opportunities for Statistical Analysis in Literary Study

In a world of increasingly rapid publishing, both digitally and by traditional publishing houses, the flood of information is overwhelming to literary researchers, readers and library acquisitions staff alike (Cain, 2016). It is difficult to gain a sense of overall content or trends across such vast amounts of text, especially in a timely manner. Thankfully, information science methods have been developed that allow for reading "at a distance" (Moretti, 2015) with the assistance of computational processes and statistical software environments. One method that aids in distilling the content of vast amounts of textual material is topic modeling, a form of natural language processing and statistical analysis that gleans "topics" or related word groups by looking at the probability that words occur in the same documents (Neatrour et al., 2018).

This study will focus on popular fictional material, specifically a dataset of 1,136 *New York Times* bestsellers and popular genre fiction published between 2005 and 2016. Content analysis of thousands of recently published novels here proves significant in identifying themes that resonate with current social and political issues, across multiple genres; examples include topics like gun violence, the environment, health-care related content, digital technology, religion and mass media and information. While many literary researchers are studying these socially relevant themes in individual texts, a

natural language processing approach through topic modeling will make it possible to distill patterns across many texts at once. This study aims to utilize text mining capabilities to supply an overall picture of recent, popular literature by identifying the scope and prevalence of topics across multiple genres, authors and years. It is difficult for any single literary researcher to read thousands of novels at a normal reading pace, especially in a timely manner or when working with contemporary literature that is still being published. It is also difficult to recognize trends and commonalities across vast amounts of text and across existing research, often siloed by genres, authors and years. In this study, computational tools are used to aid in literary study by scaling up the scope of analysis.

Within the nebulous and developing field of digital humanities, topic modeling falls within the larger umbrella of "distant reading," or forms of analysis that provide distance and abstraction to view texts, genres and periods in new light. Distant reading methods include network analysis, stylometry and forms of natural language processing, like topic modeling. Distant reading, the brainchild of Franco Moretti, has undergone a much-needed transformation recently, with scholars calling for broader representation of diverse literary works, more focus on context to lessen the method's reductive tendency and new methods of analysis that critique power rather than reinforcing it; distant reading "after Moretti" now offers a more compelling alternative way to approach literary material (Klein, 2018).

However, taking into consideration criticisms of distant reading's claims to absolute, quantifiable proof, it is more useful to approach this computationally derived survey of the literary landscape not as an undeniable, mathematically true new reading,

but as an additional birds-eye outlook to complement existing and continuing literary scholarship. This study is exploratory in nature – motivated by broad questions with an element of the "playful intervention of the machine" (Liu, 2012, p. 21). The strength of quantitative analysis of fiction is that it offers a new perspective: that of the machine. This new perspective is fruitful in that computational analysis here identifies perhaps-expected topics, surprising topics, and topics so ubiquitous they are forgotten within research and are near-invisible to the human reader or researcher. By using an unsupervised model, this study takes a slightly posthuman vein; here the computer is elevated from a place as just a tool serving a researcher's goals, to that of an active participant in a new interpretative paradigm where the computer and human researcher share in "discovery and analysis" (Liu, 2012, p. 22). Digital humanities is not human interpretation or computational processes alone, but a new way of thinking forged in their union. This study aims to emulate this theoretical stance by utilizing an unsupervised model and a mixed-methods approach that includes qualitative interpretation. The following sections on novel conventions, ubiquitous topics and embodiment may be of particular interest to the literary researcher.

Beyond the scope of literary research, the findings and process of this study may also be of interest to information science researchers. The dataset has been curated to focus on popular—that is, commercially successful—fiction; fiction that is not always the focus of literary research. Instead, this study aims to distill topics in the media that is most-read or at least most-purchased: bestsellers, science fiction, romance, young adult and mystery novels. Many people encounter these best-selling books across the U.S. and globally; the texts that make up this dataset have collective sales numbers in the millions.

The effect of reading fictional material is a nebulous one, but studies do connect the consumption of fictional media with readers' perspectives on similar scenarios in daily life (Bishop, 1990; Adkins, 2013). These effects, even if nebulous, make studying often-under researched fictional material salient in that they will offer a cultural "snapshot" of what topics—one could argue, what information—contemporary readers are consuming. In viewing fiction as information, information science researchers may be particularly interested in the sections on sociocultural topics, including human institutions, actions and activities, and violence.

Discussions of the limitations of natural language processing approaches may also be of interest, as well as the implications of qualitative interpretation, viewing fiction as information, and the assumptions within computational tools. Technology, of course, is not "self-designing" nor born "innocent" (Haraway, 2016, p. 11, 65), and while this study does not analyze the topic modeling approach itself, it does address the limitations and human-created nature of the quantitative methods it utilizes. The study, again, does not make claims to mathematically "correct" readings of the literary landscape. Instead, this study provides a "snapshot" from a different angle by quantitatively distilling topics within a dataset of 1,136 novels at the level of the entire novel, as well as at the more granular level of all of the novels' sentences (over 9 million), then qualitatively interpreting topics and comparing the topics across genres and years.

1.2 *Research Questions and Hypotheses*

- Research Question 1: What 60 topics appear most frequently in popular novels, when topic modeling is applied to a dataset of 1,136 commercially successful

novels published between 2005 and 2016, where the "document" or unit of analysis is the entire novel?

- Hypothesis 1: The novel-level topics will be familiar literary themes such as love, death, adventure, friendship, health or family.
- Research Question 2: What latent topics might natural language processing identify that are not immediately clear to the human reader, especially at the sentence-level, where the unit of analysis is all of the sentences (over 9 million) found in the same 1,136 novels?
- Hypothesis 2: The sentence-level topics may have unexpected results, which are more visible to machine processing than to the human reader who may have assumptions about reading and literary content.
- Research Question 3: Do the 60 novel-level topics appear equally across genre lines? What topics are more frequent in which genres?
- Hypothesis 3: The novel-level topics will vary based on genre and will most likely align with genre conventions (for example, "intimacy" in romance, "space travel" in science fiction).
- Research Question 4: Do the 80 sentence-level topics appear equally across genre lines? What topics are more frequent in which genres?
- Hypothesis 4: The sentence-level topics will also vary based on genre, most likely also aligning with genre expectations, but perhaps with unexpected distribution across genres as well (for example, a high weight of "intimacy" in science fiction, or more surprisingly, a high weight of "space travel" in romance).

- Research Question 5: Do the 60 novel-level topics change or fluctuate significantly over time, particularly between 2013 and 2016 where the dataset is most heavily weighted?
- Hypothesis: Some novel-level topics may change over time, following trends in what is selling and popular (for example, more novels with vampires following the success of *Twilight*, or dystopias after the success of *The Hunger Games*).

CHAPTER 2

THE MECHANICS OF READING AT SCALE

2.1 *Topic Modeling Applications in Information Science and Digital Humanities*

The text mining technique applied in this study is topic modeling. The specific topical and statistical model of interest in this study is Latent Dirichlet allocation (LDA), developed by David Blei, Andrew Ng, and Michael Jordan in 2003. The model was created in order to distill "short descriptions" of texts that allow for quick and efficient processing of large amounts of text, while preserving "the statistical relationships," useful for "classification, novelty detection, summarization, and similarity and relevance judgements" (Blei et al., 2010, p. 993). An assumption of the LDA model is that texts are all modellable, as they are a finite mix of words where all discrete words have a probability of occurring with other words, when compared across documents. This assumption illustrates the underlying idea that if words appear together, they are necessarily related topically—an idea which has been disputed by humanities researchers such as Lisa Rhody, in terms of figurative language (Rhody, 2012). However, the texts included in this dataset are less experimental or poetic than those studied by Rhody and assessing the texts at the sentence-level especially, where "topics" are much more granular, allows for more conclusive word-topic relationships.

Topic modeling has been applied widely across many disciplines. Studies distilling topics and themes from vast amounts have text have included assessing disaster damage through social media data (Resch et al., 2018), themes in Consumer Financial

Protection Bureau (CFPB) complaints (Bastani et al., 2019), the prevalence of antisemitism in far-right media (Barna & Knap, 2019), topics that appear across medical records (Cohen et al., 2014), commonalities among sexual assault testimonies shared online (Karami et al., 2019), media response to terrorist threats (Bonilla & Grimmer, 2013), underlying topics in U.S. national strategy reports (Mohr et al., 2013), and to distill overarching information from population studies (Marshall, 2013).

Within the field of literary criticism, topic modeling has been applied less extensively to literature, but previous projects in digital humanities include topic modeling genre in French classical and enlightenment drama (Schöch, 2019), public discourse around Dutch nationalism (Bos & Van Den, 2019), classical texts and East German dossiers (Jähnichen et al., 2019), and reported dreams and personal narratives (Hendrickx & Onrust, 2019). More similarly to this study, LDA models were applied to a corpus of 19th-century British, American, and Irish works of fiction by Matthew L. Jockers and David Mimno in 2013. Jockers and Mimno were able to distill hundreds of topics (which they refer to as "themes") and draw correlations between the themes identified in the works and external factors like "author gender, author nationality and date of publication" (Jockers & Mimno, 2013, p. 750).

2.2 *Theoretical Framework*

Several cross-disciplinary theories underpin the concept that statistical modeling can be generatively applied to language to identify patterns and related word clusters. Natural language processing, the intersection of linguistics and computer science which focuses on programming to analyze vast amounts of human-generated text, has a long theoretical and technical history. Reaching all the way back to Alan Turing's pioneering

1950 paper “Computing Machinery and Intelligence,” it is helpful to keep in mind the theoretical stance that research questions about artificial intelligence should not be, as Turing argues, “can machines think?”—or in this case, “read” like a human—but instead, assessing if a computer or program can return results convincingly like a person (Turing, 1950, p. 433). Beginning in the 1980s with the advent of machine learning, research questions have shifted further to whether or not a program can return the expected results, regardless of the function occurring between the input and output. While previous linguistics projects had focused on applying existing language rules and grammars to texts, machine learning allows computer programs to instead identify patterns on their own, outside of human-defined expectations of language and meaning.

The theory behind the LDA model is that topics can be identified based on the probability that words appear in the same documents within a corpus. LDA has the strength of adaptability, as its sole focus on computing probabilities and weights rather than comparing with a set word list, which allows it to process texts that include slang, unfamiliar language or fictional words and names. It also has the strength of being extremely accurate in its ability to produce related word clusters; LDA is one of the most popular models for use in topic modeling, with 28,342 citations in Google Scholar today (April 10, 2020). The use of LDA here, as an extension of the theories of computer imitation, machine learning and statistical models more generally in natural language processing, illustrates how computers can accurately and quickly identify patterns in datasets, even complex textual data like novels.

2.3 *Research Design: A Mixed Methods Approach*

This content analysis study is a form of non-reactive research and will utilize a mixed methods approach. The study is inherently quantitative; the computational methods used statistically assess the words in the dataset and output a set of numerical weights. Quantitative data is also reported from the study, such as the average weight of topics across all novels within a certain genre. However, the inherent fluidity and complexity of language and linguistic meaning requires a qualitative, or in this case human, methodology as well. While the statistical model operates quantitatively—assessing weights of certain words that are semantically related— the resulting output of a series of statistically-related words is meaningless without qualitative interpretation. The quantitative model results in word lists, such as "gene, DNA, genetic, heredity, family, biology." This word list is an objective, quantitatively-produced finding—however it is bulky and difficult to understand without a succinct qualitative interpretation and label, such as "genetics." Qualitative methods, including triangulated topic interpretation and close reading, are then used. A mixed methods methodology is necessary to engage with the original research problem—the size of the dataset and, more broadly, the amount of material being published—in a way that is meaningful, human-interpreted and deep as well as wide.

Topic modeling is applied in this study through the use of MALLET, an open source Java-based package created by the University of Massachusetts-Amherst for topic modeling with LDA (McCallum 2002). The following sections break down the research framework into dataset overview and dataset preparation methods (2.4), topic number optimization and topic modeling methods (2.5), qualitative topic interpretation methods

(2.6) and genre comparison, trend analysis and data visualization methods (2.7). See the experiment guide in Appendix A for more specific instructions for reproducibility.

2.4 *Dataset Overview and Dataset Preparation Methods*

This study is a secondary analysis. The core textual data that is the basis for the project was generously made available by Dr. Andrew Piper at McGill University and Dr. Michael Gavin at the University of South Carolina. The dataset is not publicly available due to copyright restrictions, as the 1,136 novels were all published in the past twenty years. The dataset was originally used by Piper to asses "fictionality" or how fictional and nonfictional texts differ (Piper, 2016), as well as to assess "cultural capital" or how genre fiction differs from literary texts (Piper & Portelance, 2016). Piper's approach is semantic in nature, investigating differences based on features like frequency of parts of speech used – essentially asking *how* texts are written. The present study differs in that it is content analysis based, instead asking *what* texts are topically about.

The dataset has also been significantly reduced in size – the goals of the current study are to investigate the topics appearing in popular fiction with mass readership, so all prize winners and novels reviewed by the *New York Times* were removed. The present study sought to avoid the dichotomy between "high" literature and popular or "low/middle-brow" fiction—and is more interested in distilling the topics prevalent in popular culture, in the media readers most consume, without an assessment of literary quality.

The dataset used here is made up of novels with high commercial success, both in terms of sheer number of weeks on the *New York Times* bestseller list, and a more diverse array of genres based on different types of economically successful writing: as they argue

in the original study, these texts are "what we might call genre fiction" (Piper & Portelance 2016, p. 3). The following chart displays the breakdown of the dataset by genre: young adult, science fiction, romance, mystery and bestsellers (Table 2.1). Bestsellers, of course, can also be divided into genres, but they are not considered "genre fiction." The novels are evenly distributed across the genres (with a range of 211-248), but the sentences are much more variable due to the generally shorter length of novels within certain genres. For example, romance novels tend to be much shorter in length than science fiction novels, leading the sentence-level dataset to have a higher percentage of science fiction sentences.

Table 2.1 Breakdown of Dataset by Number of Novels and Sentences in each Genre

Genre	Number of Novels	Number of Sentences
Young Adult	231	1,521,426
Science Fiction	211	2,032,989
Romance	212	1,486,554
Mystery	234	1,884,095
Bestsellers	248	2,371,103
Total	1,136	9,296,167

To offer an overview of the dataset, a word cloud was created (n=50), to judge the books by their covers—or at least by titles—in Figure 2.1. In this rudimentary frequency analysis, where higher-frequency words appear larger and stop words have been removed, the titles suggest that "death," "darkness," "night" and "murder" may be topics in the dataset. However, "novel" is more likely a title convention used in series (for example, "an Artemis Fowl Novel"), and titles frequently lean towards salaciousness; a novel about "murder" may not mention the many other topics contained in the text in its publicity-focused title. Topic modeling offers a route to discovering the many less title-

worthy topics that appear in these novels. An additional generative future route of study could be to compare the title-words and topics distilled by topic modeling with the Library of Congress subject headings assigned to each novel.

