Data Glacier

Your Deep Learning Partner

Week 8 Deliverables

Name	Manoj Nagaraja
Email	manojn.7270@gmail.com
Country	United Kingdom
College	University of Liverpool
Specialization	Data Science

Problem description

Requires to implement various clustering algorithms using the Python programming language and apply them to cluster a given dataset. The purpose of this project is to assess the understanding of various clustering algorithms by implementing the algorithms and applying them to text clustering.

Data understanding

Data understanding is the initial step in the data mining process that involves getting a better understanding of the data that will be used for analysis.

Word embeddings are generated using advanced machine learning algorithms such as word2vec, GloVe, or fastText. They are used to capture the semantic and syntactic information of a word, which is often used in natural language processing (NLP) tasks such as text classification, sentiment analysis, and machine translation.

To analyze the data in the file, one would need to load it into a suitable programming environment, such as Python, and use appropriate libraries for data manipulation and analysis, such as NumPy or Pandas. Once loaded, one could perform various data exploration techniques, such as descriptive statistics and data visualization, to gain insights into the data and better understand its structure and properties.

What type of data you have got for analysis ?

Each line in the file represents a word and is accompanied by 300 features that describe the meaning of that word. These features are typically referred to as "word embeddings" and are numerical representations of the meaning of a word in a high-dimensional space.

What are the problems in the data (number of NA values, outliers , skewed etc) ?

Missing Values: There may be missing values in the word embedding data, which can cause problems for some machine learning algorithms. You may need to impute missing values using techniques such as mean imputation, median imputation, or regression imputation.

Outliers: There may be outliers in the word embedding data, which can skew the results of statistical analysis or machine learning models. You may need to detect and remove outliers using techniques such as z-score or IQR methods.

Skewed Data: The distribution of values for some or all features of the word embedding data may be skewed, which can cause problems for some machine learning algorithms. You may need to transform the data using techniques such as log-transformation, power-transformation, or box-cox transformation.

Dimensionality Reduction: There may be a high number of features (300 in this case) in the word embedding data, which can cause problems for some machine learning algorithms. You may need to perform dimensionality reduction using techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE).

It's important to note that the specific problems in the data may vary depending on the characteristics of the data itself, and the specific goals of your analysis. Therefore, it's important to carefully explore and preprocess the data to ensure that it is suitable for your specific analysis.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

Missing Values: NA values can be handled by either removing the rows or columns with NA values, or by imputing the missing values. The choice of approach depends on the amount and pattern of missing data. If the missing data is small, it is better to impute the values to retain as much data as possible. If there is a large amount of missing data, it may be better to remove those rows or columns.

Outliers: Outliers can be handled by either removing them or by transforming the data using techniques such as log transformation, Box-Cox transformation, or Winsorization. Removing the outliers should be done carefully, as outliers may contain important information about the data. It is important to identify the reason for the presence of outliers and to decide whether to remove them or not based on domain knowledge.

Dimensionality Reduction: Dimensionality reduction techniques can be used to reduce the number of features in the dataset. Principal Component Analysis (PCA) is a popular method for reducing the dimensionality of the dataset. PCA identifies the directions in which the data varies the most and projects the data onto a lower-dimensional space while retaining as much of the variability as possible. Other techniques for dimensionality reduction include Linear Discriminant Analysis (LDA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Non-negative Matrix Factorization (NMF)

Github Repo link : https://github.com/ManojN7270/Final-Project-week_7-to-week_13.git