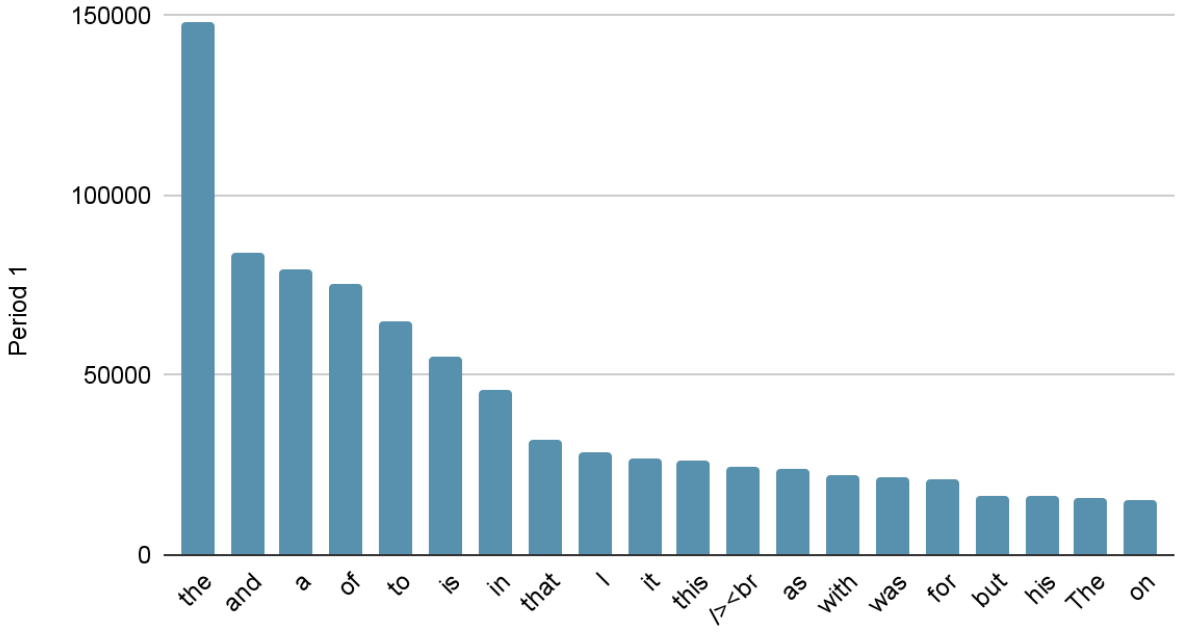
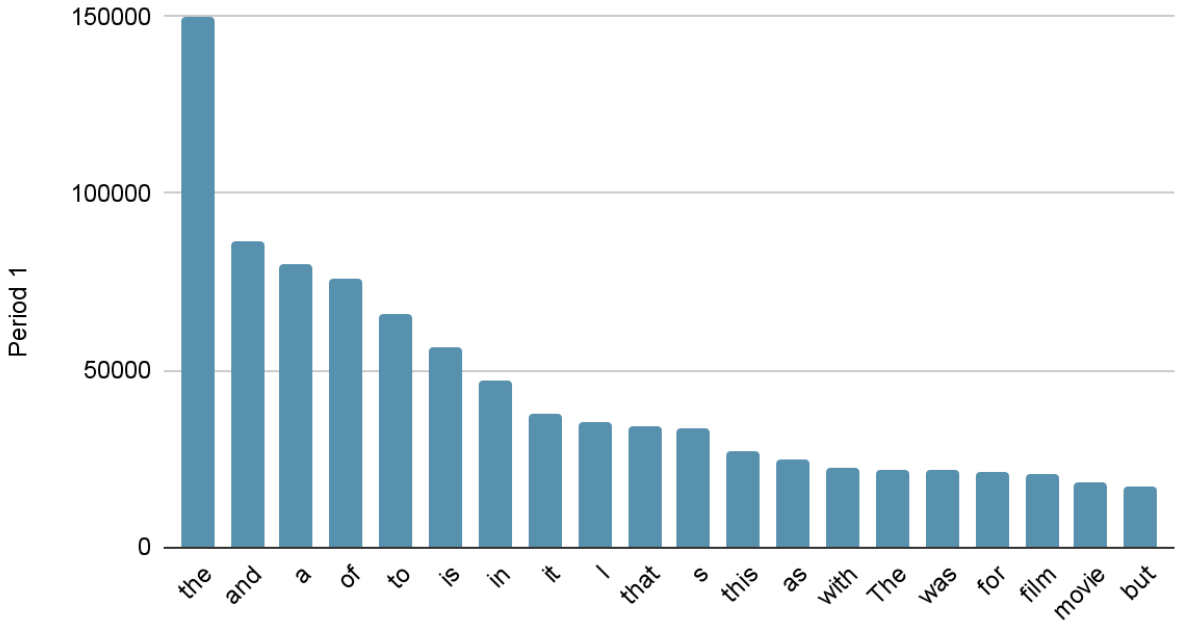