Figure 2.1 Frequency Analysis Word Cloud of the Words in the Novel Titles

Due to copyright restrictions, the dataset was received in the form of sentence-word matrixes in RDA (R) format. These matrices were translated back into text file formats using an R script, where the sentences now appear line by line (sentence by sentence), but with the word-order "scrambled." This way, statistical, quantitative methods could be used without copyright infringement. To assess the novel-level topics, these individual-novel "scrambled" text files were imported to Mallet using the import-directory function and converted into Mallet format. To assess the sentence-level topics, more dataset preparation was required, including adding headers (novel title, genre) and separating the sentences into multiple Excel files to accommodate the large 9 million-line size, before converting the entire set of 9 million to one text file and using the Mallet

import-file command to consider each line as an individual document, and convert to Mallet format. During this process, additional data cleaning was also preformed including removing unnecessary content from the dataset: words like “chapter,” “back” and “prologue” as well as copyright information or unrelated text at the end of novels.

2.5 *Topic Number Optimization and Topic Modeling Methods*

After test iterations, additional stop words were also added using the Mallet extra-stop-words function. Topic modeling is exceptionally good at identifying character relationships within novels, so character names that appeared in multiple novels together in series were identified as topics. While character relationships are interesting and certainly an important part of novels, identifying how many novels have a character named “Robert” is not exceptionally helpful. Using a convenient online database of baby names, an additional stop words text file was created and used to decrease the number of character names appearing in topic lists.

Once the files were converted to Mallet format, topic modeling was applied through several iterations with increasing numbers of topics to optimize the topic number for each experiment (both novel-level and sentence-level). While topic optimization can be done computationally, notably with the *ldatuning* package in R or through visualizing an elbow curve, trial and error in goldilocks fashion was used here in order to obtain a topic set that contained both semantically straightforward and seemingly unintelligible topics. As Ted Underwood explores in his blog on topic modeling for literary analysis, there is more to learn for the literary researcher in “ambiguous topics” than in readily understandable topic lists (Underwood, 2012). What he calls “problematic topics”—those which are seemingly unintelligible or not immediately clearly semantically linked —can

point to what he calls a "discourse" or a "rhetoric," or something researchers do not yet understand (2012). These topics were also generative moments for the application of close reading methods by looking to the most-related documents, to identify abstract topics (like *perception & communication*) as well as straightforward topics (like *cellphone use*).

For the novel-level experiment, 20 topics proved far too vague and broad with topics that mostly just separated the novels by genre: for example, the topic *space travel*. 40 topics proved more interesting and slightly more granular, with the too-coarse topic *bodies* in the 20-topic list now split meaningfully into two separate concepts: *body parts* and *sex*. However, the topics were still too vague, or frequently conglomerates of multiple ideas, like the topic *moments*. At 60 topics, the results were more granular, and there was also a mix of expected topics (like *family*), abstract topics (like *paranormal*), and unexpectedly ubiquitous topics (like *cars*). At 80 topics, many of the topics appearing in the 60-topic set were again present, but with a higher percentage of completely unintelligible topics—only 28 out of 80 were useable as the topics began to separate out specific series. A 100 topics setting was also tested and was far too specific with frequent topic-splitting that would have skewed results. 60 was chosen as the optimal number of topics for the novel-level experiment.

The same process was implemented for the sentence-level experiment. 20 topics was again far too coarse—with the words "police," "president," and "judge" all appearing in the same conglomerate topic—while 100 topics was far too specific, with more than one topic for the exact same body parts. 80 topics was chosen over 60 topics, again to

include more abstract topics like *joy* and *judgement* that did not appear in the slightly more straightforward 60-topic list.

Using the Mallet implementation of LDA topic modeling, each the dataset was modeled with the entire novel as the unit of analysis, and then the sentence as the unit of analysis. Further research could pursue modeling novels with the paragraph as the unit of analysis; the format of the dataset made this impossible within the scope of this study. The output settings used were topic-keys, which records the topic wordlists for qualitative interpretation, and topics-composition, which records the weight of each topic within each document – within an entire novel or a specific sentence depending on the experiment. See the experiment guide in Appendix A for specific commands used.

2.6 *Qualitative Topic Interpretation Methods*

Qualitative methods were then used to assess, interpret and label topics from the wordlists. In order to make these often-intuitive assessments less subjective, each wordlist was qualitatively interpreted by two coders. The top 10 highest-weighted words for each topic, in descending order, were the basis for the qualitative assessment. Each coder independently assessed the topic wordlists and determined the relationship between the words in each list, then submitted a brief label for the topic. The coders also analyzed the 3 most-related documents (novels or specific sentences) by applying close reading methods for sentences or assessing the plot of novels from summaries, especially for labeling "problematic" or seemingly unintelligible topics (Underwood, 2012). The most-related sentences and novels were determined using the topic-composition files.

Including synonyms, the coders had 83% agreement for the novel-level topic analysis, and 90% agreement for the sentence-level topic analysis. For disagreements, the

two coders employed another round of annotation. For examples of qualitative interpretation and topic labeling, see an example of an abstract topic from the sentence-level experiment with most-related sentences (Table 2.2) and a straightforward topic from the novel-level experiment with most-related novels (Table 2.3).

Table 2.2 Example of an Abstract Topic Label Interpretation

Topic Label	Topic Word List
Goodness, Badness & Judgement	knew wanted thought wrong happened hurt bad good people happen telling felt meant part needed understand
Most Related Sentences:	
(1) “Something’s really wrong and I want to talk about it, but I think talking to you would be the world’s worst idea because you always overreact, make it about you, or, in this case, wouldn’t believe me anyway.” - *Subway Love* by Nora Raleigh Baskin (2) “There were a lot of people in this world who thought they were better than everyone else, but I couldn’t ever remember hearing anyone actually admit they felt that way.” – *Glimmerglass* by Jenna Black (3) “Surely it was good, telling someone who understood that strange and terrible world.” – *The Iron Witch* by Karen Mahoney	

Table 2.3 Example of a Straightforward Topic Label Interpretation

Topic Label	Word List
American Military	man men president sir military war people government general army team security minutes american colonel united russian officer intelligence country
Most Related Novels:	
Dead or Alive by Tom Clancy *Cross Fire* by James Patterson *A Soldier’s Duty* by Jean Johnson	

All topics are explored more fully in the results sections. To see a full list of all topic wordlists and topic labels see Appendix B.

2.7 *Genre Comparison, Trend Analysis and Data Visualization Methods*

To compare across genres, the weight of all novels or sentences within a genre were averaged and compared. Trends were assessed by averaging and comparing the topic weights per year. To aid in comparison and analysis, topic weights were visualized in bar, column and line graphs, as well as pie charts.

CHAPTER 3

EXPERIMENT RESULTS, 1,136 NOVELS

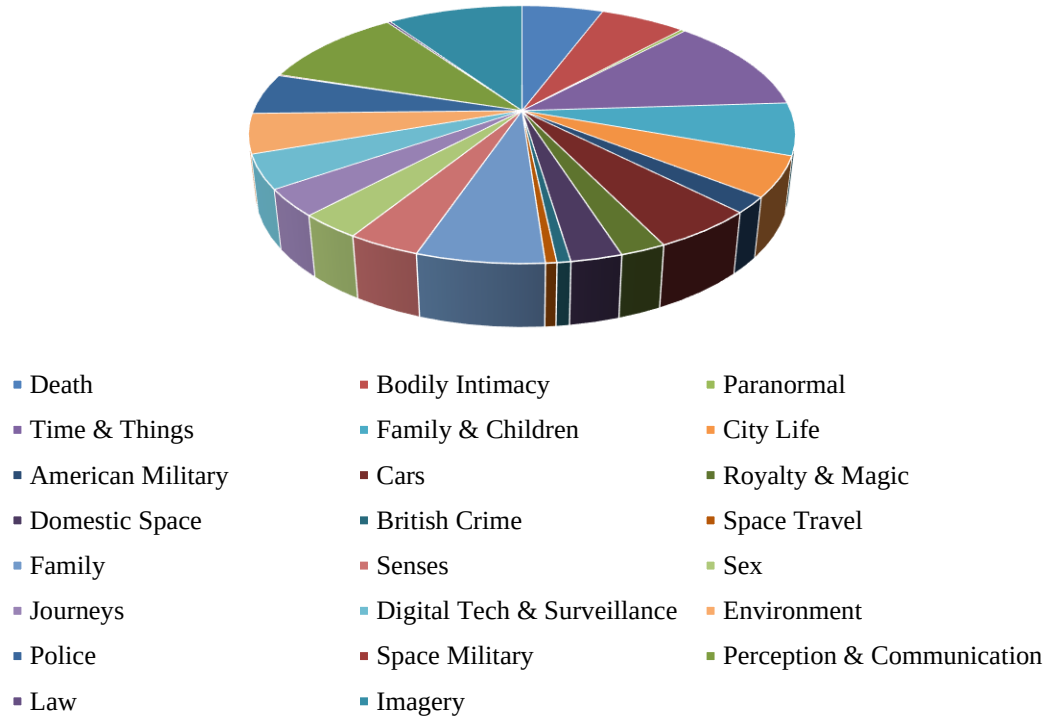


Figure 3.1 Pie Chart of Novel-Level Topic Distribution Across the Dataset

3.1 *Popular Topics and Most-Related Novels*

Of the 60 topic wordlists generated at the novel-level, 23 topics were selected as semantically meaningful. While the sentence-level experiment was far more successful at generating meaningful topics and results, the novel-level topic modeling still offers insight into high-level themes within novels as a whole. Aligning with the original hypothesis, many of the following novel-level topics are familiar literary themes—such

as *death* and *family*, though some familiar topics are slightly more literal, such as *bodily intimacy* rather than "love" or *journeys* rather than "adventure." There are also more surprising topics, such as *digital technology & surveillance, cars* and the *paranormal*. Generally, the topics are still quite vague and at times genre-specific, which is to be expected when assessing an entire novel as one large "bag of words" document.

In order to provide a sense of the topics and the dataset itself, the following chart (Table 3.1) details each novel-level topic and the corresponding 3 most related novels.

Table 3.1 Novel-Level Topics and Their 3 Most-Related Novels

Topic	**Most Related Novels**
Death	*My Sister's Grave* by Robert Dugoni (2014, MY) *I See You* by Clare Mackintosh (2016, MY) *Dreaming Spies* by Laurie R. King (2015, MY)
Bodily Intimacy	*Robopocalypse* by Daniel H. Wilson (2011, SF) *The Lincoln Lawyer* by Michael Connelly (2005, BS) *This Was A Man* by Jeffrey Archer (2016, BS)
Paranormal	*Fins Are Forever* by Tera Lynn Childs (2011, YA) *Horrorstor* by Grady Hendrix (2014, MY) *Tall, Dark and Panther* by Milly Taiden (2015, ROM)
Time & Things	*The Daring Exploits of a Runaway Heiress* by Victoria Alexander (2015, ROM) *Pride and Prejudice and Zombies* by Seth Graham-Smith (2009, BS) *Murder at the Brightwell* by Ashley Weaver (2014, MY)
Family & Children	*The Last Bookaneer* by Matthew Pearl (2015, MY) *The Last Girl* by Joe Hart (2016, MY) *The Last Man* by Vince Flynn (2012, BS)
City Life	*The Scorch Trials* by James Dashner (2010, YA) *Sea of Shadows* by Kelley Armstrong (2014, YA) *Insurgent* by Veronica Roth (2012, YA)
American Military	*Dead or Alive* by Tom Clancy (2010, BS) *Cross Fire* by James Patterson (2010, BS) *A Soldier's Duty* by Jean Johnson (2011, SF)

Cars	*Reconstructing Amelia* by Kimberly McCreight (2013, MY) *The Agreement* by Se. E. Lund (2014, ROM) *Escape Clause* by John Sandford (2016, BS)
Royalty & Magic	*Christmas Sweater* by Glenn Beck (2008, BS) *Something Like Fate* by Susane Colasanti (2010, YA) *The Iron Queen* by Julie Kagawa (2011, YA)
Domestic Space	*Leah* by R.J. Lewis (2015, ROM) *CHANCE* by Deborah Bladon (2015, ROM) *Carter* by R.J. Lewis (2015, ROM)
Crime	*A Wanted Man* by Lee Child (2012, BS) *Leaving Time* by Jodi Picoult (2014, MY) *The Appeal* by John Grisham (2008, BS)
Space Travel	*Calamity* by Brandon Sanderson (2016, SF) *Tankborn* by Karen Sandler (2011, SF) *Rapture* by Kameron Hurley (2012, SF)
Family	*Ashen Winter* by Mike Mullin (2012, SF) *Exo* by Steven Gould (2014, SF) *Reached* by Ally Condie (2012, YA)
Senses	*Bringing Down Sam* by Leslie Kelly (2012, ROM) *Shooting for the Stars* by Sarina Bowen (2015, ROM) *Unlucky 13* by James Patterson (2015, MY)
Sex	*Taming the Highlander* by May McGoldrick (2016, ROM) *The Ruby Circle* by Richelle Mead (2015, ROM) *Known: A Bone Secrets Novel* by Kendra Eliott (2016, ROM)
Journeys	*The One Safe Place* by Tania Unsworth (2014, YA) *The Note* by Teresa Mummert (2013, ROM) *This Dark Road to Mercy* by Wiley Cash (2014, MY)
Digital Technology & Surveillance	*Behind Closed Doors* by B. A. Paris (2016, MY) *Til Death* by Bella Jewel (2015, ROM) *Carter Reed* by Tijan Meyer (2013, ROM)
Environment	*Sawyer* by Nicole Edwards (2015, ROM) *Never Too Late* by Micalea Smeltzer (2015, ROM) *Chambers and The Shadow* by James Runcie (2012, MY)
Police	*Zone One* by Colson Whitehead (2011, SF) *Robogenesis* by Daniel H. Wilson (2014, SF) *Smuggler's Moon* by Cynthia Wright (2014, ROM)
Space Military	*A Night Without Stars* by Peter F. Hamilton (2016, SF) *11.22.63* by Stephen King (2011, BS) *Existence* by David Brin (2012, SF)

Perception & Communication	*More Than Him* by Jay McLean (2014, ROM) *Where the Road Takes Me* by Jay McLean (2015, ROM) *The Note* by Teresa Mummert (2013, ROM)
Law	*Captive of Kadar* by Trish Morey (2015, ROM) *Murder at Merisham Lodge* by Celina Grace (2016, MY) *Kill Someone* by Luke Smitherd (2016, MY)
Imagery	*Elizabeth is Missing* by Emma Healey (2014, MY) *The Girls* by Emma Cline (2016, BS) *Defiance* by Lili St. Crow (2010, YA)

Looking just at the most-related novels, there are already a few surprises. For example, none of the most-related novels in the *bodily intimacy* topic are from the romance genre, while two of the three most-related novels in *digital technology & surveillance* are from this genre. Upon further investigation, these three novels insidiously involve the control and surveillance of a romantic partner. Across many of the topics, a pattern emerges that texts with the highest weight of a given topic have plots that involve the *absence* of that topic. The concept of ubiquity is explored more fully in the next chapter, but hints are visible here; at the novel-level ubiquitous aspects of life do not register, but their absence receives enough "coverage" in the texts to be detected by natural language processing. For example, the romance novel *Captive of Kadar* has the highest weight in the *law* topic, and is about a woman's illegal imprisonment, or the absence of law. Similarly in the *police* topic, the novels are about the absence of the police in dystopian settings; in *Zone One* the "sweepers" or survivors patrol New York City for zombies. Similarly in the *city life* topic, characters in these young adult texts inhabit desolate, crumbling urban landscapes, like the ruins of Chicago in *Insurgent*; while the city setting may be ubiquitous enough to not merit high numbers of descriptions in most novels, the absence or alteration of a familiar topic calls for more frequent and statistically significant amounts of description and word usage.

