Notes to Sentiment Analysis

Sentiment analysis is used to detect the underlying sentiment of a text. Therefore, it classifies the text as either positive or negative.

First, in order to visualize the most frequently used words in the reviews to give us a short overview, we created two so-called word clouds, one in English and one in Norwegian. To just have the important words in this representation, the stop words were excluded from the analysis. What is most striking here is that in both plots predominantly positive words are printed. The largest printed and thus most frequently used words in English are apartement, Oslo and place and further clean, comfortable, helpful and easy. Also in the Norwegian plot there are mainly positive expressions like "anbefale" (engl. recommend), "fin" (engl. fine) and "flott leilighet" (engl. great apartement).

To obtain a more informed analysis, we then performed a detailed sentiment analysis for each review. To do this, we used the *transformers* package, more specifically we used the *pipeline* function with the *task* argument set to *sentiment-analysis*, which is used for classifying sequences according to positive or negative sentiments (s. documentation). The used model is DistlBERT, a small, fast, cheap and light Transformer model. This distilled version of BERT, has 40% less parameters and runs 60% faster while preserving over 95% of Bert's performances as measured on the GLUE language understanding benchmark. (?, s. documentation).

The results are stored in a new data frame and applied in feature engineering to generate the new feature "proportion of negative reviews".

Word clouds:

- generated word clouds out of the 77,000 reviews; one in english, one in norwegian
- what is conspicuous: mostly positive words in both
- "biggest" and therefore most often used words in english: apartement, Oslo, place and further clean, comfortable, helpful and easy
- in norwegian: "leiligheten" (apartement), Oslo, "anbefale" (recommend), "flott" (great), "alt" (everything)

Sentiment Analysis

- used "transformers" package; pipeline function
- pipeline "sentiment-analysis", distilbert algorithm
- now use this to decide if a review is positive or negative
- save in new dataframe
- used to expand the listings dataframe with the new column: fraction negative reviews