Final Report (milestone_03)

Denitsa Vasileva & Lucy Mosquera

2020-03-15

Contents

1 Introduction

1.1 Motivation

Project Gutenberg is a free online source which provides free access to more than 60,000 books- mostly classics @project_gutenberg. Its purpose is to crete digital copies of books in the public domain and thus make them more accessible and foster a love of reading to new generations of readers. The Gutenberg project stores troves of information about each available book- including both metadata about the author and work itself as well as popularity, difficulty and readibility metrics for each book.

1.2 Research Questions

1. Explore the changes in the prevalent sentiments and subjects in books in Project Gutenberg change based on publication year?
2. Explore the changes in the prevalent sentiments and subjects in books in Project Gutenberg change based on author gender, location,etc.?
3. Explore what characteristics are associated with an increase in a book's popularity for download on Project Gutenberg. Potential variables to include are publication year, length, formats the book is available in, subject matter, and reading difficulty.

2 Exploratory Data Analysis

2.1 What are the names of the columns (i.e. the variables we can use in our analysis)?

The variables have the following names:

congress.classifications, languages, subjects, title, type, downloads, id, rank, url, author.birth, author.death, author.name, publication.day, publication.full, publication.month, publication.month.name, publication.year, formats.total, formats.types, automated.readability.index, coleman.liau.index, dale.chall.readability.score, difficult.words, flesch.kincaid.grade, flesch.reading.ease, gunning.fog, linsear.write.formula, smog.index, polarity, subjectivity, average.letter.per.word, average.sentence.length, average.sentence.per.word, characters, polysyllables, sentences, syllables, words, language.en, language.de, language.es, language.fr, language.it, language.la, language.nl, language.pt, language.ru, language.tl

2.2 What is the distribution of years of publication for the books in Project Gutenberg?

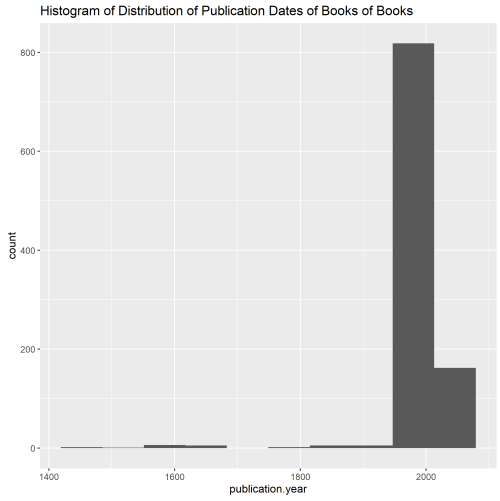

Figure 1: Histogram of publication year for all books on Project Guttenberg.

Figure 1 shows the publication dates for the 1006 books present in the Project Guttenberg dataset. From this figure we can see that there are some older books, and then a large cluster of books from the 2000's.

2.3 What are the correlations between key variables of reading difficulty and text structure?

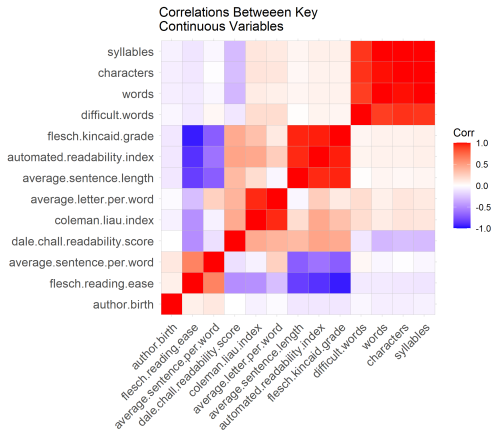

Figure 2: Correlogram of the correlation between key variables

Figure 2 shows the correlation between different key variables that describe the books as well as characterize their difficulty using metrics. In this figure we can see that there fairly high positive correlation between the number of words, number of characters, number of syllables, and number of hard words in a book as well as between the average sentence length, automated readability index, and Flesch Kincaid grade. There are also reasonably strong negative correlations between average sentence length, automated readability index, Flesch Kincaid grade and the average number of sentences per word and the Flesch reading ease index.

2.4 What is the relationship between year of publication and reading difficulty as measured by the automatic readability index?