However, for the most part the topics fall along genre conventions in familiar ways: *American military* is dominated by masculine blockbuster adventure series, *space travel* appears in science fiction, *crime* and *death* in mystery, and romance novels take the highest-weighted spots in *sex*, the *domestic space* and *perception & communication*, perhaps due to plot-level focuses on communication as courtship. Some topics are more dispersed across genres, in that they are a part of all novels: *journeys*, for example, or *imagery*. The *environment* weights highly in texts with rural settings across genres, but with a focus again on difference and alteration—many of these characters are moving "back" to rural areas from urban ones, warranting more description. So, while the topics themselves were familiar and many fall along genre lines, the way the topics manifest in texts quantitatively is at times unexpected.

3.2 *Topics Compared Across Genres*

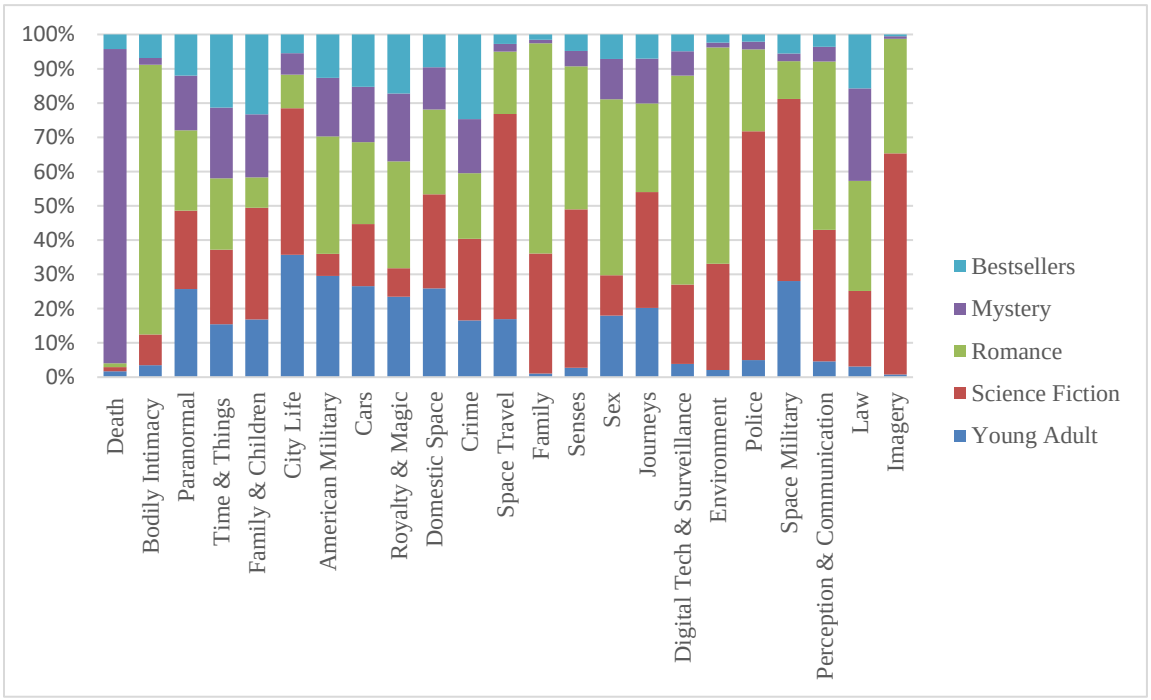
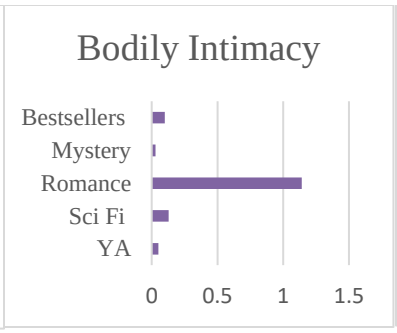


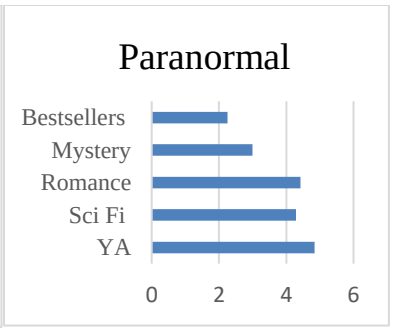
Figure 3.2 Average Novel-Level Topic Weights by Genre in Stacked Column Graph

Turning to look at the weights of the topics across all texts within each genre—rather than just the most-related three novels—there are again expected results and a few surprises. A stacked column visualization (Figure 3.2) makes clear the most dramatic genre-topic relationships and offers a quick snapshot of the distribution. *Death* is dominated aggressively by mystery. This however does not mean no characters die in young adult or science fiction, instead it may point to how death is often euphemized or glossed over in young adult fiction, or the clean Stormtrooper-blaster kinds of deaths that appear in science fiction. Topic modeling is able to distill the weights of how much a topic is described and to what quantitative extent, but not necessarily when topics occur in metaphorized or implied ways. While the most-highly weighted texts in bodily intimacy were from the science fiction, overall romance is quite dominant. Science fiction unsurprisingly takes up much real estate in space travel, space military and police – but also corners the market on imagery. This does not mean other genres lack imagery—indeed the topic list *imagery* is just a certain type of imagery—but also points to the frequent lavish descriptions of alien planets. Again, in statistical analysis, quantity is key.

To view more granular differences, in Figure 3.3 all topics are visualized in individual bar graphs, illustrating the average weight of that topic in each genre. While the *paranormal* is highly weighted across the dataset, it is particularly high in romance, science fiction and young adult—unsurprisingly given the fantastical nature of science fiction, and the trends in recent years for paranormal activity in young adult and romance (particularly vampires and werewolves). Vague topics like *time & things* and ubiquitous topics like *cars, the domestic space, law* and *crime* have fairly even weights, though the


Death
Bestsellers
Mystery
Romance
Sci Fi
YA
0
2
4
6


Bodily Intimacy
Bestsellers
Mystery
Romance
Sci Fi
YA
0
0.5
1
1.5


Paranormal
Bestsellers
Mystery
Romance
Sci Fi
YA
0
2
4
6


Time & Things
Bestsellers
Mystery
Romance
Sci Fi
YA
0
0.05
0.1
0.15


Family & Children
Bestsellers
Mystery
Romance
Sci Fi
YA
0
0.2
0.4


City Life
Bestsellers
Mystery
Romance
Sci Fi
YA
0
0.2
0.4
0.6


American Military
Bestsellers
Mystery
Romance
Sci Fi
YA
0
1
2
3


Cars
Bestsellers
Mystery
Romance
Sci Fi
YA
0
0.2
0.4


Royalty & Magic
Bestsellers
Mystery
Romance
Sci Fi
YA
0
1
2


Domestic Space
Bestsellers
Mystery
Romance
Sci Fi
YA
0
1
2


Crime
Bestsellers
Mystery
Romance
Sci Fi
YA
0
1
2
3


Space Travel
Bestsellers
Mystery
Romance
Sci Fi
YA
0
5
10


Family
Bestsellers
Mystery
Romance
Sci Fi
YA
0
1
2
3


Senses
Bestsellers
Mystery
Romance
Sci Fi
YA
0
5
10


Sex
Bestsellers
Mystery
Romance
Sci Fi
YA
0
2
4
6

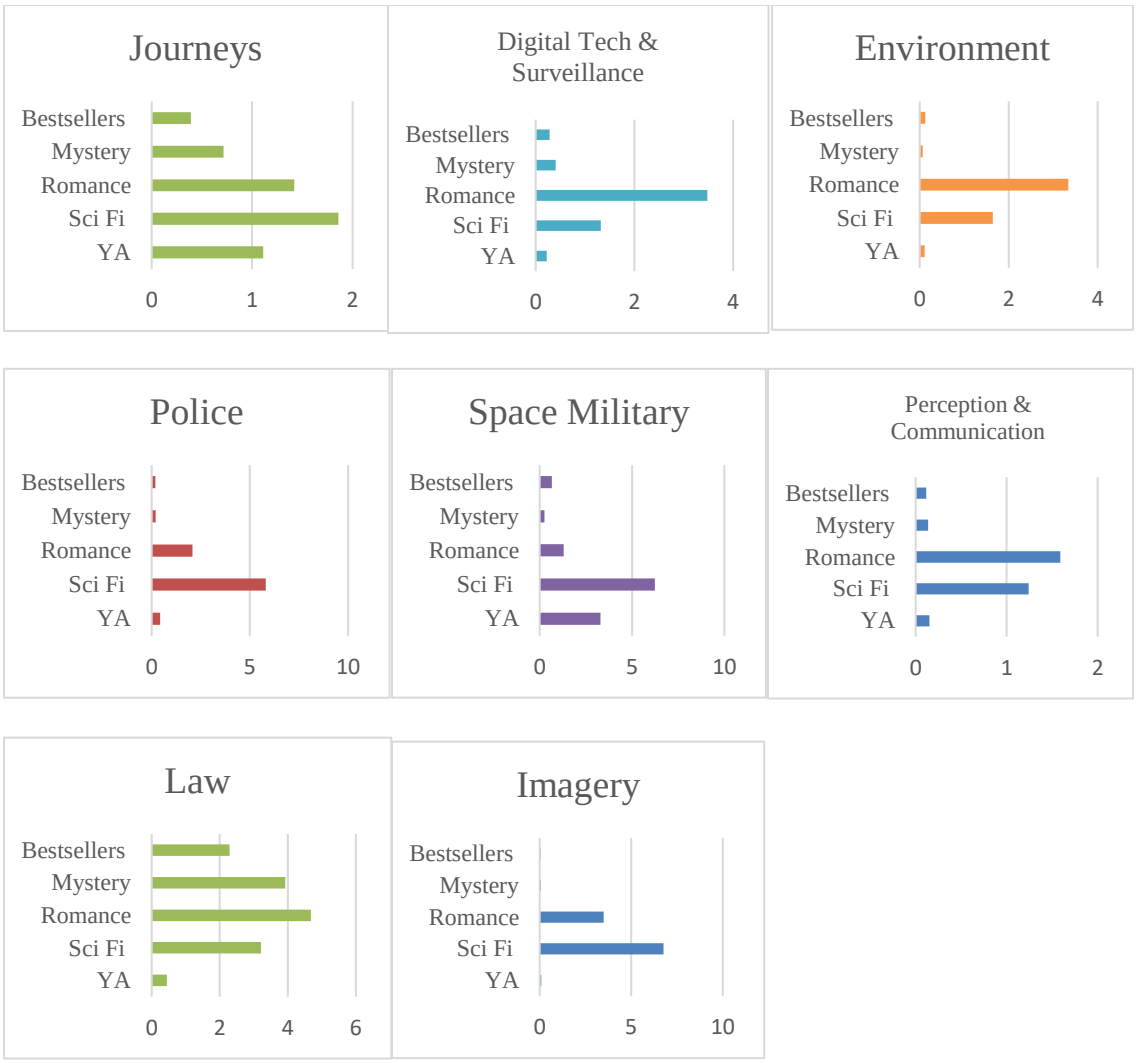


Figure 3.3. Average Novel-Level Topic Weights by Genre in Individual Bar Graphs

domestic space appears less in bestsellers and mystery novels. There is a distinction between *American military*, with higher average weights in romance and young adult and *space military* in science fiction and YA, perhaps due to differing vocabularies (for example, different weapons).

Science fiction and young adult have higher average weights in *city life*, while science fiction and romance hold higher weights in *environment*—again perhaps pointing to the amount of settings-description in science fiction, and a tendency towards rural

(romantic?) environments in romance novels. There is also a distinction between *family* where romance holds the highest weight, and *family & children,* where romance drops, and the other genres hold higher average weights; pointing to a genre difference in focusing on a couple as a unit, or a family with children as a unit. *Imagery*—and by extension, *senses—perception & communication, sex* and *digital technology and surveillance* all follow the tendencies explored in the most-related novels section.

However, other topics hold surprises: *royalty & magic* holds weight across multiple genres, even in mystery (the "royal diamond" trope, perhaps), which may also point to the fact that Fantasy is not separated out as an independent genre. However, romance expectedly still holds the highest average weight. *Journeys* too holds weight across genres, though science fiction unsurprisingly outweighs the rest with the prevalent trope of space travel. The romance genre has a higher average weight in the police topic—the focus being policemen, perhaps—and *law* is precipitously lower in young adult, perhaps due to the lawless settings prevalent in the dystopian sub-genre of young adult fiction. Much of the interpretation of these genre comparisons are conjecture—the goal of this study is to provide a high-level view of the contemporary popular literary landscape and open up questioning space for revisited and further literary analysis.

