Magic Vs Evolution*

Or more specifically language used in Harry Potter vs Darwin's Natural Selection

Shreya Sakura Noskor

April 17, 2024

This paper analyses the frequency of occurrence of certain words ("magic" and "natural") in the books Harry Potter and The Prisoner of Azkaban and Charles Darwin's Book of Evolution. Using the data provided by Gutenburg's project and Internet Archive, we were able to find the related data and analyze whether a topic of the book had any effect on the number of occurrences of each word in each text. The data found here will be related to a very important topic in Machine Learning; clustering. This is when a machine is able to see the content of data, and group them together based on what it sees.

Table of contents

*Code and data are available at: HarryDarwin

1 Introduction

This paper looks at 2 completely different texts in terms of themes; Charles Darwin's Books of Evolution (Darwin 1859) and Harry Potter and the Prisoner of Azkaban (Rowling 1999). The book of Evolution, Origin of Species by Charles Darwin was a book that came out in 1859, and it revolutionalized evolutional biology. Darwin described in this book the process of evolution by natural selection where all animals change some sort of aspect of themselves (either physical quality or behavior), that helps them survive in the world longer. The process of change throughout the million years is called evolution and natural selection is the idea that animals with desirable traits survive longer in nature. Some examples of this include the beaks of birds where food is limited or the color of fur to attract mates. Overall this non-fiction book changed the scientific world's view on how animals (including humans) work. It is a theory, but it is supported by millions of years of evidence found in fossil fuels. Harry Potter and the Prisoner of Azkaban on the other hand is a completely different book. This is a fiction book written by J.K. Rowling, that highlights the adventures of a young wizard named Harry Potter. This is the third book in the series and the reason why this was chosen specifically was because of the complete contrast in themes. Also, it is one of my favorite books to read at the time.

Now coming to the main topic of the paper. There is 2 questions that this paper aims to answer: "How do the occurrences of "natural" and "magic" differ in terms of frequency and context between Darwin's scientific work and Rowling's fantasy novel?" and "How does this explain clustering in computer science?". The analysis section will answer the first question by showing various graphs and models but the discussion section will go more in-depth towards the second question. The estimand is how often the selected words appear in each of our texts. Additionally, the purpose of this paper is not argumentative nor is it to convince the reader of some hidden analysis found in our data. The goal of this is to help the reader understand how clustering in computer science works and how it may relate to the statistical models we look at here.

Language is a key aspect of everyday life. The vocabulary we use can often affect the theme and tone that we are aiming for when communicating with our audience. However, to an average reader, this may not seem like a very important topic that needs to be investigated. This is understandable as what insights can we gain from just observing the frequency of the occurrences of such thematic words; all we can tell from it is that the 2 texts have contrasting themes which we just know from the context of the title or the summary. However, such analysis is important because it can help understand a much higher level idea used in Computer Science called clustering. Clustering is the concept in Machine Learning where the computer needs to be able to group data in categories based on a feature that the data has. This is very helpful as usually the data is unsupervised (meaning there is no label/group associated with them) so this process helps find similarities. This paper finds that with each text the theme of the text greatly impacts the occurrences of the words "magic" and "natural", which is expected but this will help motivate the concept of clustering.

This report is structured like so: Data section describing the data and the variables inspected, Model section showcasing the Regression Model used to perform the analysis. The result section shows the results of the model analysis as well as a summary for each model and lastly the Discussion section will go in-depth about what we see in the Data and Results section as well as what that means in terms of the concept we are explaining today; clustering. The Discussion section will also state possible limitations in our data.

2 Data

2.1 Source

The data utilized was from Project Gutenberg and PDF Drives and with the help of R (R Core Team 2023) we were able to create this paper. Also, the code for making the models was made referencing Telling Stories by Rohan Alexander(Wickham et al. 2019a). Other R packages were used to clean, process, and model the data such as Wickham et al. (2019b), Johnston and Robinson (2023), Goodrich et al. (2022), Ooms (2023), Wickham et al. (2023), Richardson et al. (2024), Wickham (2023), Arel-Bundock (2024).