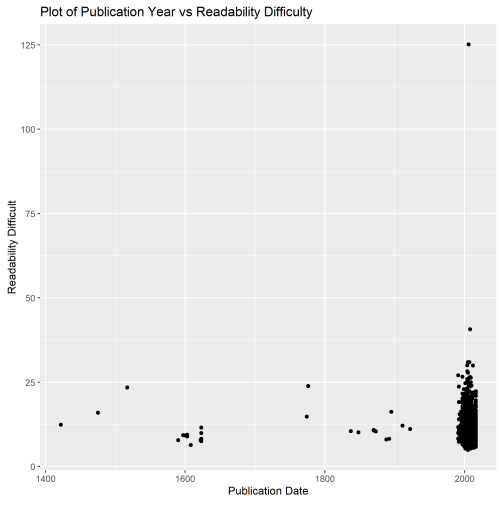

Figure 3: Scatter plot of readability difficulty vs publication year.

Figure 3 plots the readability score against the year of publication. This graph does not show any significant difference in the level of difficulty in books based on Publication Year.

Table 1: Coefficient estimates, standard error, test statistics, and p-value for linear regression of year of publication predicting sentiment index values using the Project Guttenberg data.

term	estimate	std.error	statistic	p.value
(Intercept)	0.3307099	0.0719379	4.597159	0.0000048
data$publication.year	-0.0001151	0.0000360	-3.197866	0.0014278

2.5 Sentiment Analysis

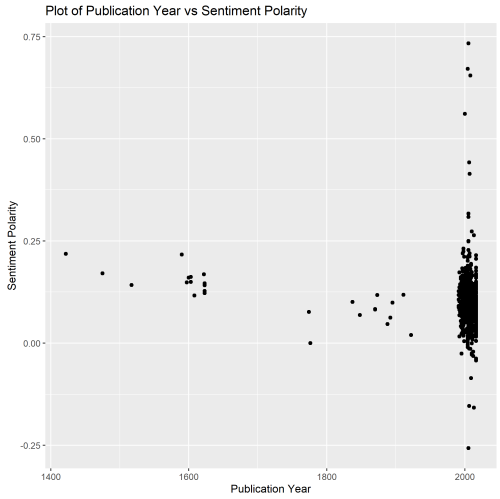

Figure 4: Scatterplot of sentiment vs publication year

Provided in the Project Gutenberg data is a sentiment polarity score that aims to quantify the positivity or negativity of a work in general. Figure 4 shows these sentiment scores against the publication year. This plot does not show a relationship between publication year and sentiment polarity.

3 Analysis

In order to answer our research questions, we used linear regression.

To answer the first question, we implemented a linear regression model on the sentimentality index of each book compared to the publication year.

4 Results

The results of the linear model of the year of publication vs the sentimentality index for each book can be seen in Table 1.

The scatter plot of sentimentality index vs year with the resulting linear regression superimposed can be seen in Figure 5.

This regression shows tha t the publication year is a significant predictor of the sentimentality index (p value = 0.0014). The coefficient for the publication year is a negative value, indicating that newer books are more negative in their general sentiment than older books (coeff = 10^{-4}), although the magnitude of this coefficient is quite small indicating a small but significant effect.

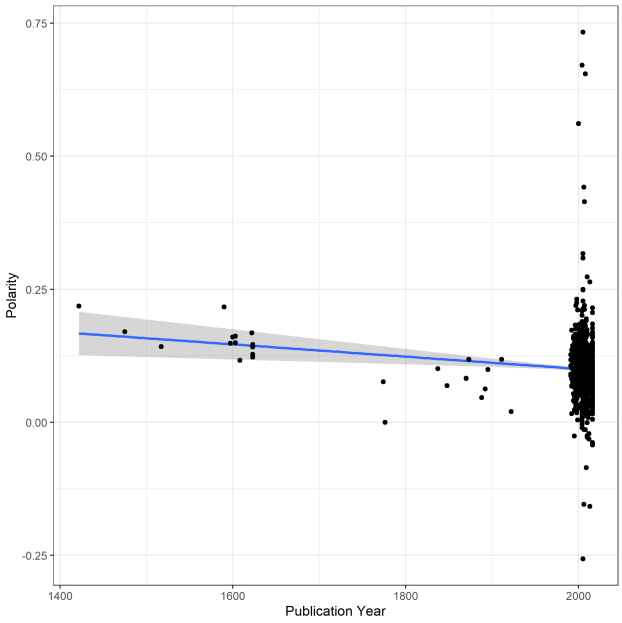

Figure 5: Scatterplot of year vs polarity.

5 Discussion and Conclusions

Throughout this analysis we have explored the Project Guttenberg classics dataset and gained valuable insights. We have shown that the sentiment or overall tone of newer books is marginally more negative than older books.