3.3 *Trend Analysis: Topics Over Time*

The dataset is disproportionally weighted towards the later years of the range (2014-2016), so trend analysis will only be explored for these years to avoid misleading results and interpretation. Figure 3.5 displays each topic's average weight across all genres, in each of the three years, plotted as a line graph to visualize increases and decreases.

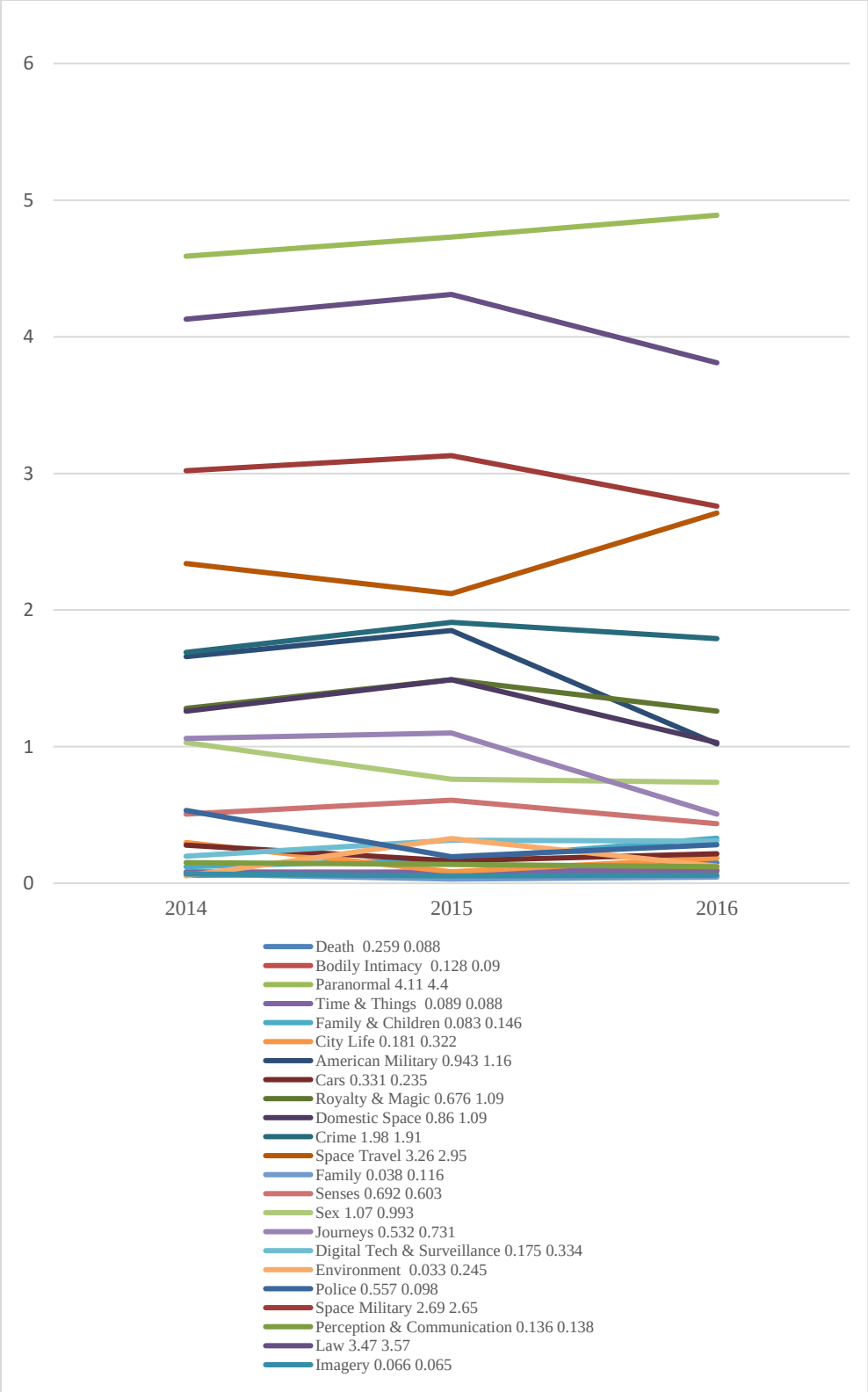


Figure 3.4 Trend Analysis Stacked Line Graph, Average Topic Weights 2014-2016

Paranormal has the most clearly increasing trend, which is expected given the great commercial success of vampire, werewolf, ghost, mythical, supernatural television, movie, and novel series in this time period. *Digital Technology & Surveillance* also has an increasing trend again unsurprising given the increasing omnipresence of digital technology in society. *Sex* has an overall decreasing trend. These are of course not *all* novels published between 2014 and 2016; this visualization simply gives a sense of the trends within this specific—though representative—dataset. The prevalence of topics that increase and then decrease or vice versa also points to the limitations of preforming trend analysis over so few years; it is difficult to really see substantial shifts in the depictions of these topics. Some topics, like *imagery*, however are extremely consistent over these years, pointing to aspects of novels that are underlying and consistent.

3.4 *Novel-Level Conclusions*

Overall, the novel-level experiment provided interesting, surface-level results, as only the most overarching topics can be distilled within documents (novels) that are made up of millions of words. At this high level, much conjecture is required to interpret results. Most results fell along genre convention lines, though there were a few surprises in average topic weights across genres, as well as surprises in how topics actually quantitatively manifest and are detected in topic modeling processes. Issues of abstraction, quantity-focused results and dataset representation were also raised, as well as the limitations of trend analysis over short periods of time. Generally, novel-level topics were helpful for an overview of the topics in novels across genre lines, but a more granular approach is necessary to identify more "latent" topics from a computational perspective.

CHAPTER 4

EXPERIMENT RESULTS, 9,296,167 SENTENCES

4.1 *Thematic Grouping and Overview of Sentence-Level Topics*

Of the 80 topic wordlists created when modeling the dataset as a corpus of 9 million sentences—each sentence considered a document—66 were determined to be semantically significant. In order to generatively compare and discuss the topics, they were (subjectively) separated into 9 thematic groups, as displayed in Figure 4.1. At the more granular level of the sentence as document, more specific topics were distilled.

The computational perspective here highlights aspects of novels and human life that are so ubiquitous as to elude attention, like human embodiment—bodies, breathing, *eating*—or everyday objects like *furniture.* Topic modeling also identified conventions of the novel form that may also be taken for granted in traditional research, like linear progression (*time passing*), forms of characterization, *thought & interiority* and settings. While what is remembered from a novel may be its plot, statistical analysis highlights how much of the actual text is made up of descriptive and background information. Topic modeling also distilled a host of socio-culturally relevant topics including human institutions like *religion* and *government,* actions and activities like *shopping* and *celebration,* depictions of violence like *gun violence* and *war*, and abstract topics like *judgement* and *human difference*. In all visualizations the color of each genre remains consistent, so it is possible to compare across thematic topic groups as well.

Table 4.1 Sentence-Level Topics Separated in Thematic Groups

Embodiment Topics	Sensory Topics	Abstract Topics
Walking Drinking Physical Intimacy Sleep Body Parts Cardiovascular Physiological Responses Eating Injury Eyes Body - Violence & Accidents Embodied Emotions	Temperature, Wind & Sensations Sound Light, Color & Sky Whispers Gaze & Looking Touch	Goodness, Badness and Judgement Joy Home & Country Life & Love Humans, Aliens & Difference Explicit Content
Novel Conventions Topics	**Settings Topics**	**Objects & Ubiquitous Topics**
Thought & Interiority Physical Characterization Gendered Characterization Time Passing Time Periods Nonverbal Communication	Urban Setting Rural Setting Indoor Settings, Rooms The Environment House Setting	Metal Animals Stuff You Can Put Stuff In Furniture Waiting Clothes Doors & Windows
Human Institutions	**Actions & Activities**	**Violence & War**
Government Religion Law Healthcare Mass Media & Information Family Money School	Reading & Writing Cellphone Use Driving Travel Digital Technology Use Work & Doing Shopping Group Social Interaction Entertainment Holidays & Celebration	War & Intelligence Agencies Gun Violence Military War Death & Families

4.2 *Embodiment Topics Across Genres*

Interestingly, the topics identified by computational methods overwhelming center the human body and its functions. Figure 4.1 illustrates the average weight of each

embodiment topic within all of the sentences of all of the novels within each genre, clustered so that it is possible to compare the relative weights of each topic across genres.

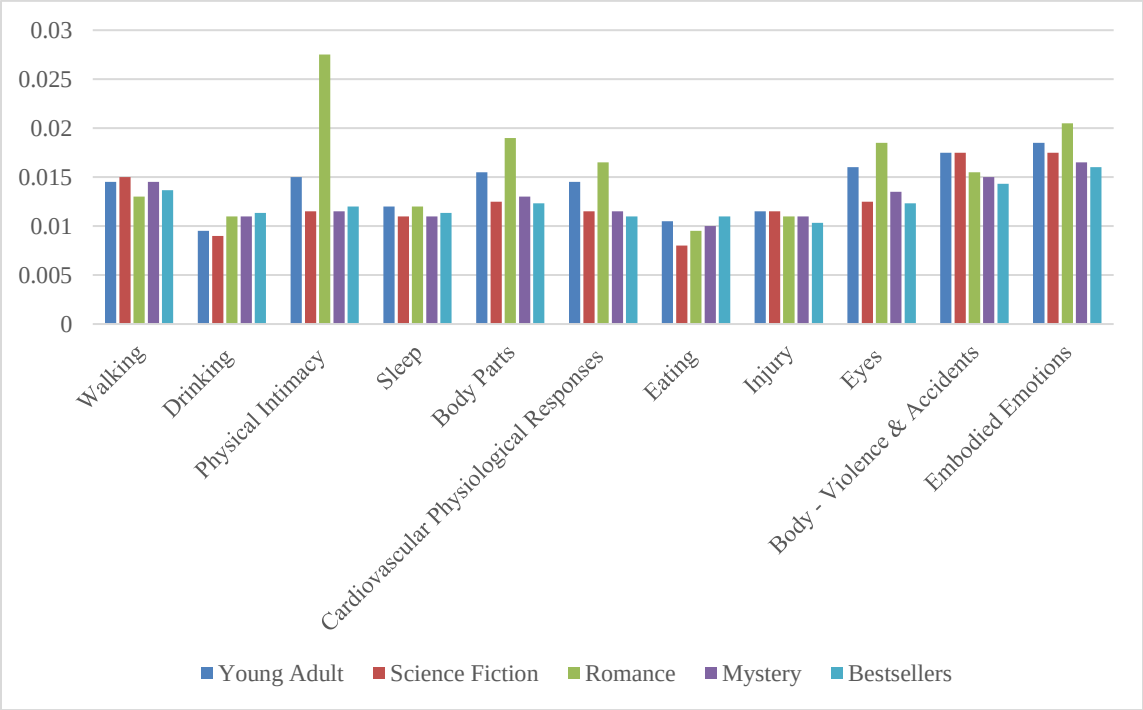


Figure 4.1 Bar Graph, Average Weights of Embodiment Topics Across Genres

Notably, the romance genre averages at much higher rates for several topics in this category, most dramatically in *physical intimacy*—not surprisingly. Young adult fiction also averages at higher rates for the same topics as romance, including physical intimacy, and perhaps relatedly, *body parts, sleep, cardiovascular physiological responses, eyes* and *embodied emotions*. Like the novel-level experiment reported, sexual activity and intimacy appear more prevalently in the romance genre and, to a lesser but still significant extent, young adult genre. *Embodied emotion* appears in this category because the original wordlist, "felt feel feeling body pain mind fear anger inside thought moment sense relief heart guilt panic stomach" makes clear the topic is the experience of feeling emotions bodily and includes multiple different emotions. Similarly, close

readings of related sentences revealed that the *cardiovascular physiological responses* topic is not tied to any one activity—including sexual activity, running and panic—but is instead united by the shared bodily responses of each of these activities.

Other, more universal, topics were more consistent across genres. While *eating* and *drinking* averaged at lower rates in science fiction, generally these two basic human functions are consistent across genres. *Sleep* is also consistent, though romance and YA are again slightly elevated: potentially in relation again to the elevated depictions of sexual intimacy. Surprisingly, despite the variability in depictions of violence seen in a later section (4.10), the weights of the topic *injury* are fairly consistent across genres – again pointing to embodiment and human fragility. The topics that related to action—*walking, body accidents & violence*—were more heavily weighted in the science fiction and young adult genres which frequently include action-heavy novels.

4.3 *Sensory Topics Across Genres*

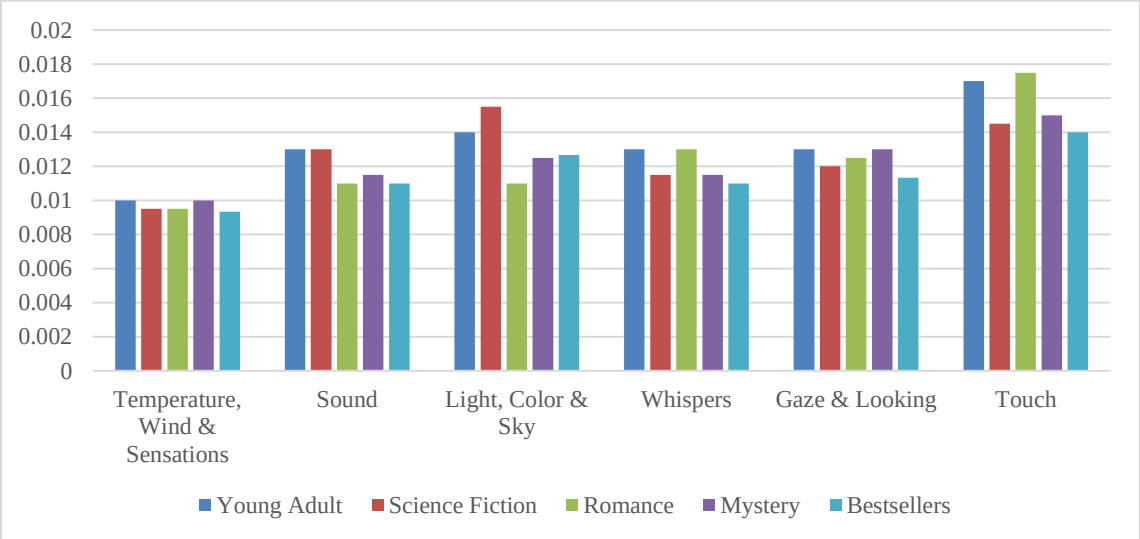


Figure 4.2 Bar Graph, Average Weights of Sensory Topics Across Genres

The sensory category was delineated from the embodiment category based on a focus on sensing rather than specific body parts, so *gaze* appears here while *eyes* appears in embodiment. Figure 4.2 breaks down the average topic weight in each genre, for each topic in the sensory category. Topic weights within this category are consistent across genres, particularly in the *temperature, wind & sensation* and *gaze & looking* topics, with quite granular differences. Sensory imagery and a character's experience of sensing their fictional world is more of an underlying topic that overarches the human experience and boundaries of genre. However, there are slight differences; similarly to the increased weight of *physical intimacy* in the previous category, *whispers* and *touch* each average more highly in the young adult and romance genres. Science fiction and young adult fiction each average higher weights in the *sound* and *light, color & sky topics*, perhaps in relation to the higher rates of the topic *imagery* in these genres in the novel-level experiment.

