|  |
| --- |
| **Data**  This dataset contains about 11k newsgroups posts from 20 different topics. This is dataset shows that if we train our model on daily news posts then we can easily figure out what topics are trending and take business and political decision accordingly  **Preprocessing Techniques**  For data preprocessing the first step we took is to remove emails, newline and extra spaces using regular expression. Then we tokenized the corpus which is breaking the sentence into a list of words for further cleaning purposes. After tokenizing we removed all punctuations, unnecessary characters and stopwords from the tokenized corpus. Then we created bigrams and perform lemmatization on the corpus. After all cleaning of the corpus, we transformed the corpus into a trainable format which is a dictionary.  **LDA Model**  LDA’s approach to topic modelling is it considers each document as a collection of topics in a certain proportion. And each topic as a collection of keywords, again, in a certain proportion  Once you provide the algorithm with the number of topics, all it does is rearrange the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of the topic-keywords distribution.  A topic is nothing but a collection of dominant keywords that are typical representatives. Just by looking at the keywords, you can identify what the topic is all about. The following are key factors to obtaining good segregation topics:   1. The quality of text processing. 2. The variety of topics the text talks about. 3. The choice of topic modelling algorithm. 4. The number of topics fed to the algorithm. 5. The algorithms tuning parameters |

|  |
| --- |
| Now we have everything required to train the LDA model. In addition to the corpus and dictionary, you need to provide the number of topics as well.  Apart from that, alpha and eta are hyperparameters that affect the sparsity of the topics. According to the Gensim docs, both defaults to 1.0/num\_topics prior.  Chunk size is the number of documents to be used in each training chunk. update\_every determines how often the model parameters should be updated and passes are the total number of training passes  **Parameters of LDA**  **Alpha and Beta Hyperparameters –** alpha represents document-topic density and Beta represents topic-word density. Higher the value of alpha, documents are composed of more topics and lower the value of alpha, documents contain fewer topics. On the other hand, higher the beta, topics are composed of a large number of words in the corpus, and with the lower value of beta, they are composed of few words.  **Number of Topics** **–** Number of topics to be extracted from the corpus. Researchers have developed approaches to obtain an optimal number of topics by using Kullback Leibler Divergence Score. I will not discuss this in detail, as it is too mathematical. For understanding, one can refer to this[1] original paper on the use of KL divergence.  **The number of Topic Terms –** Number of terms composed in a single topic. It is generally decided according to the requirement. If the problem statement talks about extracting themes or concepts, it is recommended to choose a higher number, if the problem statement talks about extracting features or terms, a low number is recommended.  **Number of Iterations/passes –** Maximum number of iterations allowed to LDA algorithm for convergence.  **Model evaluation metrics**  Model perplexity and topic coherence provide a convenient measure to judge how good a given topic model is. In my experience, the topic coherence score, in particular, has been more helpful. |

|  |
| --- |
| **Choosing the right number of Topics**  Here approach to finding the optimal number of topics is to build many LDA models with different values of the number of topics (k) and pick the one that gives the highest coherence value as shown in the graph below. ( 5 topic has more coherence score)    After finding the dominant topics we save the current model to reuse it for further analysis which changes from business to business.  **Conclusion**  We built a basic topic model using Gensim’s LDA and visualize the topics using pyLDAvis. You saw how to find the optimal number of topics using coherence scores and how you can come to a logical understanding of how to choose the optimal model. Finally, we saw how to aggregate and present the results to generate insights that may be more actionable. |