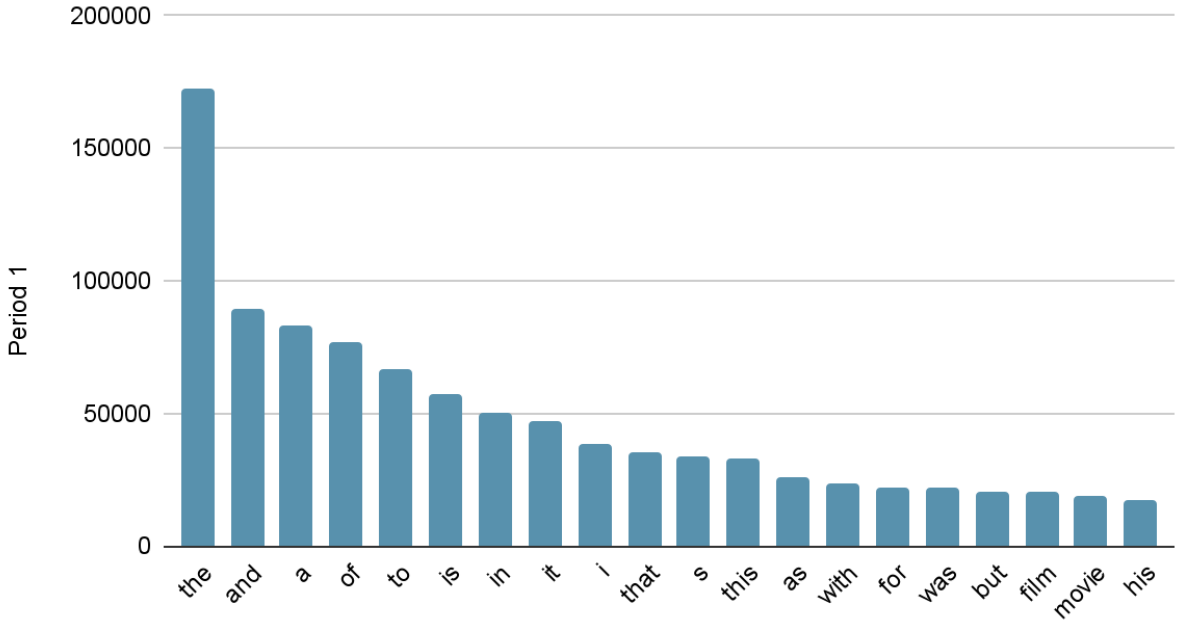
Original POS reviews



