Modeling

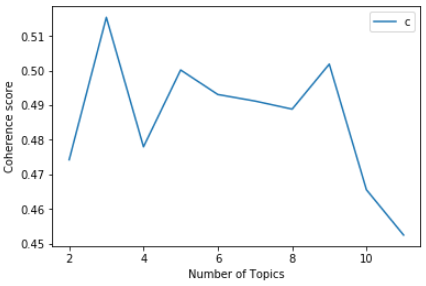
After data preprocessing, we explored three models to better capture the latent topics: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). All of those three models were trained on the Kindle Family to extract the topics so that we can better understand the relationship between topics and reviews.

LSA

LSA is a basic model employs Singular Value Decomposition with document-words as input and the topics with respect to each words and documents as output. By choosing the number of topic, we decide the elements in the V and U matrix, which can be interpreted as the relationship between topics and documents or the words.

The LSA has one very strong advantage is that it is extremely efficient when applying. The reason is that it uses singular value decomposition so that the return is based simply on the decomposed matrixes. When the immediate topic extraction is required, it is one good method. However, it has multiple cons that needs to be taken care of. The first is that it lacks the interpretable embedding, where the coefficients associate with each word have two signs. Statistically speaking, negative sign refers to a lack of the word. However, when analyze it within real world context, it tends to be not interpretable. Meanwhile, it requires relatively large data set to be accurate.

The number of topics is selected based on the coherence score shown as below. The number associated with the optimal occurrence score is three, which is the input for number of topics.



PLSA

PLSA is a probability version of LSA, which takes consideration of conditional probability of document given topic and probability of words given a topic. Instead of returning coefficients, it returns the conditional probability for U and V matrix.

The PLSA is more flexible than LSA given that it involves the word and document conditional probability. However, it is even harder to explain the results from PLSA because the numbers are the conditional probabilities which lack the interpretation in the matrix. Moreover, as a probability-based model, it tends to overfit with the increasing number of documents in the input data. Another disadvantage is that when new data come in, since there is no parameter for the probability function.

For PLSA, it is hard to determine number of topics since coherence sore does not work well on this model. Our team manually experimented with different numbers and set the number as five since four topics among them is interpretable based on the words with top conditional probability.

LDA

The model that is most popular currently is the LDA modeling. It is a Bayesian version of PLSA using dirichlet priors for the document-topic and word-topic distributions. According to Wikipedia, it allows sets of observations to be explained by [unobserved](https://en.wikipedia.org/wiki/Latent_variable) groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics (<https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation>).

It has numerous advantages comparing to the previous two models. The first is that it can be better generalized than PLSA model, which means that when new data points are input to the model, it can generate the relating vector representing their topic mixtures in coefficients. It is more accurate than both PLSA and LSA, and easier to interpret since each topic is characterized by the words they are most strongly associated with. This is due to the fact that all the coefficients responding to each word are all positive numbers. As a result, the larger the value is, the more strongly the word is relating to the topic. However, it is slower in training process and still require human intervention to label the topic.

The number of topics is also determined by the coherence score, where the optimal score is achieved when the number of topics equals to three.

