Topic Modeling Towards Data Science-NLP

Unsupervised learning model

The proposal of Project 4 bootcamp data science T5

Objective:

Explore the articles highlighted in TowardsDataScience (TDS). This project breaks down all 35000+ blog posts on TDS into 10 separate documents using two methods.

1. A document-term matrix is reduced in dimensionality to 30 subtopics using Non-Negative Matrix Factorization
2. Document Vectors are extracted and clustered based on their cosine similarity.

Motivation:

Expand my topic modeling to organize new documents into their respective bins. This method can be expanded to training a model in a different domain to classify new documents.

Tools:

Using NLP and clustering methods to find patterns in TDS articles.

Topic modeling was performed using TF-IDF and NMF on the entire corpus. A separate doc2vec model was then trained on the corpus extracting document vectors from the articles using Gensim

Organize the documents into groups based on semantic similarity, I used KMeansClusterer to group together the document vectors..