4.4 *Novel Conventions Topics Across Genres*

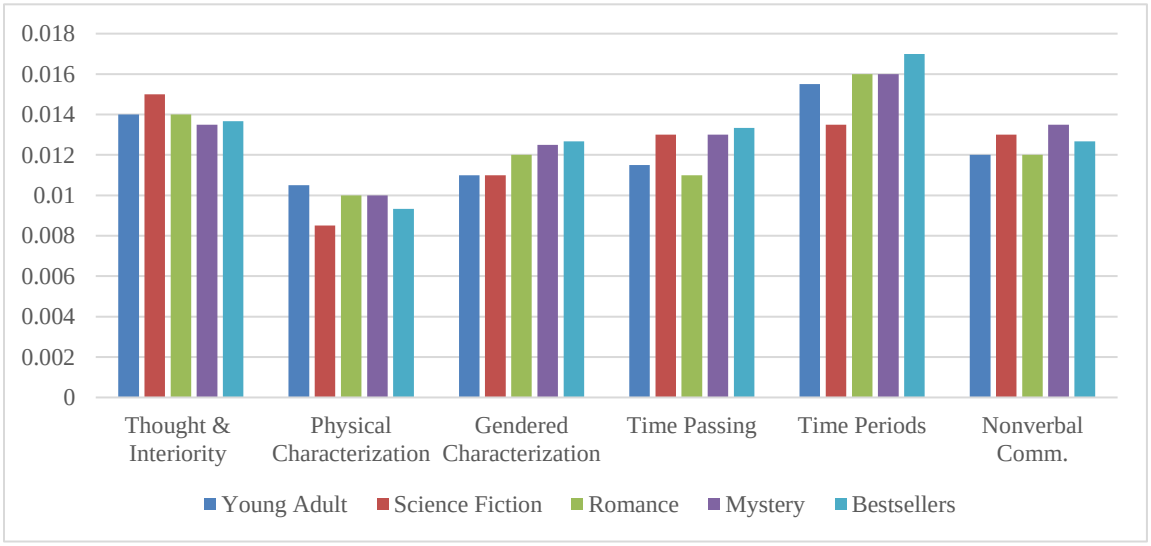


Figure 4.3 Bar Graph, Average Weights of Novel Conventions Topics Across Genres

Topics in this category are again more consistent across genres because they are reflective of the format of the novel form itself; these topics could be considered the "parts" that make up a novel. There are slight differences: despite the overall high weight of *time passing*—an indication of narrative linearity—sentences from the romance genre average at a lower rate, perhaps because romance novels are generally shorter and sometimes happen in short windows of time. Another indicator of linear narratives, *time periods*, while again generally high-weighted, is lower in the science fiction genre perhaps because demarcations of time like "day" and "night" may be less prevalent in a novel set in space.

Characterization, both *physical characterization* and *gendered characterization*, is certainly a necessary part of the novel form in any genre, though slight variability is present; the decreased average weight of *physical characterization* in science fiction may point to a greater diversity in the bodies being characterized, as an alien character may not fit the same topic-list words like "hair." Though this study does not delve into the kinds of gendered characterization (future studies into sub-topics of this topic could be fruitful), the fact that gendered characterization appears as a topic points to the prominent role gender plays in how characters are described, though perhaps less (quantitatively) in young adult and science fiction characters.

Thought & interiority, indicative of the frequently first person or kinds of third person points of view, is prevalent in all genres and slightly more so in science fiction. While dialogue, another feature of novels, does not appear as a topic due to the variability of what is actually being said in a sentence of dialogue, *nonverbal communication* does

register. Highest in the mystery genre, this could point to secrecy and hidden messages not said out loud.

4.5 *Settings Topics Across Each Genre*

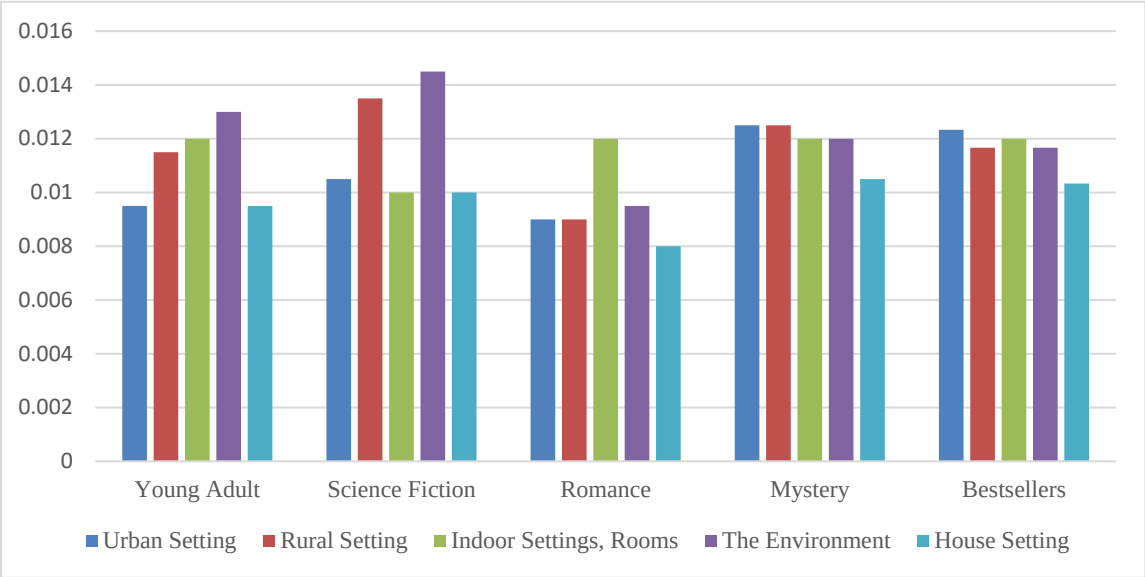


Figure 4.4 Bar Graph, Average Weights of Settings Topics Compared by Genre

This visualization (Figure 4.4) is slightly different, with bars clustered by genre rather than by topic to better compare which kinds of settings-related sentences appear within each genre. Again, this experiment is at the level of the sentence; while the primary setting of a novel overall may be rural or urban, this analysis identifies how frequently, on average, sentences of a genre contain descriptions of or references to the many possible settings encountered in any given novel. These topics are also generally weighted less on average overall (see Figure 4.4's y axis). Despite a high average weight in the *city life* topic in the novel-level experiment, in this experiment it is revealed that more sentences in the young adult genre are dedicated to describing the natural *environment*, *indoor settings* and *rural settings*. Generally, each genre contains sentences

that describe all of the settings-related topics, as no genre is constrained to a single setting. Surprisingly, science fiction sentences are more likely to describe natural *environments* or *rural settings* than urban ones—though this may be because nature is more similarly depicted—that is, using the same words—in speculative fiction in contrast to new age cities. Less surprisingly, many sentences in romance novels occur in *indoor settings*. Mystery and bestsellers are more equally distributed, with sentences dedicated to describing all settings options. Generally, the *house settings*—that is, sentences related to the exteriors of houses—are the least frequent, and urban, rural, environmental and indoor settings occur across all genres.

4.6 *Objects & Ubiquitous Topics Across Genres*

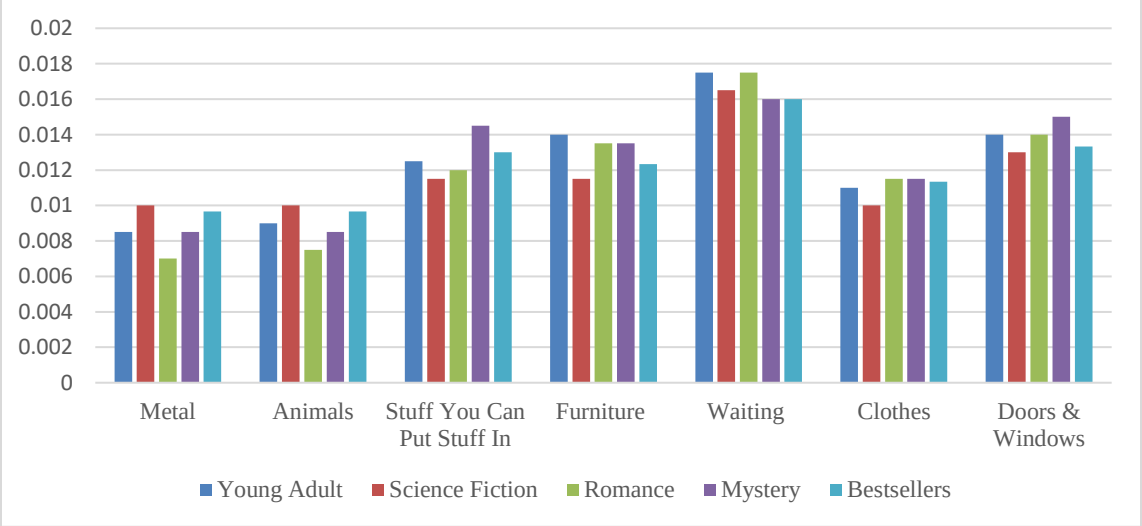


Figure 4.5 Bar Graph, Average Weights of Objects & Ubiquitous Topics Across Genres

The topics in this category are some of the most surprising; how often do human readers think of *metal*, *waiting* or *furniture* as prevalent features within novels? *Waiting* has the highest weight across genres, pointing again to linearity but also to the human experience of waiting, as well as the action that may be more focused on by a casual

reader or researcher. *Furniture* and *clothes* are pretty consistent across genres, as more universal background information—though slightly less weighted in science fiction, where clothing and furniture may be more futuristic or described with differing vocabularies. The *stuff you can put stuff in* category is made up of words describing objects with a sense of interiority, secrecy or hiddenness—like bags, boxes, pockets and containers. Unsurprisingly, mystery out-ranks other genres here, with its conventional focus on secrecy and disclosure. *Doors and windows*, while prevalent in all genres, also has a slight spike in mystery, calling to mind "whodunnit" narrative tropes. *Animals* and *metal*, while interesting that they were topics at all, are more difficult to interpret; these topics appear less on average in the romance genre, but more science fiction and bestseller genres.

4.7 *Human Institutions Topics Across Genres*

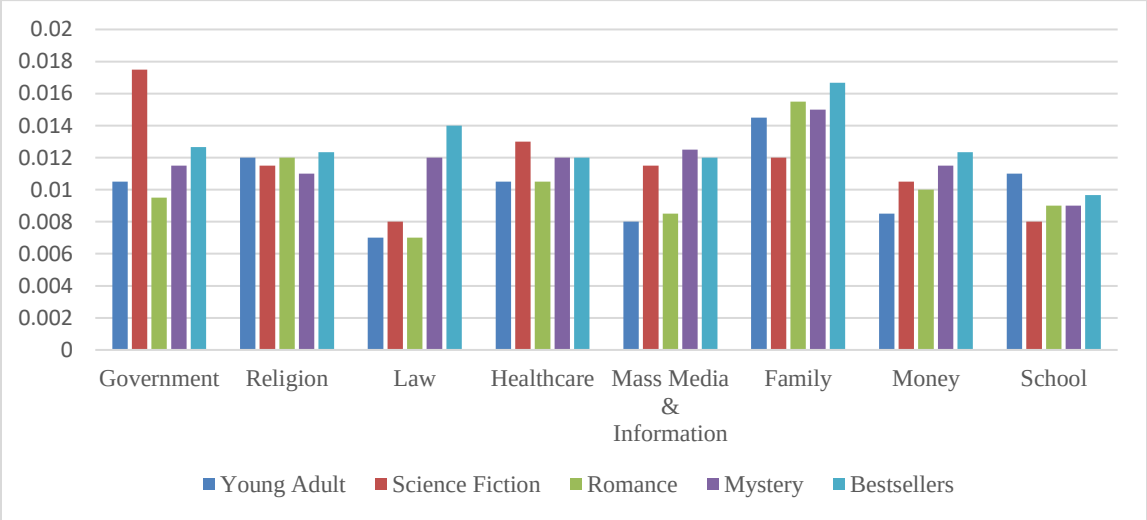


Figure 4.6 Bar Graph, Average Weights of Human Institutions Topics Across Genres

While all topics are cultural in nature, the topics in the following sections (4.7-4.11) were subjectively considered as more socially and culturally relevant phenomena.

The most dramatic genre-difference in the first socially-relevant category, Human Institutions, is within the *government* topic, perhaps due to the popularity of the speculative fiction subgenre of science fiction. *Law* also holds dramatic differences along genre lines, with bestsellers and mystery holding much higher averages—unsurprisingly given the tropes of murder and running as fugitives in both fiction categories. *Family* is also highly weighted across genres, particularly in bestsellers, though less so in science fiction where a social unit may not be described in the same "mother, father, children" vocabularies. *School* as a topic is also fairly consistent, with an unsurprisingly higher weight in the young adult genre. *Money* is also equally distributed, with romance not averaging higher despite the trope of extremely wealthy suitors, while mystery and bestsellers average at higher rates again perhaps as a reflection of crime-focused plots. *Religion* as a topic is consistent across genres, as a more underlying factor of characters' lives than as the topic of novels themselves.

For the information science goals of this study, the depiction of mass media & information is of interest, averaging higher in mystery and bestsellers where characters may be researching or watching the news in crime-focused plots (or, in trope fashion, seeing themselves on television). Mass media also averages at higher weights in science fiction, perhaps again because of the sudden apocalyptic plots or government-controlled dystopian worlds featured frequently in science fiction. Despite the consistency of *injury* across sentences in all genres in the embodiment section (4.2), average weights of the healthcare topic, while high overall, are more variable, with more appearing in the science fiction genre.