2.2 Variables and Measurement

The book of Evolution by Charles Darwin (Darwin 1859) is collected from Project Gutenberg (Johnston and Robinson 2023), which is a source known for its reliability in offering free EBooks as it collects the texts from the source. Harry Potter and the Prisoner of Azkaban by J.K. Rowling was collected from the Internet Archive (Archive 2022). Internet Archive is a very large platform, however, with millions of users who are able to upload anything they want. To combat this slightly, I compared a couple of pages from the Internet Archive to my own personal copy, and I was able to see that there are no changes in terms of content. I extracted the lines from each of the text and counted up the occurrence of each variable and used that as my form of measurement. The reason why I chose this as my data set was because it perfectly displays what I want to show; how the difference in topic affects the vocabulary and therefore aiding machines to categorizing them. Now for the variables. We first cleaned each text of any leading white spaces that would hinder our analysis and broke it down into sentences. Then I cleaned it up so that there are 4 major columns: text, natural_count, magic_count, and word_count. This was done to both texts. The "text" column shows the actual lines of the book. The `natural_count` shows the number of times the word "natural", or "species" appeared in that line. Throughout this paper we will refer to it as the word "natural", to not confuse the reader and also to keep things concise. But keep in mind that we are looking at both of these values because they are similar to each other, especially in the context they are used in the Book of Evolution (Darwin 1859). Next is the `magic_count` which takes a look at the number of times the word "magic", "miracle" and "wizard" shows up in each book. Again, in this paper, we will refer to this as the number of times the word "magic" appears due to the overlapping theme of the words and to not confuse the reader. Lastly the variable `word_count` just counts the amount of lines in that word. This helps find important info like averages and predict if larger sentences have a higher frequency of each of the words.

2.2.1 Distribution of each word with each text

This section will briefly look at the correlation between the words and each individual text.

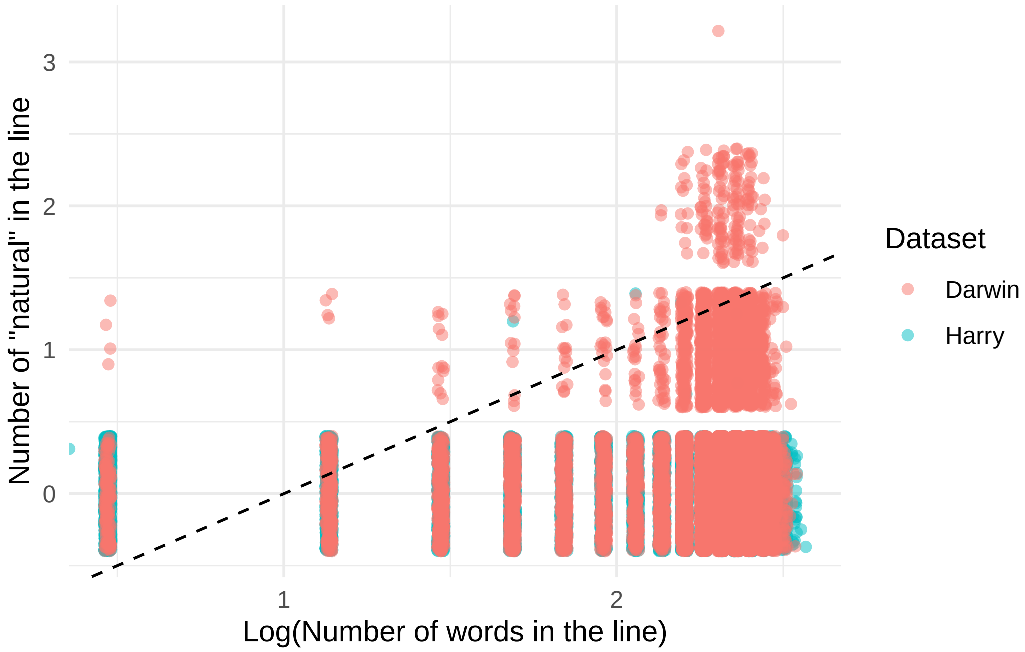


Figure 1: Comparison of number of "nature"-realted words and the number of words in a line

