The analysis of 1984 in 11 different languages

My topic for this project was the reading cleaning and analysis of 11 different language version of George Orwell’s book 1984. I chose this topic because I was interested in the length and content of books in different languages.

My hypothesis was that the most common letter in Latin languages is e. I used 1984 as a research medium to investigate my hypothesis. I used 9 different Latin-based alphabet books and also Russian and Chinese. I retrieved the length of the book, length of the book cleaned without all the punctuation marks, average length of words, most common character and most common word.

My chosen dataset is George Orwell's book 1984. I specifically chose this book because I did research on which books where the most translated worldwide, it was close to the top and I have already read the book, so I knew the novel. I collected these books by searching on the internet. I verified all the books whether they were the actual book or not by translating parts of it to match up with the English version.

First, I needed to decide on a book to analyze, collect the book in the different languages, code the python script for the data cleaning and extracting, code a visualizer, write the report and design the website.

Final results

A lot of the parts of the code worked out as expected. I was able to determine the lengths of the books get the most common characters and words. In my original plan I also wanted to find the average word length and this worked out perfectly except for the russian version which gave an answer that was way off of 48 word length. The biggest challenge of my project was the cleaning of the data as regular expressions were a bit hard to figure out but it worked in the end, although it is slow for some reason as the disk goes to 100%. Another challenge that arose was writing all my data about the books into a csv file so that they were labeled correctly and would be able to be extracted by the graph making script.