MACHINE LEARNING - UNSUPERVISED LEARNING AND FEATURE ENGINEERING

Categorizing Trends in Science

Name: Tourad Baba

Date: 2024-03-26

GitHub Link: https://github.com/TouradBaba/Trends_in_Science

Table of Contents

GitHub Link:
https://github.com/TouradBaba/Trends_in_Science

1. Introduction

In the realm of data science, exploring and understanding the underlying patterns in datasets is paramount. This case study delves into the analysis of a large data set to uncover insights using various techniques such as data preprocessing, TF-IDF transformation, clustering, and dimensionality reduction. By applying these methodologies, we aim to gain deeper insights into the dataset and extract meaningful information for further analysis and decision-making.

1.1-Overview of the Problem Statement

The dataset under consideration is extensive, containing textual information along with other features. The primary challenge lies in effectively processing and analyzing this large-scale text data to uncover hidden patterns and structures. Additionally, the dataset may contain missing values, inconsistencies, or noise that need to be addressed through data preprocessing techniques. Furthermore, given the dataset's size, efficient algorithms and techniques are necessary to handle the computational complexity associated with its analysis.

1.2-Objectives of the Case Study

1. **Data Exploration**: The first objective is to thoroughly explore the dataset, examining its dimensions, data types, and unique categories. This exploration will provide a foundational understanding of the dataset's structure and characteristics.

2. **Data Preprocessing**: The second objective is to preprocess the text data within the dataset. This involves tasks such as removing punctuation, stop words, and performing lemmatization to clean and standardize the text data for further analysis.

3. **TF-IDF Transformation**: The third objective is to transform the preprocessed text data into TF-IDF (Term Frequency-Inverse Document Frequency) vectors. This transformation will enable the conversion of text data into numerical representations suitable for machine learning algorithms.

4. **Clustering Optimization**: The fourth objective is to determine the optimal number of clusters for clustering analysis. This involves exploring techniques such as the elbow method and silhouette score to identify the most suitable number of clusters that best capture the inherent structure of the data.

5. **Clustering**: The fifth objective is to perform clustering using KMeans and MiniBatchKMeans algorithms. This step aims to group similar data points together based on their features and evaluate the clustering performance using the Davies Bouldin Index.

6. **Dimensionality Reduction and Visualization**: The final objective is to reduce the dimensionality of the data using TruncatedSVD and visualize the clusters in lower-dimensional space. This visualization will provide insights into the structure and relationships within the data, aiding in its interpretation and understanding.

2. Data Exploration

Data exploration serves as the foundation for any data analysis task, providing critical insights into the dataset's characteristics and structure. In this section, we embark on a comprehensive exploration journey to gain a deep understanding of the dataset.

2.1-Checking Dataset Dimensions

The first step in our exploration process involves examining the dimensions of the dataset. By determining the number of rows and columns, we gauge the dataset's size and complexity, which is essential for planning subsequent analyses and preprocessing steps.

2.2-Handling Missing Values

Missing values are a common occurrence in real-world datasets and can significantly impact the integrity of analyses. In this step, we meticulously address missing values using appropriate techniques such as imputation or removal. By ensuring the completeness of our data, we mitigate potential biases and inaccuracies in our findings.

2.3-Exploring Data Types

Understanding the data types of different features is paramount for effective preprocessing and analysis. Through thorough inspection, we identify the nature of each column, distinguishing between categorical, numerical, and textual variables. This understanding enables us to tailor preprocessing techniques to suit the specific characteristics of each feature.

2.4-Analyzing Unique Categories

Categorical variables often play a pivotal role in datasets, encapsulating valuable information about different attributes or classes. In this phase, we delve into the unique categories present within each categorical feature. By examining the diversity and distribution of categories, we gain insights into the dataset's underlying structure and potential patterns.

2.5-Visualizing Category Distribution

Visualizing the distribution of categories within categorical variables provides a comprehensive overview of their prevalence and spread. Leveraging various visualization techniques such as bar charts, we elucidate the frequency and proportion of different categories. This visual representation facilitate intuitive understanding and aid in identifying dominant trends or outliers.

2.6-Word Cloud Visualization

Textual data represents a rich source of information, particularly in natural language processing (NLP) tasks. To unravel the semantic nuances embedded within the text, we employ word cloud visualization. By generating word clouds from textual data, we visually depict the frequency of words, highlighting prominent terms and themes. This approach enables us to discern key insights and uncover underlying patterns within the dataset's textual content.