Figure 1 shows the number of time the word "natural" or "species" show up in each line. Due to the fact that both books are not small -it has around 300-600 pages-, what we see is that there is a lot of data points. The dashed line is $y = x$ and it shows that if there were points on that line it means there is an average of 1 occurrence of the words per line. However for the most part we see that it is is below it meaning that not every line contains the words “natural” or “species”. The bottom block shows that exactly where majority of the red and blue dots are concentrated around 0 meaning most lines in the text do not have these words. Additionally the faint blue shows that there is even fewer lines in Harry Potter that contain the word natural. Lastly the fact that some of the red dots are above the $y = x$ line means that there are lines with more than one occurrence of the words “natural” and “species”.

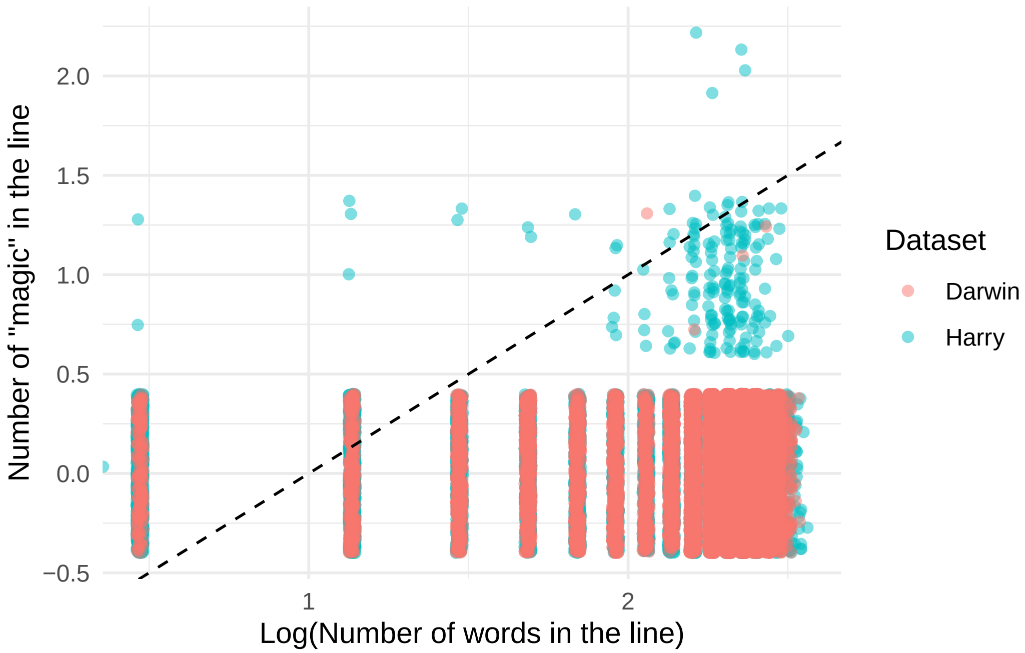


Figure 2: Comparison of number of “magic”-realted words and the number of words in a line

Figure 2 on the other hand shows the number of time the word “magic”, “miracle” or “wizard” show up in each line. The dashed line is $y = x$ and it shows that if there were points on that line it means there is an average of 1 occurrence of the words per line. However for the most part we see that it is is below it meaning that not every line contains the words “magic”, “miracle” or “wizard” . The bottom block shows that exactly where majority of the red and blue dots are concentrated around 0 meaning most lines in the text do not have these words. Additionally the faint red shows that there is even fewer lines in the Book Of Evolution that contain the word “magic”. Lastly the fact that some of the blue dots are above the $y = x$ line means that there are lines with more than one occurrence of the words “magic”, “miracle” or “wizard” .

