The aim of this project was to:

1) explore how entropy changes based on the literary genre
2) compare entropy values between two authors
3) estimate the information rate for Polish

Datasets

In total I used 4 datasets:

- all of Adam Mickiewicz's works available on wolnelektury.pl
- the same for Juliusz Słowacki
- a dataset of 36k Polish tweets gathered over a year (from my Research Lab)
- chrisociepa/wikipedia-pl-20230401 · Datasets at Hugging Face

NOTE: I used a subset of the last two datasets, to match the amount of data for Mickiewicz and Słowacki. When tokenizing the text by characters (so letters + spaces) and by words, for Mickiewicz I got respectively 1M and 170k tokens, and for Słowacki: 2.3M and 384k tokens. So, for the Twitter and Wikipedia data, I took the first 2M letter- and 300k word- tokens.

1) Entropy in different genres - Mickiewicz

I divided the works of Mickiewicz into 4 categories: epic, drama, poetry and syncretic (which consisted entirely of ballads). [more information about data gathering and properties is given in the presentation and the notebook]

Entropy on the character level did not differ between categories (4.46 [bits] for the ballads, 4.48 for the rest). This result seems intuitive: the subset of letters remains fixed and the overall language structure (stable letter frequency) is not sensitive to the genre.

There were differences within the word-level entropy. The highest entropy (11.91) was obtained for epic literature - where the sentences are long and the vocabulary varied. Drama pieces had lower entropy (11.39) - as the text is divided into dialogue parts, with names of characters and certain phrases recurring. For poetry, the entropy was lower still (11.16) - with the rhythm and structures of poems reducing word diversity. Ballads had the lowest entropy (10.57) - as a syncretic genre, they combine the repetitive format of dramas and poems.

The entropy calculated on words from all of the works combined was 12.10. It is higher than the entropies above, because each genre might have specific words, which are frequently used. But in the combined corpus, the pool of words is more diverse (less predictable).

2) Entropy between two authors

In order to check whether the observed entropy relationships were exclusive to Mickiewicz, I compared his works to those of an author from a similar time period - Juliusz Słowacki. Since he did not write ballads, only the first 3 literary genres were contrasted.

Entropy calculated on letters differed slightly between genres, but the one from all Słowacki's works combined was 4.48 - identical for Mickiewicz (for the reasons given above).

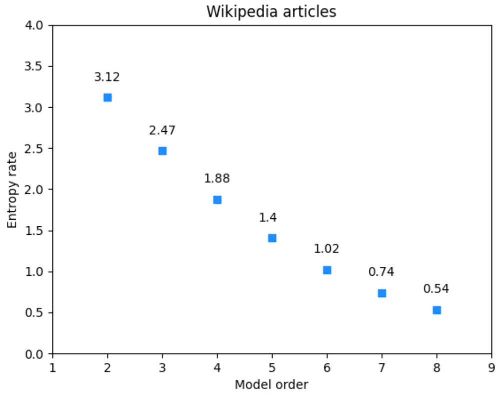
When it comes to entropy obtained for words, the highest one was for poetry (11.73) - while it was the lowest for Mickiewicz. That is because, 11 out of 81 Słowacki's works in this genre, were instances of narrative poetry. The text is written in metered verse, but does not necessarily have many rhymes and repetitions, as in standard poems. This quality is nicely reflected in the higher entropy.
The relationship between epic and drama was the same as before - that is, the entropy for the former was higher than for the latter.

