Mark White

Info 4670

11/17/24

**Cluster Analysis**

**1. Introduction:**

* Clustering is a data analysis technique that groups similar objects into clusters based on their characteristics. This assignment allows us to work with clustering methods to demonstrate how helpful they can be in analyzing data.

**2. Data Description:**

* The data set includes attribute columns such as Company, Location, Date, Employee Status, Job Title, Pros, Cons, Ratings, Advice to management, Voice notes, and URL source to to the full review.
* The dataset has a combination of structured data and unstructured data. Where this project focuses on the “Pros” column, which provides feedback from employees

**3. Clustering Methods:**

* **Clustering Overview:** Define clustering and explain how it groups similar objects into clusters. Mention real-world applications like market research and document classification.
* Clustering - groups similar objects into their cluster based on characteristics. Other real-world applications can be seen in healthcare, as it can be important to group patients with similar symptoms for a proper diagnosis or treatment. Furthermore, facial recognition may need similar data sets or patterns to execute properly
* **Types of Clustering Methods:**
  + **Partitioning Method (e.g., K-Means):** This method partitions data into a defined number of clusters (k). It also assigns points to the nearest cluster center and recalculates the centroid.
  + **Other Methods (e.g., Hierarchical, Density-Based, etc.):** Hierarchical Cluster builds a tree of clusters, merging/splitting them based on distance. Density-based clustering forms clusters based on areas of high density and proceeds to identify outliers.