3 Model

We use a Poisson Regression model to show the correlation between the number fo words in a text and the number of times the word “magic”, “miracle” or “wizard” appears or “natural” and “species” appear.

3.1 Natural and the 2 texts

Define y_i is the number of times “natural” appeared in each text and the explanatory variable is the number of words in the line. This means that we have 4 models in total with the y_i’s being, number of times the word “natural” and “species” showed up in The book of Evolution or Harry Potter and the Prisoner of Azkaban. We predict to see that there is a positive correlation in Darwin’s text but not in Harry Potter due to the difference in topics.

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{Number of Words}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
& && (5)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

I predict to see that there is a positive correlation between the words “natural” and “species” in the Book of Evolution but a negative one in Harry Potter. This is due to the fact that the main topic of Charles Darwin’s book is the concept of natural selection affecting different species. There are very limited ways to describe that process without using thoes 2 specific words. On the other had Harry Potter is a fictitious book that has very little to do with species and even little to do with nature. Hense in whatever context they do appear, it was not in the way it is used in Darwin’s book.

3.2 “Magic” and the 2 texts

Define y_i is the number of times “magic”, “miracle” or “wizard” appeared in the text and the explanatory variable is the number of words in the line.

$$\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (6)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{Number of Words}_i && (7)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (8)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (9)\\
& && (10)
\end{aligned}$$

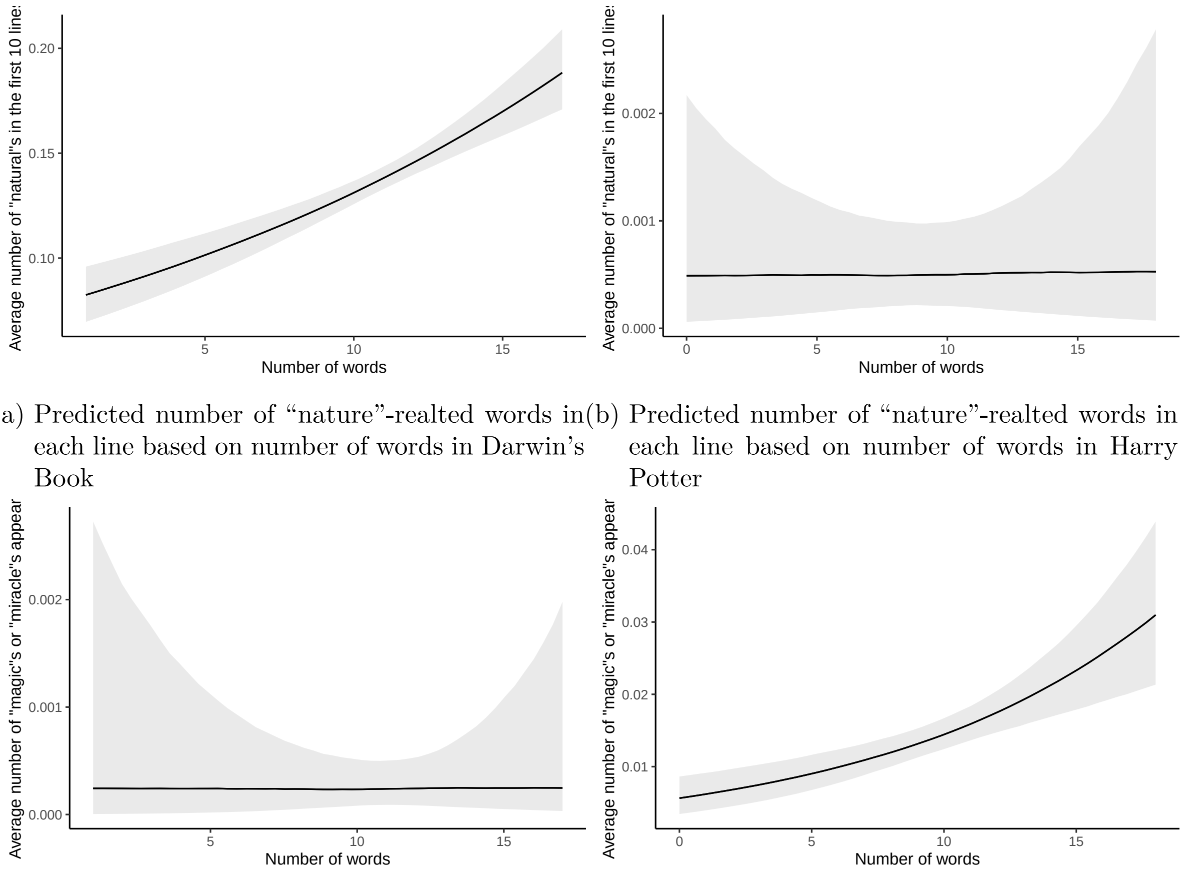
We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.2.1 Model justification