4.8 *Actions & Activities Topics Across Genres*

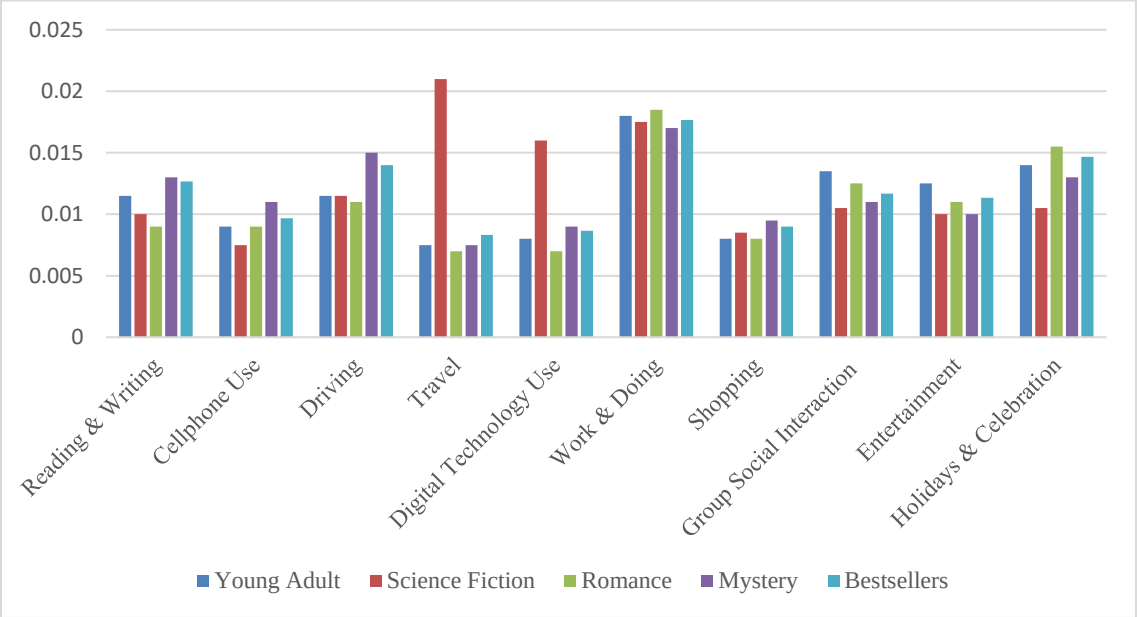


Figure 4.7 Bar Graph, Average Weights of Actions & Activities Topics Across Genres

Work & doing, in its vagueness, is the most highly weighted human activity across genres in this category. Science fiction holds unsurprisingly high averages in *travel* (space travel) and *digital technology use*, while the more social topics (*group social interaction*, *entertainment* and *holidays & celebrations*) appear more frequently in the young adult, romance, and to an extent bestseller, genres. The functional activities—*driving* and *shopping*—have higher averages in the mystery and bestseller genres. While science fiction dominates the broader *digital technology use* topic, on average more sentences from the mystery genre feature *cellphone use*. Of interest to the literary aspects of this study, depictions of *reading & writing* are also prevalent enough to appear as a topic, most frequently in mystery and bestseller novels.

4.9 *Violence & War Topics Across Genres*

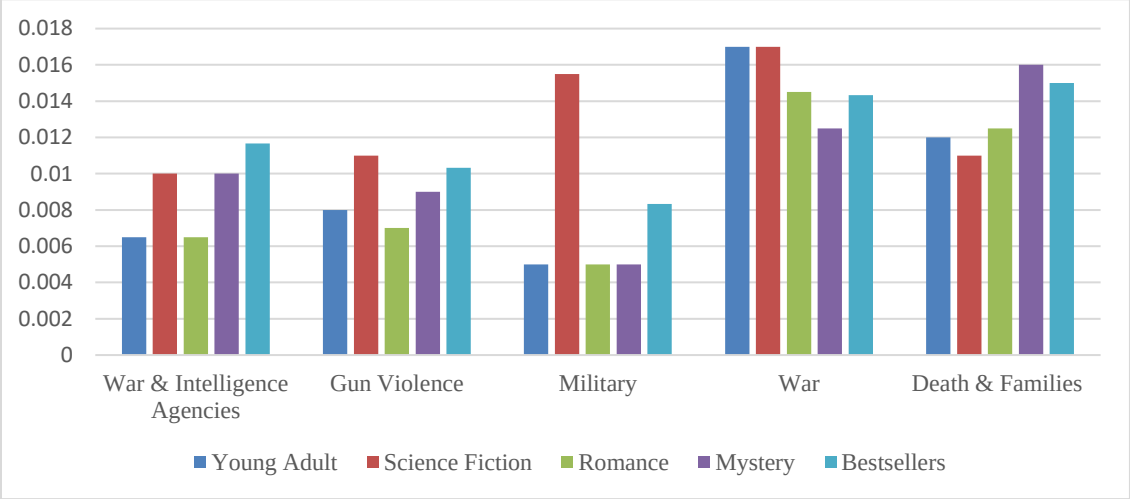


Figure 4.8 Bar Graph, Average Weights of Violence & War Topics Across Genres

This topic modeling-based study is again exploratory in nature—it cannot speak to the nature of the depictions of the topics that are distilled, just that they appear at all and to what frequency. Future computational studies could shed light on the nature of depictions of socially relevant topics like the ones in this study, for example through the use of sentiment analysis to determine at scale whether topics—like violence, here—are described positively or negatively. Regardless, the fact that so many topics, at high averages, contain violence points to the overall prevalence of violent content in recently published popular fiction across genres. *War* as a topic appears at higher averages in young adult and science fiction genres, while a *military* structure appears dramatically more in science fiction. More specifically, *war & intelligence agencies* average more highly in bestsellers than young adult or science fiction, with the prevalence of espionage-focused plot lines. *Gun violence*—with or outside of war—as a topic appears similarly most often in science fiction and bestsellers, though each of these topics do

appear at significant average rates in all genres. Explorations of the impact of violence (as well as presumably non-violent deaths, though the most-related sentences all related to violence)—the *death & families* topic—appears at the highest average weights in mystery, where plot lines often begin after the violence or murder has taken place.

4.10 *Abstract Topics Across Genres*

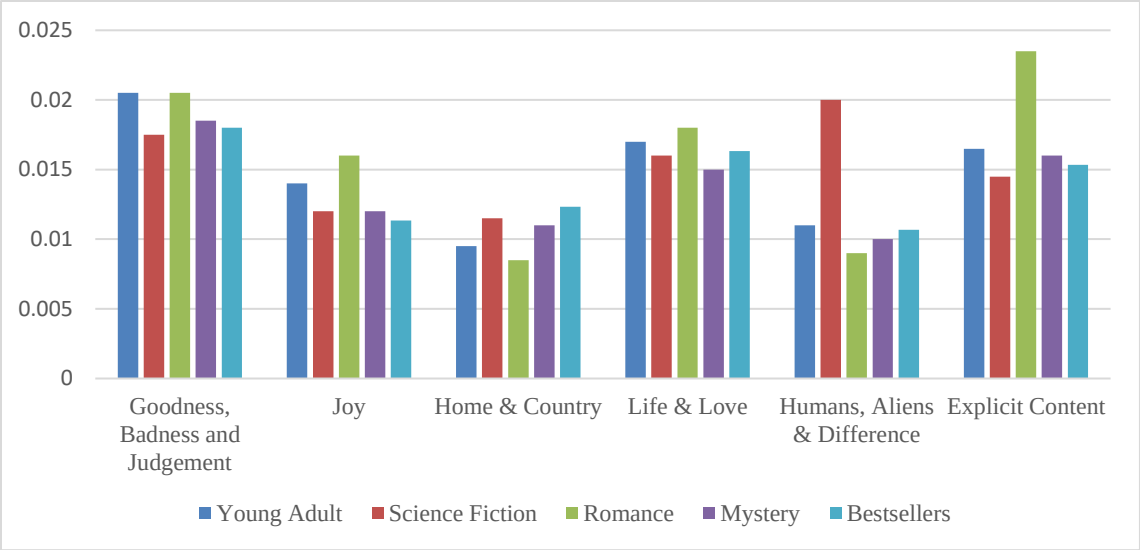


Figure 4.9 Bar Graph, Average Weights of Abstract Topics Across Genres

The final category of topics gleaned from the sentence-level dataset were the most semantically ambiguous topics, which required close reading of the most-related sentences. Explicit content appears here rather than in embodiment because the related sentences involved both sexual activity—which explains the highest weight in the romance genre—and a more general use of expletives. While humans, aliens and difference appeared most often in science fiction, related word groups were also used to describe other kinds of human difference (without the presence of aliens), such as racial difference in other genres' sentences. The more vague topic *life & love* averaged highly across genres, slightly more so unsurprisingly in the romance genre, with *joy* following

similar patterns across genre lines. The broad topic *home & country*, which in some contexts could also be called nationalism, also appeared across genres slightly more so in the bestseller category. *Goodness, badness & judgement* also appeared across genres at high averages, especially young adult and romance genres.

4.11 *Sentence-Level Experiment Conclusions*

Compared to the novel-level experiment, this experiment's results were much more specific and variable across genres. Overall, the most highly weighted topics were goodness, badness & judgement (.045), work & doing (.043), embodied emotions (.042), waiting (.041), explicit content (.04), life & love (.039), time periods (.038), bodily violence & accidents (.037), war (.036), touch (.035) and family (.035). These highest-weight topics are a mix of familiar literary themes, and less expected features like novel conventions and ubiquitous experiences and a focus on embodiment, human bodies and tactile intimacy.

CHAPTER 5

THE EDGES OF COMPUTATIONAL ANALYSIS

5.1 *The Limitations of Text Mining*

While this study offers a distant, alternate view of the literary landscape in recent years, it is inherently a limiting and abstracting process. There has been no shortage of heated debate in literary studies—and in the larger field of digital humanities—about the appropriateness and limitations of applying large-scale content analysis methods to fictional text. In recent years, scholars within the #transformDH collective have called for broader representation of diverse literary works in text mining studies, more focus on context to lessen the methods' reductive tendencies, and new methods of analysis that critique power rather than reinforcing it (Klein, 2018). This study has endeavored to reduce these limitations by relying on an unsupervised machine learning technique rather than set expectations that reinforce dominant expectations (such as a binary gender approach). The study has also aimed to have a representative sample of diverse popular texts and to reapply context through the utilization of qualitative methods as well as quantitative methods.

The mixed methods approach—that is, utilizing both information science methods and close reading methods—employs the strengths of both fields, and aims to limit the limitations of each approach. As literary researchers James F. English and Ted Underwood have argued, quantitative methods must have a "place at the table" in studies

of fictional material, without fulfilling concerns in the humanities that they will completely overrun, devalue, or eclipse other literary methods of analysis and interpretation (2016, p. 292). Far from being "incompatible," content analysis and close reading, for English and Underwood, can operate together: "long-term trends" only making sense when viewed in the light of "local histories" (2016, p. 292), in this case, most-related novels and sentences.

However, this study is still inherently limited in that it relies completely on generalizations across mass amounts of text, losing the nuance of traditional literary studies. The texts chosen are also limited—they do not represent all American readership, rather focusing on which texts have been *purchased* most often. The texts in the dataset also reflect publishing trends as much as readership, in that they are still limited—as American publishing is—by a higher percentage of white, able-bodied, cisgender, heterosexual characters and stories. This study also does not include nonfiction texts, or online media outside of print texts and is focused only on texts published in English. Even within these constraints, the dataset, while representatively sampled, is only a slice of the entire body of popular fiction published in recent years.

The study is also limited by the assumption of topic modeling that words appearing the same sentences are inherently related, which may not always be the case in all sentences. For example, a topic which was removed from sentence-level wordlist included a conglomeration of words all related to the word "arms": arms of chairs, bearing arms and human arms, which is not actually a single cohesive topic. The study is also limited in that it is difficult to prove that topics in fictional material influence readers, rather than the topics only reflecting the opinions of authors or publishers. This

assumed relationship is more relevant for investigation in a psychology study. However, future information science studies could also greatly strengthen findings like these by applying the additional research methods of survey and focus groups to better assess how actual readers interact with and think about quantitatively analyzed fictional material. In light of these limitations, it may be helpful to consider this study not as perfect distillations of fictional trends or of American collective cultural thought, but as an additional outlook to complement existing literary scholarship on popular texts and information science scholarship focused on media and information dissemination.

Considering these limitations, future studies may consider representative texts outside mainstream publication, consider library circulation as well as sales numbers, include textual media from online platforms or texts in languages other than English, or further investigate the sub-topics of the initial topics discovered here.

5.2 *Ethical Considerations*

In a world of increasingly prevalent and accessible data, there has also been much debate around how secondary data analysis must ethically engage with the re-use of data, individuals' data protection rights, anonymity, and consent (Tripathy 2013). While this study is largely separate from many of the concerns of data re-use—if only because the textual "data" is not connected to or gleaned from individual human subjects—there are still ethical considerations. This study has sought the consent of the original researchers and will endeavor to acknowledge the original creators of the dataset, as well as all authors of the fictional texts. The study, through the scrambling of word-order in sentences, has attempted to remove the ethical implications of free access to texts under copyright. However, the study is unable to freely share the dataset, with implications for

the lack of complete reproducibility. While fictional texts, by their nature, are implied to have consented to being interpreted by any number of human readers, the authors of the texts have not consented to being included in this secondary study. In that the study is inherently generalizing, there is a chance that any text here is associated with a generalized "negative" topic or may be included in a sentence-level topic group that does not represent the work in its entirety. While striving for inter-coder reliability through the use of triangulation during qualitative interpretation, topic selection, labeling, thematic grouping and interpretation has of course still been informed by the researchers' individual interpretation, literary background and personal views.

CHAPTER 6

CONCLUSIONS

To better understand the content of recently published popular fiction, this study provides a macro-analysis of what topics appear prevalently across 1,136-novel dataset. The study offers a wide view of frequently occurring topics, as well as sense of genre and temporal patterns and comparisons. This study illustrates how text mining techniques including topic modeling can be generatively used to quickly and efficiently investigate large amounts of text that would take years to read at an average reading pace. By deploying unsupervised machine learning, the study also identifies topics and trends that a human researcher using qualitative methods and with preexisting cultural and literary expectations may not recognize or view as notable. The findings of this study offer an additional outlook on both the American contemporary literary landscape and as an abstracted snapshot of American culture, as mediated through fiction.

