# Description of the corpus on romance novel’s affective qualities

1. How did we collect the data

First the data for this research paper has been collected using API for the Guardian newspaper, and was intensions on researching the difference of stigmatization on romance novels due to its audience being women’s and being viewed as undervalued in literature qualities or of any cultural benefits to its readers.

The corpus started with 1130 observations and after the cleaning and pre-processing, the amount was reduced to 52 observations, due to subletting the review options in the headline of the document, and only including the books as the section id. Unfortunately, the code acquired from The Guardian API and processed didn’t answer our research questions about Romance novels stigmatization, even after cleaning the main key words, and removing most frequent tokens still there was not much left to be understood about this social phenomena on Romance Novels that are viewed as poor written, as shown on figure 1.

Figure

Furthermore, GoodReads books reviews API was selected to be tried to use to acquire the data on books reviews, but unfortunately the website no longer issues Developer keys, and it is required on download xml data from their API. Another route was taken within web mining Best Romance novels reviews for 2012 and 2022, and two main authors were E.L.James and Collen Hoover. This step was much longer than anticipated due to changes on RSelenium and different packages that assist within the web scrapping process. An API was not used due to the website not providing it and need extra private amazon API keys. Once solved the issues with the remote access to the page, gathering the corpus data with RSelenium to web scraping with javascrip, and using Rvest to read html, wdman to manage selenium, and magriti plus tidyverse to use pipes %>%.

The source of the corpus was the GoodReads website of books reviews and general public opinions. Goodreads is the world’s largest site for readers and book recommendations. Their goal is to assist people in flinging and share books they love.

The corpus was divided between 2012 and 2022, a total of 10 years difference and a lot of social happenings during those time. For example the advance of kindle as a reading device making the readers less judgment while reading certain contents in books, since others wouldn’t be able to view it’s content as per normal books, which covers are shown. The covid-19 pandemic and its tool on social engagement and society structures in terms of work and ways of having fun changed due to not being able to attend outside places for a while and less face to face arrangements.

For the GoodReads data, it was split between the most popular romance book of 2012, by E.L.James - Fifty Shades of Grey and for 2022 Collen Hoover – Reminders of him.

After scraping reviews from Goodreads, than it started the exploratory data analysis to get a better sense of the data, and pre-processing. The values acquired from the website reviews contains five variables:

1. X: item value
2. Book: book title
3. Reviewer: it was pulled as NA due to data privacy to the name of the profile.
4. Rating: text with the review
5. Review: it seemed to be the status on the profile of the reviewer.

Not all were relevant to the research questions, so the only column that was kept was the rating review and item value id. There was 3564 observations under ratings for E.L.James.

Within pre-processing the data was extracting the reviews from the page, cleaning the reviews with regex, populating a data frame with the relevant reviews, trying to isolating the ratings, but due to changes in the java scrip, this steps was left undone, removing duplicated reviews and saving the html file as csv. (Attach in the appendix code)

Chart, histogram

Description automatically generatedOnce these steps were taken, there was still cleaning of the text and removing unwanted symbols, punctuation, stop words, and high frequent words that provided no meaning.

## Discussion of the method or methods used to analyse the corpus under the study

Our first method was to used was text mining the distribution of the ratings per reviews, to gather any understanding, but it was not a good source of understanding. So, the sentiment analysis with positive and negative words to understand the emotional intention of the document was the next step taken to comprehend the document.

It started by creating a document-feature-matrix and getting the top features, and the distribution of ratings, to better understand the data and to get a sense of whether this data would be able to answer some of the topics. Once that was done, the sentiment plot was acquired to understand the different words between negative and positive as shown in figure 3, 4, 5 and 6.

Figure

Chart, bar chart

Description automatically generated



Within the most frequent reviews, and it portrays for both E.L.James and Collen very different sentiment within the negative aspects, but similar on positive.

Figure 3 shows the first most negative word and positive, and fiction is one of those, so to make any inferences, the key word in context was used to understand how the negative text was being formed as shown below.

Text

Description automatically generated

It could also be interesting to see more emotions classifications and sentiment score to better understand the difference between certain types of words, if there would be any anomaly in the sentiment analysis and to prove our hypothesis on the bad perception on romance novels over the last decade.