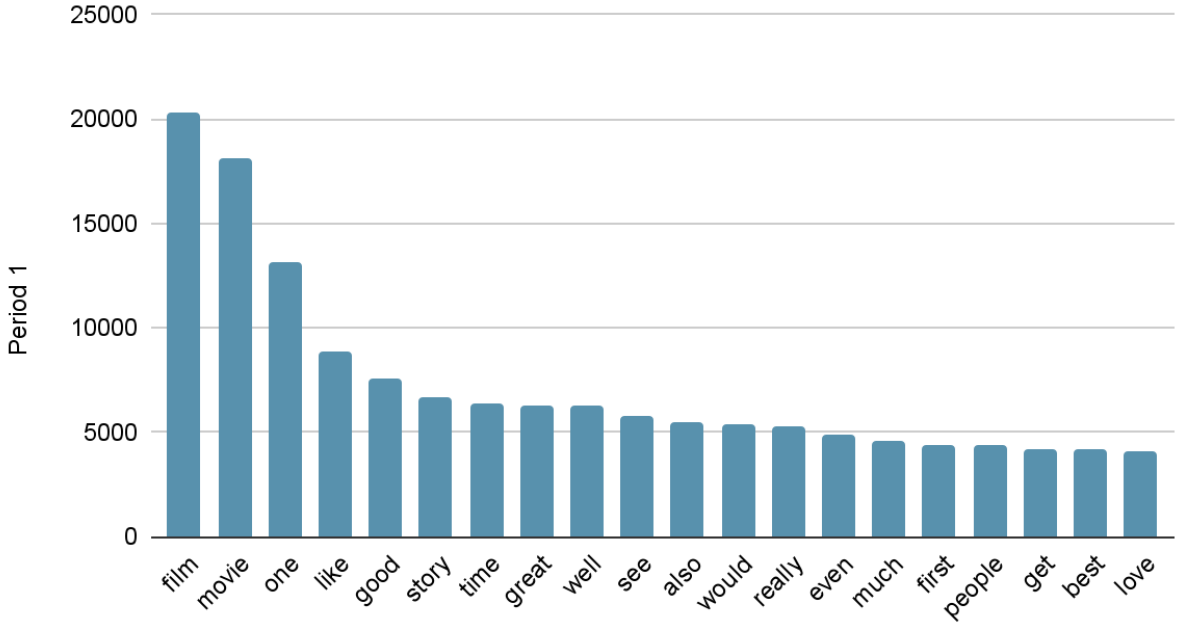
Cleaned POS reviews

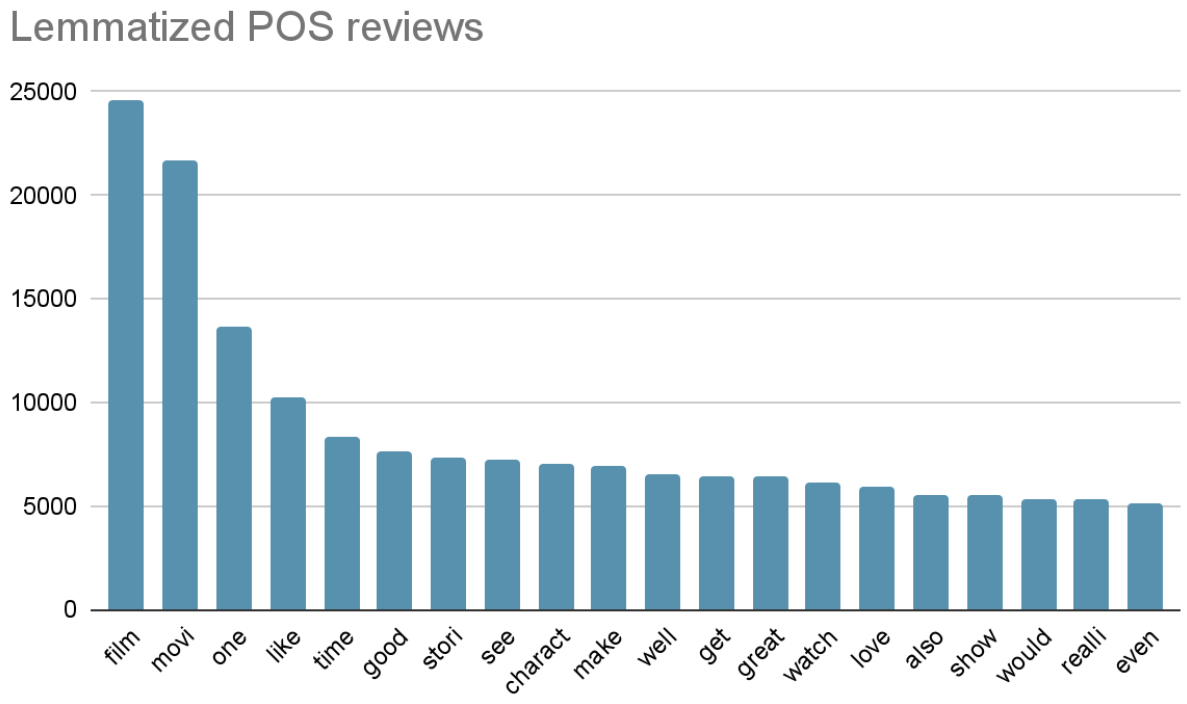


Lowercased POS reviews

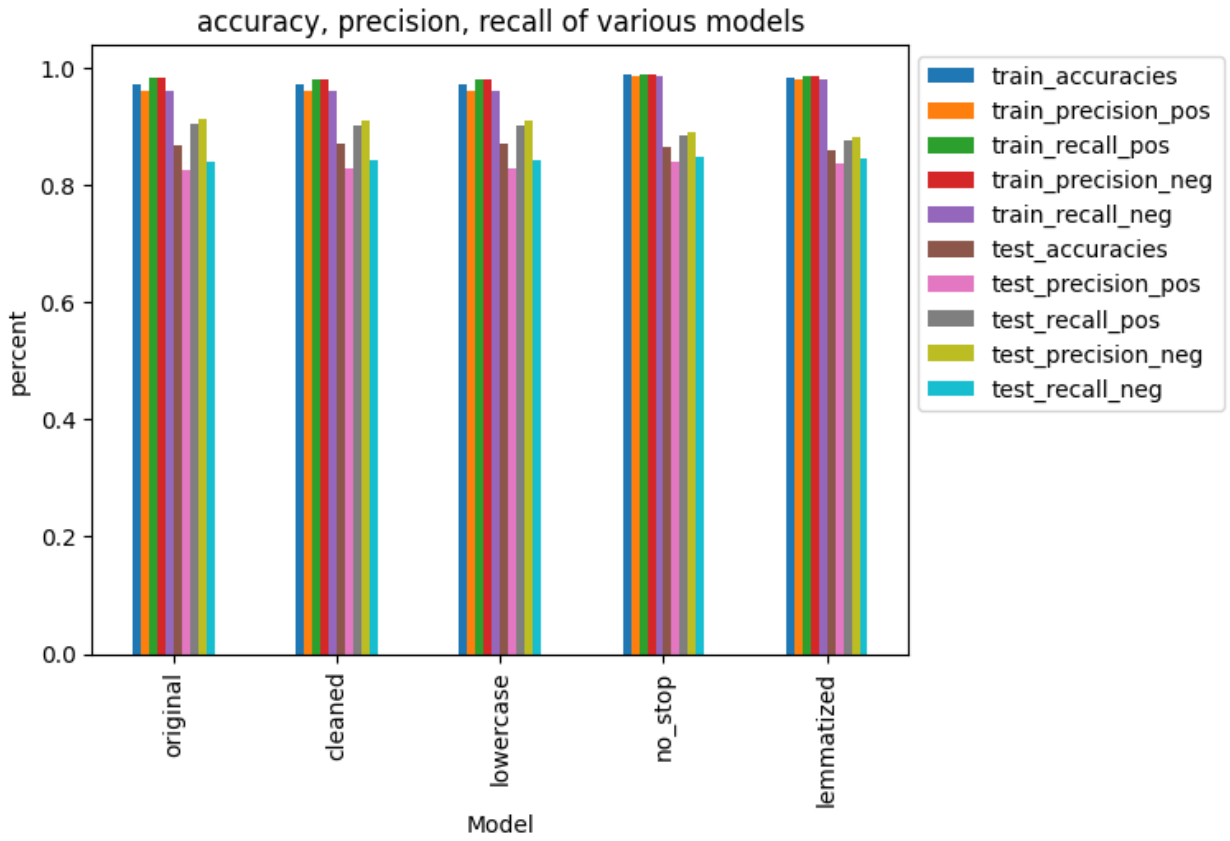


No stopwords POS reviews





The lemmatizer grouped similar words together so that they weren't separated into individual tokens. For example, "movie" and "movies" are two words that would contain the same meaning but would be otherwise separated into individual tokens. A lemmatizer would prevent this from happening. This lemma is represented in the table with a base form of "movi". I would say that the lemmatizer seemed to do a good job based on the previous tables.



1. English as a language, it has certain properties. E.g. English is a morphologically simple language. How does this potentially affect the value of lemmatization?
2. Dataset size. IMDB is fairly large! How may this affect the value of preprocessing?
3. The genre. How do people tend to write reviews, what style/grammar do they tend to employ? How might that affect the value of e.g. lowercasing?
4. …anything else that you might think about here! We look forward to hearing your thoughts!

What the graph seems to indicate is that there is very little difference between the different models that were tested in our program. As our model is very simple, changes made in the data itself are unlikely to change the results in the end. Additionally, the dataset is quite large, and in statistics the law of large numbers means that we are more likely to achieve the “average” result, which in this case is similar results from all the models. Although the model does get better scores for accuracy, precision, and recall for the lemmatized and no_stop models compared to the other models, it isn’t noticeably so or by a significant amount.