Week 10 Research Notebook

From the research topic “Universities”:

When using LDA for topic modeling on mission statements, we are essentially extracting latent features that represent the main themes or topics present in the text. These features are inferred from the patterns of word co-occurrence in the mission statements. LDA assumes that documents are generated by a mixture of topics, and topics are probability distributions over words.

To control or monitor the features being extracted, consider the following approaches:

Text preprocessing: This may include removing stop words, punctuation, special characters, and converting text to lowercase. We can also apply stemming or lemmatization to reduce words to their root forms.
Hyperparameters: to influence the model's behavior like

n_components: The number of topics to be extracted. A higher value results in more fine-grained topics, while a lower value results in more general topics.

max_iter: The maximum number of iterations for the LDA model to converge. Higher values can lead to better convergence, but may increase computation time.

learning_decay: The learning rate decay for the online variational Bayes method. Adjusting this parameter can impact the convergence of the model.

doc_topic_prior and topic_word_prior: The Dirichlet prior parameters, also known as alpha and beta. Adjusting these values affects the sparsity of the document-topic and topic-word distributions, influencing the granularity of the topics.

Feature extraction: We can control the features used as input to the LDA model by using different text vectorization techniques, such as Bag-of-Words, TF-IDF, or pre-trained word embeddings. The choice of vectorization method can impact the quality and interpretability of the extracted topics.

Model evaluation: Assess the quality of the extracted topics using topic coherence measures like UMass, UCI, or NPMI. These measures help evaluate how semantically meaningful and interpretable the topics are by analyzing the similarity between the top words in each topic.

Model evaluation in the context of text analytics is about assessing the quality and effectiveness of the text processing or analysis methods being used. It is important because it helps us understand how well our chosen methods are working and whether they are providing meaningful and useful insights or results.

Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), aim to discover latent topics within a collection of documents