I predict the opposite for this case. I predict to see that there is a positive correlation between the words “magic”, “miracle” or “wizard” in Harry Potter but a negative one in the Book of Evolution. This is, again, due to the fact that the main topic of Harry Potter is that a young wizard is finding his way in the life of magic. On the other had the Book of Evolution is a non-fictitious book that has very little to do with magic and even less with wizardry. Again the theme of the book decides the correlation between the varibales.

4 Results

In this section we see the very thing we predicted. Firstly, Figure 3a and Table 1a model frequency of “nature” in the Book of Evolution. Figure 3b and Table 1b model frequency of “nature” in Harry Potter and Figure 3c and Table 1c model frequency of “magic” in the Book of Evolution. Lastly, Figure 3d and Table 1d model frequency of “magic” in Harry Potter. We see that our prediction about Figure 3a is true as we see a positive correlation. Same with Figure 3d. However in Figure 3b and Figure 3c, instead of negative correlation we see there is zero correlation which makes sense as there is no connection between the words and their respective text.



(a) Predicted number of “nature”-realted words in each line based on number of words in Darwin’s Book

(b) Predicted number of “nature”-realted words in each line based on number of words in Harry Potter

(c) Predicted number of “magic”-realted words in each line based on number of words in Darwin’s Book

(d) Predicted number of “magic”-realted words in each line based on number of words in Harry Potter

Figure 3: Predicted plots for each book

4.1 Prediction Plots

4.2 Model Summary

Table 1: Model Summaries

(a) Model Summary showcasing the correlation coeffcient for Darwin's evolution book and the word nature

	darwin and nature
(Intercept)	−2.55
	(0.09)
word_count	0.05
	(0.01)
Num.Obs.	19 411
Log.Lik.	−8095.192
ELPD	−8096.9
ELPD s.e.	106.3
LOOIC	16 193.7
LOOIC s.e.	212.7
WAIC	16 193.7
RMSE	0.37

(b) Model Summary showcasing the correlation coeffcient for Harry Potter and The Prisoner of Azkaban and the word nature

	harry and nature
(Intercept)	−7.62
	(0.93)
word_count	0.00
	(0.10)
Num.Obs.	12 767
Log.Lik.	−52.022
ELPD	−53.6
ELPD s.e.	19.0
LOOIC	107.2
LOOIC s.e.	38.0
WAIC	107.2
RMSE	0.02

(c) Model Summary showcasing the correlation coeffcient for Darwin's evolution book and the word magic

	darwin and magic
(Intercept)	−8.31
	(1.78)
word_count	0.00
	(0.16)
Num.Obs.	19 411
Log.Lik.	−38.004
ELPD	−39.7
ELPD s.e.	17.2
LOOIC	79.4
LOOIC s.e.	34.4
WAIC	79.4
RMSE	0.01

