youssief amr

[Company name]  [Company address]

topic modelling

Introduction:

In this project, we explore the concept of topic modelling and its application to clustering customer reviews.

By leveraging unsupervised machine learning techniques, we aim to identify patterns and recurring themes within a set of articles and group them based on their similarity. We use methods such as TF-IDF (Term Frequency-Inverse Document Frequency) and K-means clustering to achieve our objective.

Dataset:

We begin by working with a dataset of articles, which includes various attributes such as ID, title, publication, author, date, year, month, URL, and content. The content column contains the textual data that we will analyze and convert into vectors to determine the similarity between articles.

Data Preprocessing:

Before performing topic modelling, we need to preprocess the text data to remove noise and irrelevant information.

We implement several text preprocessing techniques using regular expressions (regex) and the NLTK library, including removing special characters, numbers, URLs, and emails. We also tokenize the text into words, apply lemmatization to reduce words to their base form, and remove stop words to eliminate commonly occurring words with little semantic value.

TF-IDF Vectorization:

To convert the preprocessed textual data into numerical vectors, we employ the TF-IDF vectorization technique.

TF-IDF assigns weights to each term in a document based on its frequency within the document and its rarity in the entire corpus. The resulting TF-IDF matrix represents the articles as vectors in a high-dimensional space, where each dimension corresponds to a unique term in the corpus.

K-means Clustering:

We apply the K-means clustering algorithm to the TF-IDF matrix, aiming to group similar articles together.

K-means is an unsupervised machine learning algorithm that partitions the data into k clusters, where k is a user-defined parameter.

By iteratively optimizing cluster centroids, K-means assigns each article to the cluster with the closest centroid based on the Euclidean distance in the TF-IDF vector space.

Clustering Visualization:

To gain insights from the clustering results, we reduce the dimensionality of the TF-IDF matrix using PCA (Principal Component Analysis) and visualize the clusters in a two-dimensional space. PCA captures the most informative components of the TF-IDF matrix, enabling us to represent the articles in a scatter plot with different colors indicating the assigned clusters.

Evaluation and Interpretation:

To assess the quality of the clustering, we can use metrics like the silhouette score, which measures the cohesion and separation of the clusters.

A higher silhouette score indicates better-defined and more distinct clusters. Additionally, we can interpret the topics of each cluster by examining the articles within them, identifying common keywords and themes.

Conclusion:

Topic modelling through the application of TF-IDF vectorization and K-means clustering is a valuable technique for analyzing and organizing textual data. In this project, we have demonstrated its potential by clustering customer reviews based on similarity. By identifying common topics and themes, businesses can gain insights into customer sentiments and preferences, enabling them to make data-driven decisions to improve their products or services.

By implementing this project, we have explored the process of text preprocessing, vectorization, clustering, and visualization. We hope this serves as a useful guide for understanding topic modelling and its practical application in real-world scenarios.