Through meticulous data exploration, we lay the groundwork for subsequent preprocessing, analysis, and visualization tasks, empowering us to extract meaningful insights and derive actionable conclusions from the dataset.

3. Data Preprocessing

Data preprocessing serves as the foundation for robust and insightful analyses, involving a series of steps to refine raw data into a more structured and analytically tractable form. In this section, we delve into the intricacies of data preprocessing, encompassing feature selection and text data refinement techniques.

3.1-Selecting Relevant Features

Feature selection is a critical aspect of data preprocessing, aimed at identifying and retaining the most informative attributes within the dataset while discarding redundant or irrelevant ones. This

process involves a careful evaluation of each feature's contribution to the analysis objectives, considering factors such as predictive power, correlation with the target variable, and computational efficiency.(ChenDataBytes. (2024, January 27).)

By selectively retaining features that are most pertinent to the analysis goals, we streamline subsequent modeling tasks and mitigate the risk of overfitting, where models learn noise instead of true patterns in the data. Through principled feature selection techniques, we aim to enhance the interpretability, generalizability, and performance of the analytical models.

3.2-Preprocessing Text Data

Text data often presents unique challenges due to its unstructured nature, requiring specialized preprocessing techniques to extract meaningful insights effectively. In this sub-section, we focus on refining textual features through a series of preprocessing steps tailored to address common challenges encountered in analyzing text features.

3.3-Removing Punctuation and Stop words

Punctuation marks and stop words, such as articles, conjunctions, and prepositions, are ubiquitous in text data but often convey little substantive meaning. Consequently, they can introduce noise and hinder the accuracy of subsequent analyses. To address this issue, we employ techniques to systematically remove punctuation marks and stop words from the text, thereby enhancing the signal-to-noise ratio and improving the quality of the textual data.

3.4-Lemmatization

Lemmatization is a linguistic technique employed to reduce words to their base or root form, known as the lemma. (Raviraj. (2023, August 27)). By lemmatizing words, we unify variations of the same word, such as different tenses, conjugations, or inflections, into a single canonical form. This process not only reduces redundancy within the text but also standardizes the vocabulary, enabling more accurate comparisons and analyses across documents.

Through meticulous data preprocessing, we aim to transform raw data into a more structured and refined format conducive to exploratory analysis, modeling, and interpretation. By selectively retaining relevant features and refining textual data using advanced NLP techniques, we lay the groundwork for extracting actionable insights and uncovering hidden patterns within the dataset.

4. TF-IDF Transformation

TF-IDF (Term Frequency-Inverse Document Frequency) transformation, As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. (Ramos, J. (2003, December))

4.1-Vectorizing Text Data

Vectorizing text data involves converting textual content into numerical vectors that machine learning algorithms can process. This transformation enables quantitative analysis of textual features, facilitating the application of various statistical and machine learning techniques to text-based datasets.

In vectorizing text data, we encode each document as a vector, where each dimension represents a unique term or word in the vocabulary, and the value of each dimension corresponds to the frequency or importance of the term within the document. This process enables the representation of textual documents in a high-dimensional space, where the geometric relationships between vectors capture semantic similarities and differences.

4.2-Generating TF-IDF Matrix

The TF-IDF matrix is a numerical representation of the corpus that encapsulates the importance of each term within each document relative to the entire corpus. TF-IDF is calculated as the product of two components: term frequency (TF), which measures the frequency of a term within a document, and inverse document frequency (IDF), which measures the rarity of a term across the entire corpus. Y, D. (2024, February 13).

By multiplying the term frequency (TF) by the inverse document frequency (IDF), TF-IDF assigns higher weights to terms that are frequent within a document but rare across the corpus, thus emphasizing terms that are discriminative and informative. This weighting scheme helps mitigate the impact of common terms while highlighting the significance of distinctive terms in each document.

Through the generation of the TF-IDF matrix, we obtain a structured representation of the textual data that preserves the semantic nuances and importance of terms within each document. This matrix serves as the basis for subsequent text analysis tasks, including clustering, enabling the extraction of actionable insights from text-rich datasets.

5. Clustering Optimization

In this section, we focus on optimizing the clustering algorithm by determining the optimal number of clusters for partitioning the dataset. Clustering is a fundamental unsupervised learning technique that groups similar data points into clusters, enabling the discovery of inherent patterns and structures within the data.