(d) Model Summary showcasing the correlation coeffcient for Harry Potter and The Prisoner of Azkaban and the word magic

	harry and magic
(Intercept)	−5.18
	(0.22)
word_count	0.09
	(0.02)
Num.Obs.	12 767
Log.Lik.	−938.888
ELPD	−940.6
ELPD s.e.	57.8
LOOIC	1881.1
LOOIC s.e.	115.7
WAIC	1881.1
RMSE	0.12

5 Discussion

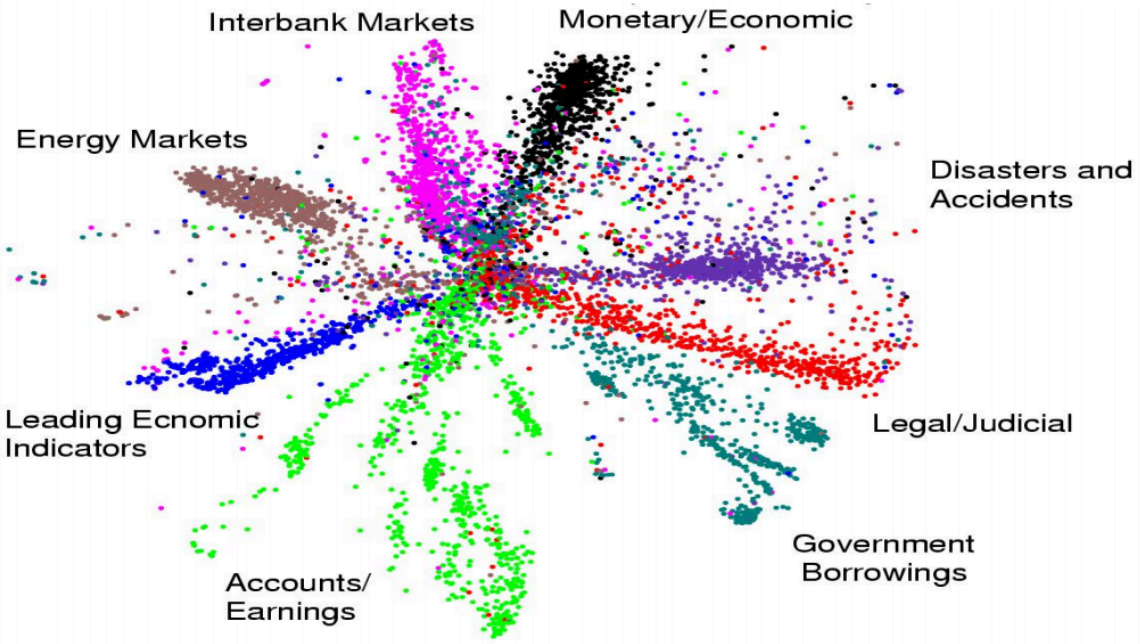
5.1 Finding and Analysis of Data

From the previous sections we see that there is a positive correlation between the number of times the word “natural” or “species” appears in Charles Darwin’s book on the Origin of Species Table 1a. The correlation coefficient is in fact 0.05 meaning that when the number of words increases by 1, there is a 5% more chance that we will see the word nature again. This may not seem much but compared to our other text Table 1b it is a lot of information as the correlation factor is 0 meaning we have no way of knowing if the next word will be “nature” or “species” or neither. Similarly, with Table 1d, we see that the correlation coefficient is 0.09 meaning that with each increase in word count, we see a 9% more chance that the word “magic”, “miracle” or “wizard” appears (contrary to Table 1c with Darwin’s book). This means that when the book is fictitious, it is more likely to see the words “magic”, “miracle” or “wizard” (Like Harry Potter (Rowling 1999)) and when it is non-fiction you are more likely to see the words “natural” or “species” (Like The Book of Evolution (Darwin 1859))