This study may be beneficial to literary researchers looking to understand the larger picture of contemporary popular fiction, to the reader looking for texts related to topics they enjoy, and to individuals interested in what trends and topics appear in fiction as a reflection of contemporary culture. It may also be beneficial to information science researchers interested in how topic modeling and content analysis methods can be applied to fictional text, or interested in the limitations and considerations necessary when working with fictional material. Ideally, this study could be useful to publishers both in

identifying popular topics people are purchasing, and in identifying the gaps in what is being published.

Some key conclusions of this research are as follows:

- Computational tools can be generatively used to vastly broaden the scope of literary analysis, but results must still be interpreted through qualitative means.
- The novel may be quantitatively analyzed at both the level of the entire novel and the level of the sentence but analyzing at the level of the sentence offers more granular and interesting results.
- Computational analysis can be utilized to identify latent, ubiquitous topics from a different "machine" perspective, that a human researcher may ignore or miss.
- Natural language processing methods re-center research focus on the human body, its functions and the embodied nature of fiction.
- Statistical modeling is able to identify novel conventions such as linearity, characterization and settings.
- Topic modeling was able to distill many socially relevant topics from a dataset of popular fictional media at both the novel and sentence level, including violence, surveillance and human institutions and activities.
- Topics varied widely by genre. While topic modeling reinforced some topical expectations based on genre conventions and tropes, topics also appeared unexpectedly in other genres: helping re-imagine the popular fiction landscape outside of genre-based siloes.
- Statistical analysis of a fictional dataset offers a new, birds-eye view, but also has many limitations and should not be taken as fact or scientifically "true."

This study shows the potential of text mining research to understand large-scale fictional corpora as well as the usefulness of topic modeling to investigate themes and cross-genre patterns across a decade. This study proposes a methodology not just to investigate the topical contours of the contemporary fictional topography, but also to illustrate how information science methods can be generatively applied to fictional and

artistic material, in ways focused on content and fiction-as-information, in addition to questions of style. Future data-heavy research in often under-explored fictional media could offer further insights into trends and commonalities in what cultural narratives and mis/information are being shared at scale.

.

REFERENCES

Adkins, T., & Castle, J. J. (2013). Moving pictures? Experimental evidence of cinematic influence on political attitudes. *Social Science Quarterly*, 95(5), 1230–1244. doi:10.1111/ssqu.12070

Barna, I., & Knap, Á. (2019). Antisemitism in contemporary Hungary: exploring topics of antisemitism in the far-right media using natural language processing. *Theo Web*, 18(1), 75–92. https://doi-org.pallas2.tcl.sc.edu/10.23770/tw00

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*, 127, 256–271. https://doi-org.pallas2.tcl.sc.edu/10.1016/j.eswa.2019.03.001

Bishop, R. D. (1990). Mirrors, windows and sliding glass doors. *Perspectives: Choosing and using books for the classroom*, 6(3), https://scenicregional.org/wpcontent/uploads/ 2017/08/Mirrors-Windows-and Sliding-Glass-Doors.pdf

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*. doi:10.1109/msp.2010.938079

Bonilla, T. & Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics*, 41(6), 650–669. doi:10.1016/j.poetic.2013.06.003.

Bos, M. & Van Den, J. M. (2019). Mining public discourse for emerging dutch nationalism. *DHQ : Digital Humanities Quarterly*, 10(3), 1–12. http://www.digitalhumanities.org/dhq/vol/10/3/000263/000263.html

Cohen, R., Aviram, I., Elhadad, M., & Elhadad, N. (2014). Redundancy-aware topic modeling for patient record notes. *PLoS ONE*, 9(2), 1–7. https://doiorg.pallas2.tcl.sc.edu/10.1371/journal.pone.0087555

Cain, J. O. (2016). Using topic modeling to enhance access to library digital collections. *Journal of Web Librarianship*, 10(3), 210–225. doi: 10.1080/19322909.2016.1193455

English, J. F. & Underwood, T. (2016). Shifting scales. *Modern Language Quarterly*, 77(3), 277–295. doi:10.1215/00267929-3570612.

Graham, S., Weingart, S. & Milligan, I. (2012). Getting started with topic modeling and MALLET. *The Programming Historian*. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

Haraway, D. J. (2016). A cyborg manifesto: Science, technology, and socialist-feminism in the late 20th century. *Manifestly Haraway*, 3-90. doi: 10.5749/minnesota/9780816650477.003.0001

Hendrickx, I. & Onrust L. (2019). Unraveling reported dreams with text analytics. *DHQ : Digital Humanities Quarterly*, 11(4), 1–12. http://www.digitalhumanities.org/dhq/vol/11/4/000342/000342.html

Jähnichen, P., Group, M. L., & Berlin, H. (2019). Exploratory search through visual analysis of topic models. *DHQ : Digital Humanities Quarterly*, 11(2), 1–17. http://digitalhumanities.org:8081/dhq/vol/11/2/000296/000296.html

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics,* 41(6), 750-769. https://doi.org/10.1016/j.poetic.2013.08.005

Karami, A., Swan, S. C., White, C. N., & Ford, K. (2019). Hidden in plain sight for too long: Using text mining techniques to shine a light on workplace sexism and sexual harassment. *Psychology of Violence*. https://doi.org/10.1037/vio0000239

Klein, L. F. (2018). Distant reading after Moretti. *Arcade: Literature, Humanities & the World,* https://arcade.stanford.edu/blogs/distant-reading-after-moretti.

Liu, A. (2012). The state of the digital humanities: A report and a critique. *Arts and Humanities in Higher Education*, 11(2), 8-41. doi: 10.1177/1474022211427364.

Marshall, E. A. (2013). Defining population problems: Using topic models for cross national comparison of disciplinary development. *Poetics*, 41(6,) 701–724. doi:10.1016/j.poetic.2013.08.001.

McCallum, A. K. (2002). MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu/

Mohr, J. W., Wagner-Pacifici, R., Breiger, R. L. & Bogdanov, P. (2013). Graphing the grammar of motives in national security strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics,* 41(6), 670–700. doi:10.1016/j.poetic.2013.08.003.

Moretti, F. (2015). *Distant reading.* London: Verso.

Neatrour, A. L., Callaway, E., & Cummings, R. (2018). Kindles, card catalogs, and the future of libraries: a collaborative digital humanities project. *Digital Library Perspectives,* 34(3), 162–187. doi:10.1108/dlp-02-2018-0004

Piper, A. (2016). Fictionality. *Journal of Cultural Analytics.* doi: 10.31235/osf.io/93mdj

Piper, A. & Portelance, E. (2016). How cultural capital works: Prizewinning novels, bestsellers, and the time of reading. *Post45*. http://post45.research.yale.edu/2016/05/how-culturalcapital-works-prizewinning novels-bestsellers-and-the-time-of-reading/

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography & Geographic Information Science*, 45(4), 362–376. https://doi-org.pallas2.tcl.sc.edu/10.1080/15230406.2017.1356242

Rhody, L. (2012). Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1). http://journalofdigitalhumanities.org/2-1/topic-modeling-and figurative-language-by-lisam-rhody/

Schöch, C. (2019). Topic modeling genre : An exploration of French classical and enlightenment drama. *DHQ : Digital Humanities Quarterly*, 11(2), 1–20. http://digitalhumanities.org:8081/dhq/vol/11/2/000291/000291.html

Turing, I. (1950). Computing Machinery and Intelligence. *Mind*, 59 (236), 433- 460. https://doi.org/10.1093/mind/LIX.236.433.

Tripathy, J. P. (2013). Secondary data analysis: Ethical issues and challenges. *Iranian Journal of Public Health*. 42(12), 1478–1479. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441947/

Underwood, T. (2012). Topic modeling made just simple enough. *The Stone and the Shell*, tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

APPENDIX A

RESEARCH EXPERIMENT GUIDE

For purposes of reproducibility, the following experiment guide outlines the steps of the research process in more detail.

Step 1: Dataset Preparation

- Convert the data from RDA word-sentence matrixes to text file formats, where each sentence of a novel is appended to the text file as a new line. For researchers beginning with other forms of novels, such as eBooks, this step may look different—the conversion being more focused on removing extraneous xml and html coding. Data cleaning may be time consuming for researchers using datasets created from text files produced by optical character recognition software.
- In order to perform the experiment at the level of the sentence, each sentence of a novel must be on a separate line in the text files.
- It is also helpful to follow naming conventions for each text file, including novel name, year and genre, for easier comparative analysis later.

Step 2: Data Cleaning

- Check the text files throughout the dataset, to check for anomalies, unnecessary text (like copyright information) and mismatched file names.
- The word order within each sentence will be "scrambled," making them impossible to read, so files should be verified against the actual text of the novels.

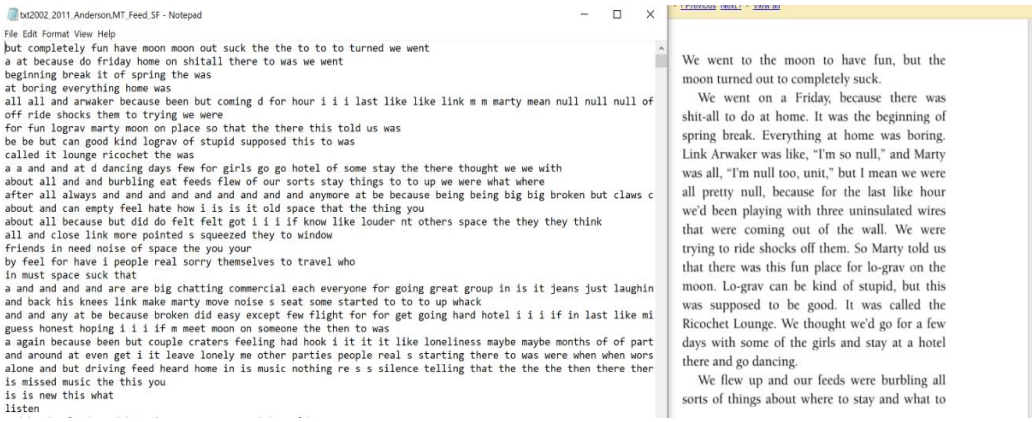


Figure A.1 Image of Text File Versus E-Book Comparison

Step 3: Install MALLET

- The original researcher used a peer reviewed tutorial by Shawn Graham, Scott Weingart and Ian Milligan to install and begin using MALLET (Graham et al., 2012). Direct link: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

NOVEL-LEVEL EXPERIMENT

Here the experiment steps converge; for steps for the sentence-level experiment skip to Step 8.

Step 4: Convert Text Files to MALLET Format

- Once installed, MALLET is used through commands typed directly into the Command Prompt. To begin the novel-level experiment, change directory to MALLET, then move the dataset (novels, each in a separate text file, in a single folder) to the MALLET directory on your device.
- Using the import-directory function, convert the text files to a single mallet-friendly format using the following command, with the data names replaced:

bin\mallet import-dir --input FOLDER NAME—output DATA.mallet --keep-sequence --remove-stopwords --extra-stopwords names.txt

- The remove-stopwords function removed common stopwords like “and” and “the.” The extra-stopwords function can be used if additional stopwords are desired; here, a text file of common names was used to remove character names from the dataset.

Step 5: Optimize the Number of Topics & Topic Model

- This step can be done computationally, notably through the *ldatuning* R package.
- Or, topic numbers can be optimized through trial and error, by running multiple topic modeling iterations with increasing topic-number settings. Increase the number of topics for each iteration using the following command:

  bin\mallet train-topics --input DATA.mallet --num-topics 100 --optimize-interval 100 --output-topic-keys novels100keys.txt --output-doc-topics novels100_compostion.txt

- This command inputs the MALLET file created in the previous step, sets the number of topics to be distilled, and selects what kinds of output files are desired. The topic-keys setting records the topic wordlists in a text file for qualitative interpretation, and the topics-composition setting records the weight of each topic within each novel in a text file.
- This command will begin the topic modeling process, and look like Figure A.2.
- Once the topic modeling process is complete (which may take several hours), opening the output files in Excel is helpful to be able to sort and average the topic weights.

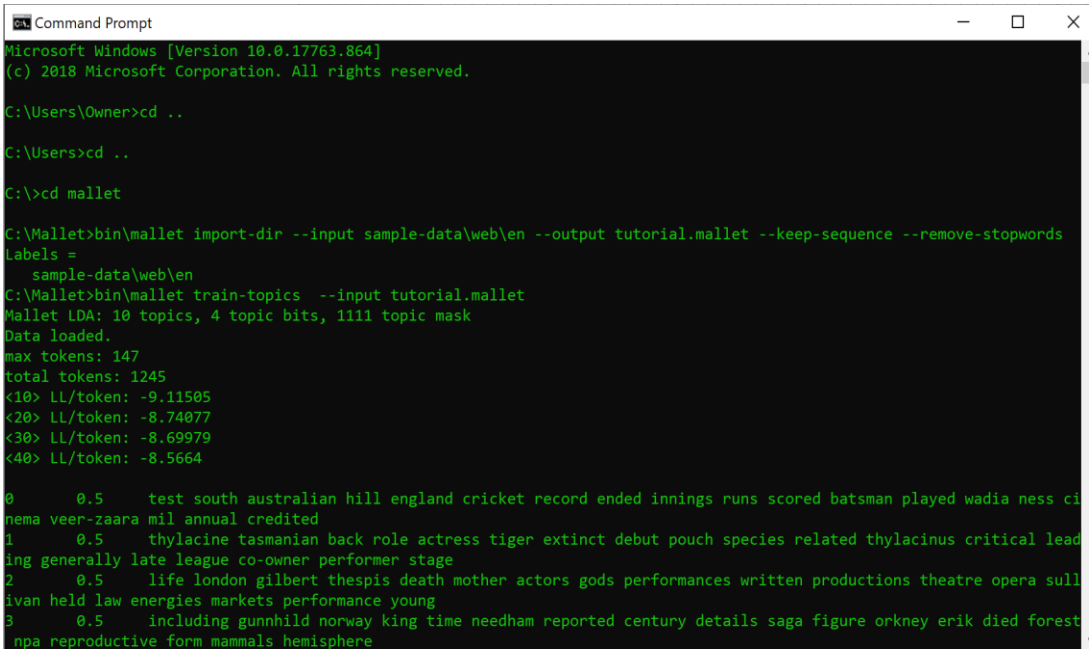


Figure A.2 Image of Mallet running Topic Modeling in Command Prompt

Step 6: Interpret Topic Wordlists & Label Topics

- See section 2.3A for more information on qualitative interpretation, or Appendix B for extensive examples of topic wordlists and labels.
- Identify the 3 most related novels to each topic, by sorting each excel column (each topic number) in the topic-composition file created by MALLET.
- Having two coders, and if necessary, a third coder for persistent disagreements, allows for inter-coder reliability and less subjective qualitative labeling.