5.1-Exploring Optimal Number of Clusters

Determining the appropriate number of clusters is crucial for effective clustering analysis. However, identifying the optimal number of clusters is often challenging and requires leveraging various techniques for evaluation.

Two common methods for exploring the optimal number of clusters are:

1. Elbow Method:

- The elbow method is a heuristic technique that evaluates the within-cluster sum of squares (WCSS) for different numbers of clusters.

- It involves plotting the number of clusters against the corresponding WCSS and identifying the point where the rate of decrease in WCSS slows down, resembling an "elbow" in the plot.

- The number of clusters at the elbow point is considered optimal, as it signifies a balance between maximizing intra-cluster cohesion and minimizing inter-cluster separation.

2. Silhouette Score:

- The silhouette score is a metric that quantifies the quality of clustering by measuring the separation between clusters and the cohesion within clusters.

- It calculates the average silhouette coefficient for each data point, which ranges from -1 to 1, with higher values indicating better cluster separation.

- The silhouette score provides a quantitative measure of cluster validity, where higher scores correspond to more appropriate cluster assignments.

By employing both the elbow method and silhouette score, we aim to identify the optimal number of clusters that maximizes the homogeneity within clusters while maximizing the distinctiveness between clusters. This optimization process ensures that the clustering algorithm produces meaningful and interpretable results, facilitating subsequent analysis and decision-making.

6. Clustering

In this section, we delve into the implementation of clustering algorithms, specifically KMeans and MiniBatchKMeans, to partition the dataset into distinct clusters based on feature similarity. Additionally, we evaluate the clustering performance using the Davies Bouldin Index (DBI), a metric that assesses the quality of clustering solutions.

6.1-Implementing KMeans and MiniBatchKMeans

- **KMeans**: KMeans is a popular clustering algorithm that partitions the dataset into K clusters by iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the mean of data points within each cluster.(Sharma, P. (2024b, February 20)) It converges when the centroids stabilize or the specified number of iterations is reached.

- **MiniBatchKMeans**: KMeans is a clustering algorithm that has been widely adopted due to its simple implementation and high clustering quality. However, the standard km suffers from high computational complexity and is therefore time-consuming. Accordingly, the mini-batch (mbatch) km is proposed to significantly reduce computational costs in a manner that updates centroids after performing distance computations on just a mbatch, rather than a full batch, of samples.(X. Zhu, J. Sun, Z. He, J. Jiang and Z. Wang, 2023)

6.2-Evaluating Clustering Performance with Davies Bouldin Index (DBI)

David L. Davies and Donald W. Bouldin devised a method known as the Davies-Bouldin Index (DBI) for assessing clusters. The evaluation based on the Davies-Bouldin Index offers an internal cluster evaluation scheme, where the quality of cluster results is determined by considering both the quantity and proximity among clusters.(Mughnyanti, M., et al (2020)).

A lower DBI indicates better clustering performance, where clusters are well-separated and internally cohesive.

By computing the DBI for both KMeans and MiniBatchKMeans clustering solutions, we can quantitatively assess their effectiveness in producing coherent and distinct clusters. This

evaluation provides insights into the quality of clustering results and guides the selection of the most suitable clustering algorithm for the dataset at hand. In our case it was KMeans.

7. Dimensionality Reduction and Visualization

In this section, we focus on reducing the dimensionality of the dataset using Truncated Singular Value Decomposition (TruncatedSVD), a technique commonly used for high-dimensional data. Subsequently, we visualize the reduced-dimensional data in both two and three dimensions to gain insights into the underlying structure and patterns.

7.1-Reducing Dimensionality with TruncatedSVD

Truncated Singular Value Decomposition (TruncatedSVD) is a dimensionality reduction technique commonly used in conjunction with term-document matrices, such as TF-IDF matrices, to capture latent semantic structures in textual data. It decomposes the original matrix into lower-dimensional representations while retaining the most important information.

Unlike traditional Singular Value Decomposition (SVD), which computes the full decomposition of the matrix, TruncatedSVD approximates the decomposition by retaining only the top k singular values and corresponding singular vectors, where k is the desired number of dimensions for the reduced representation. By discarding less significant singular values and vectors, TruncatedSVD effectively reduces the dimensionality of the data while minimizing information loss.

