Notes to Feature Engineering: Reviews

In order to extract the most important information from the reviews, we have performed the following analyses.

First, we used the *langdetect* package to determine the language of each review. With this information, we tried to get some insights about the internationality of the guests. Since English and Norwegian are the most commonly used languages, we then created two so-called word clouds in this languages to visualize the most frequently used words in the reviews and to give us a short overview of the given ratings. To just have the important words in this representation, a list of given stop words are used to extract them. 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).

The results of the language detection are stored per *listing_id* in a new data frame called *reviews_features*. This data frame contains the number of reviews per listing, the median length of a review, the number of different languages, as well as a list of languages in which the reviews for that apartment were written. The percentages of Norwegian and English reviews are also recorded for each listing id. Finally, this data frame was added to the *listings* data frame on which feature selection is performed later.

In addition, to obtain a more informed analysis of the reviews, we also performed a detailed sentiment analysis for each review. Sentiment analysis are used to detect the underlying emotion of a text. Therefore, it classifies the text as either positive or negative. 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 and light Transformer model, a distilled version of BERT algorithm, which achieves significantly faster results. (s. documentation). Finally, we used the sentiment analysis to determine the ratio of negative reviews to the total number of reviews per listings id, which is also added to the reviews_feature data frame.