Topic Modelling of high and low ratings

In this section we use topic modelling to find common threads between reviews of score 5 and reviews of score 1, with the aim of discerning any patterns in the terms that contribute to the topic.

We explored several techniques for topic modelling namely NMF, SVD and LDA.

Non-negative matrix factorization

The NMF model decomposes its input matrix into two smaller approximate product matrices that only contain nonnegative values these are iteratively adjusted until they more closely result into the input matrix due to this process the features are clustered as the error value is reduce during each iteration.

NMF can take input matrices that have been processed by both term frequency and TF-IDF

Singular value decomposition

The SVD model decomposes the input matrix into its constituent parts in the form of 3 matrices. SVD acts as a feature reducer removing terms that are not important to the overall corpus.

Latent Dirichlet Allocation

LDA assumes that all topics follow a Dirichlet distribution across the documents in the corpus this leads to the probabilities of context between words being preserved. LDA groups together terms that occur together often into a topic which at times may not lead to topical grouping.

Observations

NMF TF

When using NMF with a TF input there is a clear trend in the topics it is very easy to find an overarching commonality between the terms and possible “topic header” for example topic 9 of the score 5 group and topic 11 of the score 1 group are both about hair products. There is also a trend in the use of positive words in the Score 5 group and negative words in the Score 1 group, while the words No and Not appear in the Score 5 Group more telling words like bad, disappointment, weak and burn are seen in the Score 1 Group in comparison Score 5 group has no such words instead words like loves, like, quality and favourite are seen but some positive terms are seen in Score 1’s topics.

NMF IDF

The results of NMF IDF are just as good as those seen in NMF TF.

SVD

SVD’s topics tends to have several improperly assigned terms in the Score 5 group topic 16 which seems to cover to topic of beverages has a out of place term “dog” this improperly assigned terms occurs in all topics with some topics having no clear topic.

LDA

LDA has mixed results there is no clear positive or negative arrangement of terms in the Score groupings. Some topics like Score 5 group topic 0 is clearly about Oil while theory topics like topic 0 of Score 1 Group has many unrelated terms in them making topic assignment unclear.

The NMF model has the most coherent results all topics have a clear theme and the Score groups do not have any contamination of positive and negative terms. LDA picks up word groups that occur commonly with one another leading to mis grouped terms while SVC approximates the groups very loosely.