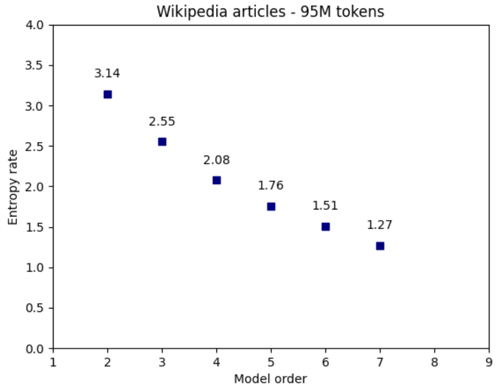
3) Information rate estimation

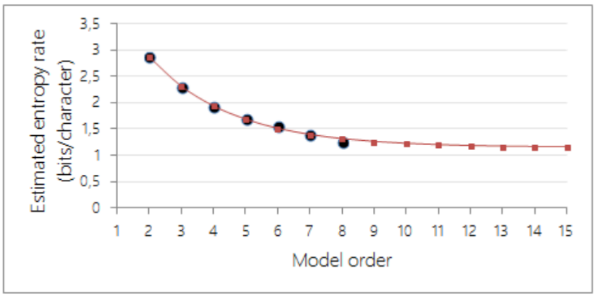
Estimating the information rate was based on this article: An experimental estimation of the entropy of English, in 50 lines of Python code | Code crumbs, by Clément Pit-Claudel, which utilizes simple n-order Markov models.
The unit of source entropy here is bits per character, so letters + spaces, which were left so that, e.g. for ("t", "h", "e"): 4-grams of "they", "them", "thei[r]" will be allowed, as well as "the ".

The plots generated for the 4 datasets [in the presentation] look very similar. Although the number of characters varied - 46 for Mickiewicz, 43 for Słowacki, 36 in tweets, and 125 in Wikipedia (letters from languages other than Polish) - the general trend was not skewed by this. This striking similarity is not what I had hoped to find, considering high literature from 200 years ago was compared with specific social media text and scientifically written entries. Moreover, a quick analysis of the most common words showed that they are shared across the corpora ("i", "się", "nie", "na", "w",...). However, the amount of data is certainly too small for entropy estimation in bits per word.

The importance of having more data can be seen in the plots. The one for Mickiewicz has the steepest slope - as there are half as many tokens as in the other datasets.
Also, when I ran the estimation on 95M tokens [as in the article] from Wikipedia data, the plot looked even more similar to the one for English. Although, the entropy rate still seems to fall too fast with bigger model orders.







4) Additional: Jensen-Shannon distance

To use this metric, the probability arrays have to be of the same size. So, for each pair of texts, I first extracted the common letters/words, and only used them.

On the character level, corpora were almost entirely preserved, when letters not shared by the pair were excluded (99.92-100% tokens were left). The distances were small: on average 0.02 in Mickiewicz-Słowacki (depending on the genre); around 0.04 in both Mickiewicz-Twitter and Słowacki-Twitter; 0.05 in those pairs but with Wikipedia; and 0.03 in Twitter-Wikipedia.

Once again, this could show the stability of letter frequencies in a language, disregarding the text's source and characteristics.

On the word level, the number of tokens substantially diminished.

In the Mickiewicz-Słowacki pair, 72-86% tokens for Mickiewicz remained, and 65-82% for Słowacki. The biggest distance of 0.16 was obtained for their poetry - perhaps again illustrating the fact that Słowacki wrote narrative poems, and Mickiewicz did not. The Jensen-Shannon distance for drama was also bigger: 0.1. For epic it was 0.04, and 0.06 for all of their works combined.

Compared to tweets, the distance for Mickiewicz (0.09) was higher than for Słowacki (0.06), with more tokens retained for the authors than for the tweets (73% Mickiewicz - 65% tweets; and 74% Słowacki - 70% tweets).

For Wikipedia, the relationship was reversed - 0.09 for Słowacki (68% tokens - 58% for Wikipedia); and 0.07 for Mickiewicz (67% tokens - 50% for Wikipedia).

The distance between Twitter and Wikipedia data was 0.12 (76% Twitter and 71% Wikipedia tokens remained).

This analysis is completely additional, as it does not seem very informative. The characteristic words for each dataset were removed, as they were not shared. And probably as a result of that, the distance values are very small and not reflective of the true underlying relationships between the texts.