Notes to Sentiment Analysis

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

First, to visualize the most frequently used words in the reviews, we created two so-called word clouds, one in english and one in norwegian. To just have the important words in this plot, the stopwords are excluded from the analysis. Most conspicuous here: there are mostly positive words printed in both plots. The biggest printed and therefore most often used words in english are apartement, Oslo and place and further clean, comfortable, helpful and easy. Also in the norwegian plot are mainly positive phrases like "anbefale" (engl. recommend), "fin" (engl. fine) and "flott leilighet" (engl. great apartement).

But to get a more justified analysis we also preformed a sentiment analysis in detail for every review. We therefore used the *transformers* package, more precise 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 saved in a new data frame and within the feature engineering used to generate the new feature "fraction 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