Step 7: Data Analysis & Visualization

- Replace topic numbers with qualitative topic labels (from previous step) in the composition file created by MALLET for clarity; the original composition file is a bit difficult to discern, as it looks like Figure A.3.

Figure A.3 Image of the first 10 Columns of a Topic Composition File

- Implement filters for the row of topic labels in Excel, so that each topic can be sorted by weight.
- Separate novel titles by genre (using the filenames), then average all of the weights of each topic in each novel within a genre, and record.

  Repeat for each genre.
- Sort novels by year of publication (using filenames), then average all of the weights of each topic in each novel within a given year, and record.

  Repeat for each year.
- Visualize datapoints overall, by genre and by year using bar and line graphs.

SENTENCE-LEVEL EXPERIMENT

Step 8: Data Preparation, Sentence-Specific

- In order for sentences to be recognizable after topic modeling, each sentence needs to have (at least) two identifying headers at the start of the sentence: the novel the sentence came from, and what genre the novel belongs to. Headers were applied manually in this study using the Excel drag-down function during the data cleaning step but could also be done computationally.

- In this study, the sentences were separated by genre into 2-3 Excel sheets per genre, for ease of data analysis later. Excel sheets can only accommodate a million and a half rows each, so multiple files are necessary for large datasets.
- For MALLET analysis by sentence, all sentences need to be in a single text file. Save each Excel sheet as a text file in a single folder.
- Use the following command in the Command Prompt to merge all text files in the folder into one merged text file:

  copy *.txt FILENAME.txt

Step 9: Convert Merged Text File to MALLET Format

- In the Command Prompt, again change directory to MALLET, then move the dataset—now a single text file of millions of sentences, each on their own line—to the MALLET directory on your device.
- Using the import-file function this time, convert the single, large text file to a mallet-friendly format using the following command with the data names replaced:

  bin\mallet import-file --input FILENAME.txt —output DATA.mallet --keep-sequence --remove-stopwords --extra-stopwords names.txt
- The remove-stopwords function removed common stopwords like “and” and “the.” The extra-stopwords function can be used if additional stopwords are desired; here, a text file of common names was used to remove character names from the dataset and therefore the topic lists.

Step 10: Optimize the Number of Topics & Topic Model

- Repeat the steps in Step 5 but using the MALLET file created in Step 9.

- If you run into a data size issue, increase the memory cap in the Mallet batch file in the Mallet bin folder, as seen in Figure A.4.

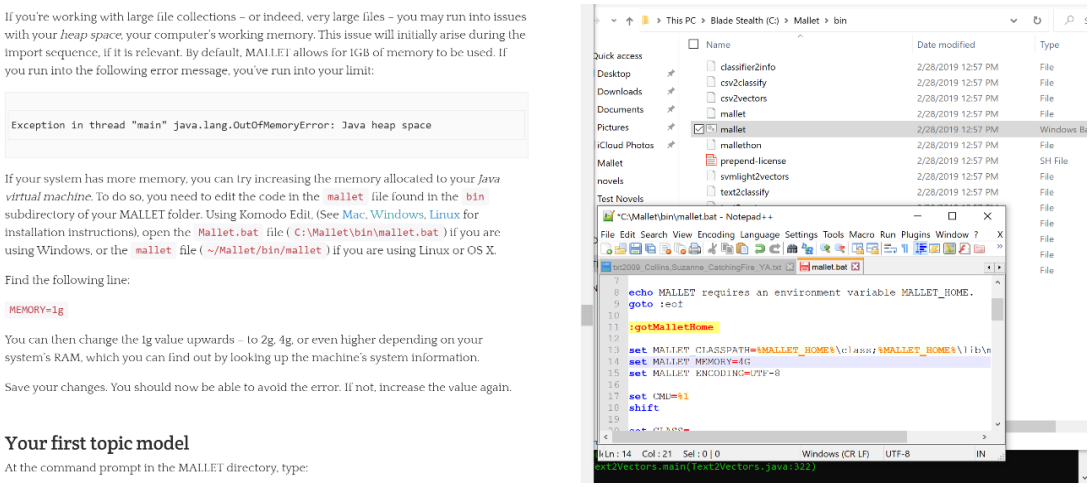


Figure A.4 Image of the Mallet batch file, with Memory Cap Highlighted

Step 11: Interpret Topic Wordlists & Determine Most-Related Sentences

- See Step 6, Section 2.3C and Appendix B for topic labeling.
- However, to determine the most-related sentences, more steps are required at the sentence level. Using a text editor that can handle large file sizes, like EmEditor, select ranges within the topic-composition file (the weights of each topic, for each sentence) and move the data to the Excel sheets with the original sentences.
- Then, the weights can be sorted to find the most-related sentence (rather than the less helpful sentence number, in the original giant topic-composition file).
- The sentences are still scrambled—so the original sentence (in order) has be found in a digital version of each novel using the navigation function.

Step 12: Data Analysis & Visualization

- Replace topic numbers with qualitative topic labels from Step 11 for clarity (the top row of each Excel sheet).
- Average all of the weights of all of the sentences within an Excel sheet for each topic/column, and record. Average these weights per topic with the other Excel sheet(s) for each genre, and record.
- Compare the average weights of each genre, for each topic, and visualize using bar graphs.

APPENDIX B

TOPIC WORDLISTS AND LABELS

Table B.1 Novel-Level Topic Labels and Original Wordlists

Topic Label	Word List
Death	people time dead make kill blood world human kind body face hard bad die stop
Bodily Intimacy	eyes hand body face hands lips felt gaze heart hair head fingers breath arms chest mouth wanted voice skin words
Family & Children	mom dad time school good thing home yeah phone talk make night house car kind feel day lot girl people
City Life	street city apartment money york room hotel office called american morning phone call coffee world building glass man train bar
American Military	man men president sir military war people government general army team security minutes american colonel united russian officer intelligence country
Cars	car man guy door house looked truck big good lot people road put town gun street side left front place
Royalty & Magic	man men king blood lord sword great fire rose father magic black kill queen world power prince dead stone fight
Domestic Space	miss mother mrs house man room woman lady home day morning husband bed girl sister night good father
British Crime	mum bit car round road london house thought realised police looked inspector put tea told day man
Space Travel	ship space earth captain ships system asked time orbit crew planet station suit big mars alien star
Family	father people mother life time knew years day told wanted thought felt world made love house home asked children family
Senses	eyes head hand feel face door hands asks room time voice takes makes hear make side open find turns turn
Sex	f*cking f*ck shit love eyes good mouth hell lips body head sex time mine a*s god hard yeah night hand
Journeys	man stone began returned appeared remained hand continued side turned head rose lay silence passed mind state spoke woman
Digital Tech & Surveillance	world human control system data screen space work level security air inside machine building team access computer wall power information

Environment	water fire trees long time river ground snow day sky boat night sun wind days big light made tree rock
Police	phone police time case call man door told car office asked home found years woman room work house find detective
Military	captain lieutenant system fleet sir ships admiral ship people star commander looked nodded lot
Perception & Communication	looked asked head eyes turned hand door room knew felt time face made nodded pulled stood thought side shook left
Law	judge defense case jury dodge years lawyers court trial
Imagery	face girl dark white light black eyes hair hands air red thing hand floor open glass mouth feet inside long

Table B.2 Sentence-Level Topic Labels and Original Wordlists

Topic Label	Word List
Walking	stairs steps door walked front step past turned side building walk walking stepped floor stood moved forward room hall moving
Temperature, Wind & Sensations	air smell cold hot smoke water fire heat warm scent smelled cool skin felt fresh wind breath cigarette burning blood
Sound	sound heard hear loud noise voice sounds ears voices coming music made silence room laughter quiet footsteps sounded louder scream
War & Intelligence Agencies	men states security team agent army war military american united president russian intelligence special cia force man government knew service
Gun Violence	gun shot fire men man fired rifle weapon guns weapons shoot pistol pointed hit soldiers head firing bullet shots bullets
Thought & Interiority	mind change thought place knew find changed things point time thing safe secret problem subject put matter keeping fact wanted
Metal	metal cut end long iron piece black small steel chain gold made knife wire pieces silver rope held loose tied
Reading & Writing	read book paper reading page written books write letter letters wrote pages writing words note notes found list names desk
Urban Setting	street city park north station west town south train road building east house car drove drive home left place apartment
Drinking	coffee water glass drink bottle table cup wine tea beer set sip poured kitchen drank bar put empty drinking glasses
Goodness, Badness & Judgement	knew wanted thing told thought truth things wrong happened hurt bad good people happen telling felt meant part needed understand
Physical Characterization	hair eyes face dark long black skin brown man red looked white head nose short blond pale gray woman blue
Cellphones	phone call cell number rang message called calls text ring picked bell put left answer button answered pocket checked ear

Physical Intimacy	lips mouth hand fingers kiss tongue mine teeth hands body lip kissed skin cheek feel neck pressed touch hard head
Animals	dog horse dogs horses animals animal wolf ayla cat wild cave bear big camp bird pack people men dead tail
Gendered Characterization	man woman young girl men women looked boy older beautiful girls tall age guy pretty younger lady thought handsome met
Police	police case detective scene found crime murder killer evidence told bosch officer find investigation called asked body victim call dead
Joy	smile face gave expression grin turned lips looked smiled nod laugh made small mouth smiles quick giving surprise smiling half
Driving	car lot truck road front parking pulled driver parked street cars drove side house van turned drive seat driving stop
Light, Color & Sky	light sky sun dark lights blue eyes bright red white black green darkness moon night stars glow window clouds lit
Whispers	voice words hear tone heard spoke word low sounded whisper sound speak speaking calm quiet asked soft ear barely whispered
Time Passing	years ago minutes ten hundred twenty days couple thousand hours times thirty months fifteen half time long seconds past weeks
Government	people government war president world law public political rules power country system family order part military support security great state
Religion	god love man jesus mother father sake hope true good christ lord heart heaven thought church soul angel give great
Sleep	night bed sleep asleep time morning sleeping slept lay hours room woke wake fell awake day lying dream long fall
Travel	ship space power earth system station air plane flight energy crew pilot fuel control small speed orbit gas ships surface
Body Parts	hands arms head shoulders legs chest body face arm neck wrapped knees pulled held put waist tight holding fingers crossed
Law	judge case court trial lawyer witness murder law evidence jury prison defense client attorney state police lawyers charges criminal jail
Stuff You Can Put Stuff In	bag pocket pulled box paper put inside reached hand small handed found open table picked opened desk drawer plastic purse
Home & Country	people city town family place world country lived house home live island small years land york living war south part
Environment	water wind snow boat feet sea ground tree trees leaves sand rain river ice beneath surface waves beach grass rock
Cardiovascular Physiological Responses	breath heart deep throat chest hard breathing felt stomach cleared feel air breathe beat inside slow fast body sigh words
Eating	eat food eating bread plate ate kitchen ice cream chicken dinner eggs bite cheese table bowl made breakfast meat chocolate
Healthcare	brain damage doctor hospital test blood death time medical physical body kind pain doctors mental people patient cancer care

Furniture	sat chair table seat sitting bed leaned side front forward sit desk feet stood edge wall couch floor head window
Digital Technology Use	computer screen camera data video system security cameras set access image code control information images device found radio showed display
Life & Love	life love world people live time wanted day thing man lives part knew rest death loved thought real place person
Rural Setting	road side trees house path small river edge hill buildings fence wall stone view tree woods line building forest high
Work & Doing	things lot people work hard life make thing good thought talk thinking start find working stuff day remember mind
Humans, Aliens & Difference	human world earth people humans life time planet history species technology alien race universe kind body ancient power great mind
Injury	blood face skin hands red left sweat nose wound body tears cut covered mouth broken neck wiped flesh fingers hand
Gaze	looked eye turned caught glanced face glance mirror watch eyes window gaze watching corner stared man shoulder staring watched contact
Death & Families	father mother son dead died man told killed wife brother family sister daughter husband knew death house thought friend found
Mass Media & Information	office news information work press security meeting staff president report media job police story senator public business company called director
Embodied Emotions	felt feel feeling body pain mind fear anger inside thought moment sense suddenly sudden time relief heart guilt panic stomach
Waiting	time long wait stay longer wanted place make run moment find minute move thought hold day give leave needed sit
Eyes	eyes closed face tears looked open gaze mouth opened mine shut wide rolled met close dark blue turned stared narrowed
Explicit Content	Sh*t f*cking f*ck h*ll youre give a*s gonna yeah d*mn stupid b*tch guy man god stop thought thing crazy cr*p
Military	ship captain ships fleet system admiral sir commander enemy command force crew star fire battle officer space navy attack point
Shopping	store food shop people clothes place bought buy stuff found things supplies bags boxes house street art shopping full work
War	kill fight people power knew war men battle death control fighting killed world find thought make man fought time human
Clothes	wearing black shirt white dress wore suit jeans dressed shoes blue jacket wear pair clothes pants boots coat red tshirt
Touch	hand arm held put shoulder fingers reached finger hands left holding grabbed head pulled hold face touched gun raised wrist
Group Social Interaction	started people talking stop stopped laughing girls crowd start walking bar standing turned began men group time crying conversation hear
Family	mother father baby parents told child home children family kids

	girl sister mom knew loved daughter wanted care thought dad
Nonverbal Communication	moment looked silence nodded stared stood long silent paused turned thought waited sat watched man staring waiting stopped watching spoke
Money	money pay paid business cash make bank buy credit company price work worth card give job account lot people cost
Indoor Settings	room living kitchen left house walked hall bedroom bathroom table sitting small waiting office floor dining empty found door rooms
Entertainment	game play playing played music movie song games watch watching show team dance football band ball player time stage star
Time Periods	day morning night time home late work hours early days afternoon tomorrow week hour spent today evening left half good
Body - Violence & Accidents	feet head ground hit floor hard fell forward side wall man body face foot air hand arm dropped hands threw
Doors & Windows	door open opened front closed shut inside doors room walked window stepped lock locked car pushed turned knock heard side
Holidays & Celebrations	night party day time dinner home friends wedding family house tonight birthday happy christmas meet wanted told year nice mother
Settings - Houses	room walls wall floor white windows glass small large table ceiling covered looked painted house stone red window building front
School	school high class year college students student teacher kids professor university day years girls classes grade work friends time art