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

NMF IDF

SVD

LDA