This technique is particularly useful in natural language processing tasks, such as document clustering, topic modeling, and recommendation systems, where high-dimensional text data need to be processed efficiently. TruncatedSVD enables researchers and practitioners to analyze and visualize large text corpora in a more computationally tractable manner, facilitating insights and discoveries in the underlying structure of the data.

7.2- 2D Visualization

- We visualize the reduced-dimensional data in a two-dimensional space, allowing us to observe the clustering structure and relationships between data points. This visualization facilitates the interpretation of clustering results and provides insights into the overall distribution of data points.

7.3- 3D Visualization

- Additionally, we extend the visualization to three dimensions to explore the dataset's structure further. By visualizing data in three dimensions, we gain a more comprehensive understanding of the spatial relationships and potential clusters within the dataset.

Through dimensionality reduction and visualization, we aim to uncover meaningful patterns and structures inherent in the data, enabling better interpretation and comprehension of the clustering results.

8. Conclusion

In this case study, we embarked on a journey to explore, analyze, and cluster a dataset containing scholarly articles from various fields. Here are the key takeaways from our analysis:

1. **Optimal Number of Clusters**: Through the Elbow Method and Silhouette Score, we determined that the dataset is best represented by thirteen clusters. This number was chosen based on the optimal balance between intra-cluster cohesion and inter-cluster separation.

2. **Clustering Performance**: We evaluated the performance of two clustering algorithms, KMeans and MiniBatchKMeans, using the Davies Bouldin Index. While both algorithms demonstrated reasonable clustering, KMeans outperformed MiniBatchKMeans slightly, as indicated by lower Davies Bouldin Index scores.

3. **Homogeneous Clusters**: Each of the thirteen clusters identified in our analysis exhibited homogeneity in terms of the categories of articles they contained. For example, Cluster 0 predominantly consisted of articles related to mathematics, while Cluster 1 focused on topics in astrophysics and computer science. This homogeneity within clusters underscores the effectiveness of our clustering approach in grouping similar articles together.

4. **Dimensionality Reduction and Visualization**: We employed TruncatedSVD to reduce the dimensionality of the dataset, allowing us to visualize the clustering results in two and three dimensions. These visualizations provided valuable insights into the distribution of data points and the separation of clusters in reduced dimensional space.

5. **Interpretation of Clusters**: The clusters generated by our analysis represent distinct themes or topics within the dataset. By examining the articles contained in each cluster, we gained a deeper understanding of the underlying structure and content of the scholarly articles, enabling us to discern patterns and relationships between different fields of study.

6. **Creation of Auxiliary Resource**: Additionally, we created a JSON file mapping category aliases to their corresponding names, facilitating better understanding and interpretation of the categories within the dataset.

In conclusion, our analysis has shed light on the structure and content of the dataset, facilitating the identification and exploration of distinct themes and topics. The clustering results obtained from our analysis provide a framework for organizing and categorizing scholarly articles, thereby contributing to the broader understanding of academic research across diverse fields. Moving forward, these insights can be leveraged for various applications, including recommendation systems, topic modeling, and academic research management.

Bibliography

ChenDataBytes. (2024, January 27). Master the art of feature engineering and feature selection. Medium. https://medium.com/@chenycy/master-the-art-of-feature-engineering-and-feature-selection-e6e87f76f89b

Raviraj. (2023, August 27). Lemmatization in Natural Language Processing (NLP) with Python Example. Medium. https://medium.com/@ravirajpatil871/lemmatization-in-natural-language-processing-nlp-with-python-example-ad338bc2fa94

Ramos, J. (2003, December). Using TF-IDF to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, No. 1, pp. 29-48). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c

Y, D. (2024, February 13). A Comprehensive Guide to TF-IDF: Understanding term frequency and inverse document frequency. Medium. https://medium.com/@digvijay.qi/a-comprehensive-guide-to-tf-idf-understanding-term-frequency-and-inverse-document-frequency-5b7a9bc01539

Sharma, P. (2024b, February 20). The Ultimate Guide to K-Means Clustering: Definition, Methods, and Applications. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

X. Zhu, J. Sun, Z. He, J. Jiang and Z. Wang, "Staleness-Reduction Mini-Batch KMeans," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2023.3279122.

Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. In IOP Conference Series: Materials Science and Engineering (Vol. 725, No. 1, p. 012128). IOP Publishing. DOI 10.1088/1757-899X/725/1/012128