Constrained Clustering of Text with Textual Explaination

Advanced Topics in Machine Learning Semester Project 2024 (**T05**)

Jonathan, Darpan, Mayank, Indrajit

Motivation and Problem Statement

Motivation: The 20NewsGroup dataset consists of several news clubbed together(that is approximately 18000 news from 20 different news groups).This collection is often used in text classification and text clustering. For our project we use the dataset to perform text clustering without using the labels. There are few challenges preprocessing, feature extraction, constraint generation, clustering, visualizing.

Problem Statement: Our goal from this challenge is to perform text clustering and visualizing the data to predict the cluster for the documents that it should belong to, by learning the underlying pattern from each document by using algorithm like PCK-means with the help of constraint generation without using the labels during clustering.

Data Constraints Creation and Constraints Clustering

Dataset: There are 20 Newsgroups in the dataset where , and each news group contains number of documents containing "From","Subject",''Newsgroup'',''Document_id'',''Organization'', and the body of the news.

Constraint Creation: We extract all the key phrases from every document, then we calculate the Jaccardi ndex and cosine similarity between each pair of documents. Based on the score (J.I > 0.10 = ML && cosine similarity > 0.3) and (J.I < 0 = ML && cosine similarity < 0.3) we form the mustlink/cannotlink constraints.

Constraint Clustering: Used algorithm like PCK-means which demonstrates soft clustering. We used a weight of 0.1, tolerance of 0.0001, max_iteration of 200 which means we iterate utmost 200 times until the datapoints converges to there respective cluster, if not then the iteration is forcefully stopped.

Feature Extraction

1.Tfidf
General idea of Tfidf is to count the frequency of each word in a document and find importance of the words.
The steps involved are as follows:-

Feature Extraction: We use TfidfVectorizer() api from sklearn library to generate tfidf_matrix which we consider as the features for the documents.

2.**Word Embeddings**
General idea of Word Embeddings is to generate a real-valued vector that represents the word in such a way that the words that are closer in the vector space are expected to be similar in meaning.
The steps involved are as follows:-

Feature Extraction: We use en_core_web_md api from Spacy library to calculate mean of every word embedding vector for each document

3.**Key Phrases**
General idea of extracting the words with the most relevance, and expressions from the
The steps involved are as follows:-

Feature Extraction: We use en_core_web_md api from Spacy library to generate key phrases

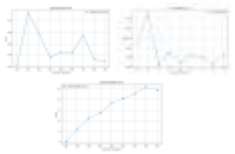
Cluster Generation: Using the Euclidean distance formula we calculate must link, cannot link constraints.and generate clusters from PCK-means

Cluster Explaination: We use Naive-Bayes classifier and LIME text classifier explaination for explaining each clusters generated for the dataset.

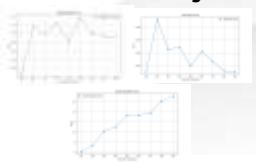
Evaluation

Silhouette Score : The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
Adjusted Rand Index: The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.
Davies-Bouldin Index: The DBI is calculated as the average of the maximum ratio of the within-cluster distance and the between-cluster distance for each cluster.

Tfidf



Word Embeddings



Key Phrases