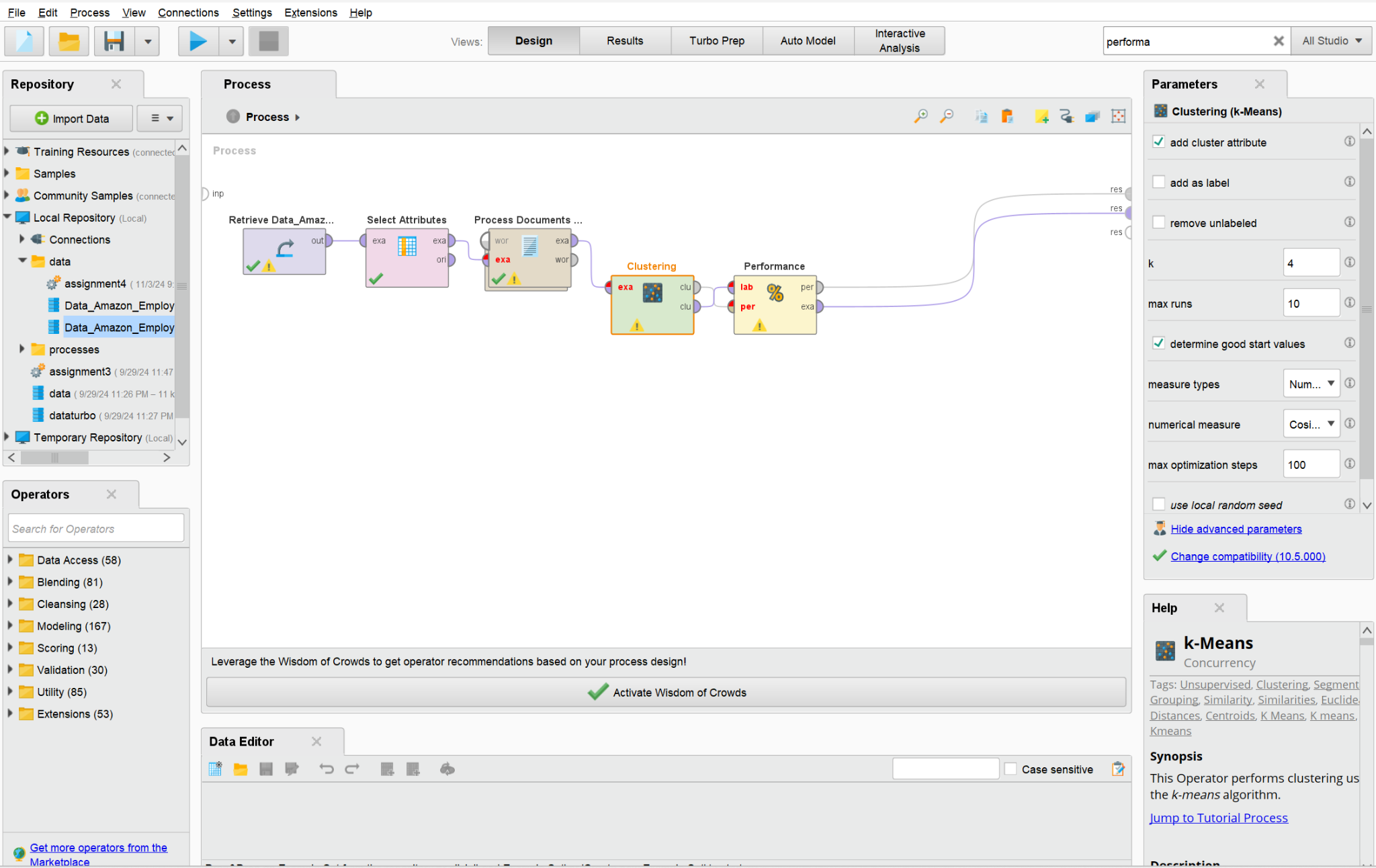
**4. K-Means Clustering:**

* **Algorithm Overview:** The k-means algorithm partitions data into (k) clusters by minimizing cluster distance. It Initializes k cluster centers randomly, assigns each point to the nearest cluster center, and updates the cluster center by avera
* **Implementation:** K is set to 4, and operators tokenizing, transforming cases, filtering stopwords, and creating TF-iDF vectors are included in the implementation. The distance metric used
* **Advantages and Disadvantages:** Advantages include its easy implementation and great for large datasets. Disadvantages include sensitivity to cluster centers and outliers.

**5. Operators Used:**

* **Select Attributes:** This operator focuses on the analysis of the Pros column and will filter out irrelevant data.
* **Process Documents from Data:** The operator transforms text into word vectors using the TF-IDF method, ensuring meaningful numerical representation
* **Filter Examples:** This operator is used to filter data based on specific criteria such as release year and other context.
* **K-Means Clustering:** This operator groups the data into 3 clusters using Cosine Similarity to measure distance, providing insight into textual patterns.
* **Cluster Distance Performance:** The operator evaluates the cluster quality by measuring the compactness and separation of clusters based on centroid distances

**6. Final Analysis Diagram:**

* 

This design process starts by importing data from the repository. The following operator selects the relevant attribute (Pros column) using the select attributes operator. The text data is processed through the Process Documents from the Data operator, performing processes such as tokenization, case transformation, stopword removal, and filtering tokens by length. A word vector is then created using the TF-IDF method within this operator. Furthermore, the K-means cluster is implemented along with the performance operator to evaluate the quality of clustering results by assessing the distance between data points and their cluster centroids.

**7. Cluster Results:**

* **Centroid Table and Interpretation:** The centroid table gives the center points of each cluster by summarizing the most representative characteristics of the data in that group. It examines the high frequency of words in each centroid, identifying themes and topics defining each cluster.
* **Cluster Sorting and Topic Assignment:** After sorting the clusters in descending order, you come across clusters such as career growth and opportunities, work environment and team dynamics, work-life balance and challenges, company reputation, and satisfaction.

**8. Re-running the Analysis with Different k-Values:**

* Change the k value (between 5 and 10) and re-run the analysis. Compare the results for different k values (e.g., k=4, k=7), including differences in the size and number of clusters.

**9. Experience and Learnings:**

* This assignment helps me become more comfortable with Rapid Miner and learn the different operators along with their functions.

**10. Conclusion:**

* Cluster analysis can help identify distinct themes in your data set offering valuable insights. Clustering allowed for an effective way to group similar data and contained steps such as tokenization, stopword removal, and other operators to provide proper clustering results. Overall clustering is very useful for large datasets