5.2 Importance to ML

Now to the driving and main point of the paper: how does all of this analysis relate to clustering? As explained previously we perform clustering on data sets that have no labels. Meaning we have a bunch of data and we don’t know how to begin to categorize them. The easiest way to motivate this is by thinking of online libraries. When reading a book you are often told what the similar topics are and what similar books are as well. Although it may be likely this information is gathered through people reading and providing their analysis this is not very efficient. We must compensate the readers as well as compile the data. We must also assume that the data provided by the user is true as cross-checking will require more resources. Then we must hope that we have enough data to begin categorizing the books. When it requires this much effort to do something by hand we always try to find a much simpler solution with computers. The solution answers our second research question: clustering. Clustering using the help of the model analysis we did earlier will take away all the inefficiencies of doing these things by hand. By training a computer we can first feed it all the books we want to categorize. Then the computer will read through each of the books and calculate the correlation coefficient between the number of words and some keywords that we chose. In this paper the keywords happened to be “natural” and “species” or “magic”, “miracle” and “wizard”. With this, they are able to somewhat tell the theme of the book. If we were to feed it another Harry Potter book the correlation coefficient of how often the words “magic”, “miracle” and “wizard” came up may be close to the coefficient for Harry Potter and the Prisoner of Azkaban. The computer is able to tell that the 2 books are related to each other, at least in terms of theme. We are then able to categorize the books based on topics. This is a very simple case however. In reality, the people running this model must first optimize the perfect group of words that allows us to group the data.

The finished product may look something like this where each dot is a text and we are able to identify what each of the different categories are (image from Krishnan and Gao (2022)):



Now this is the very general and basic idea of clustering using Poisson regression. I am unsure if this model has been applied yet (as there are multiple other ways to categorize such data) but, there does seem to be evidence supporting that this process is used in practice or theory by this paper written in the Journal of Econometrics (MacKinnon, Nielsen, and Webb 2023).

5.3 Weaknesses and next steps

Not all papers are perfect and neither is mine. Some limitations that can be included is that when copying each line of the textbooks, the cleaning process I used also left the first couple of index pages which include information about the authors and the chapters. This may cause a slight skew in our data as the content of the book does not include unnecessary information. The next weakness is that I am not an expert in Machine Learning algorithms. I used what I know about correlation coefficients in statistics and the concept of clustering from computer science and drew a connection to help the reader understand how the concept works. With all of this in mind, the next steps are to clean the data of any other external content that does not add to the theme of the book as well as get experts on clustering to explain the process better. While on that road exploring the frequencies of other words may also help narrow down the themes of the text as we don't always have one theme related to one book.

5.4 Concluding Remarks

Statistics and Computer Science are very broad subjects in the worlds. They are also very useful so it is even better to combine them together to accomplish a task more efficiently. This paper shows one of those intersections and is able to explain it to the user using an analogy. So to answer the initial 2 research question, yes there is a correlation between a text and key thematic words and that analysis is transferable to machine learning, to help categorize data. Although there is some limitations of this paper but the key ideas still apply.

References

Archive, Internet. 2022. "Harry Potter and the Prisoner of Azkaban." 2022. https://ia902505.us.archive.org/26/items/Lindas-bookshelf/Hugo%202000%20Nominee%20Novel%20-%20J.%20K.%20Rowling%20-%20Harry%20Potter%20and%20the%20Prisoner%20of%20Azkaban.pdf.

Arel-Bundock, Vincent. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Darwin, Charles. 1859. *On the Origin of Species*. London: Murray.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

Krishnan, Rahul, and Alice Gao. 2022. "CSC 311 Fall 2022: Introduction to Machine Learning." https://www.cs.toronto.edu/ rahulgk/courses/csc311_f22/logo.png.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. 2023. "Testing for the Appropriate Level of Clustering in Linear Regression Models." *Journal of Econometrics* 235 (2): 2027–56. https://doi.org/https://doi.org/10.1016/j.jeconom.2023.03.005.

Ooms, Jeroen. 2023. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Rowling, J. K. 1999. *Harry Potter and the Prisoner of Azkaban*. London: Bloomsbury.

Wickham, Hadley. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.