Topic modelling using LDA (Latent Dirichlet Allocation)

We did topic modelling on abcnews.csv dataset with 49771 rows. We processed headline\_texts.

First we represented dataset words using both Bag of Words (BOW) technique and tf-idf (Term Frequency-Inverse Document Frequency), then applied LDA to cluster topics.

In order to improve LDA model performance by changing model hyperparameters, we construct 360 LDA models by different parameters and corpus. We create two **tf-idf corpus** and **bow corpus** and changed number of ***topics*** and ***alpha*** and ***beta*** parameters. We get the Coherence score for each model to evaluate accuracy of the model. The project was run under colab framework using TPU runtime type and last more than 2:30 hour to complete.

The results saved in a .csv file named “lda\_tuning\_results.csv”. Figure 1 show coherence score vs number of topics from 5 to 10 with fixed Alpha and Beta parameters equal to 0.01 through both tf-idf and bow corpus. We can see that coherence score based on tf-idf vectorization is higher than Bag of Words and the highest score when number of topics is 8.

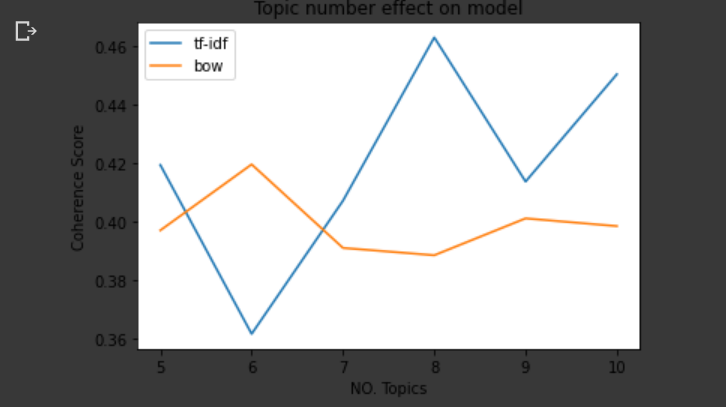


Figure 1.

By searching the highest coherence score in this file we find that

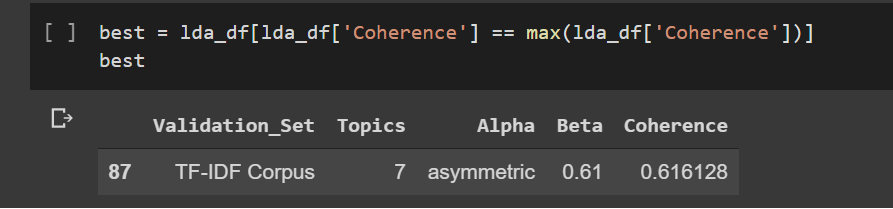


Figure 2. The best coherence score

Figure 2 shows the parameter values achieving the maximum value of coherence score, then we plot a plot based on 7 topic numbers and fixed Alpha value as “asymmetric” through tf-idf and BOW corpus vectors to get the proper value of Beta parameter. See figure 3.

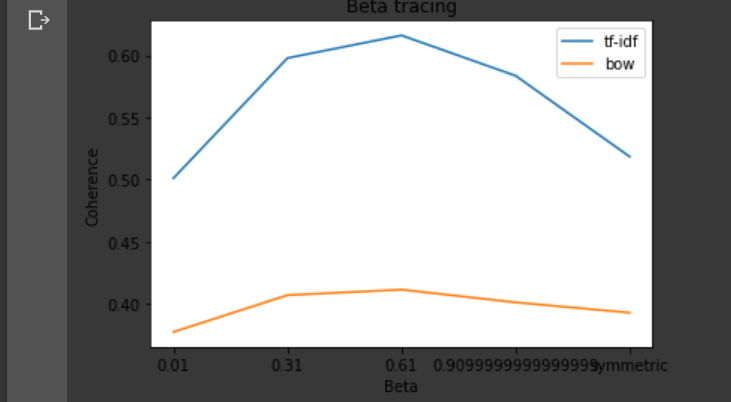


Figure 3. The most proper value of Bets is in middle of range.

We can see in figure 3 that the best value for Beta is 0.61 in the middle of range. For values higher than 0.6 the coherence becomes worse.

In all of the figures we see that tf-idf based LDA model always obtains much better coherence values than BOW corpus based model.