Certification Project

Clustering of BBC News Articles

edureka!

Certification Project

You are given a zip file which contains summaries of news from BBC. The Data is taken from Kaggle. (https://www.kaggle.com/pariza/bbc-news-summary)

The zip file contains a folder: 'BBC News Articles '

This folder contains 5 sub folders, named:

1. Business
2. Entertainment
3. Politics
4. Sports
5. Tech

Each of these subfolders contains text files which have summaries of different news articles.

These are the tasks which you have to perform:

- Read all the files from all subfolders and store their summaries in a single CSV file. Name CSV File as: "BBCNewsArticles.csv"

The CSV should contain:

Article
<The text from file 1 >
<The text from file 2>
<The text from file 3>
……

- Randomly arrange the data

- Preprocess each article using Text Preprocessing

- On the preprocessed text, perform Vectorization using 3 types of vectors.
    1. "PresenceAbsenceVector": Converts Article to vectors using Presence and Absence of Words
    2. "CountVector": Converts Article to vectors using Count of Words

3. "TF-IDFVector": Converts Article to vectors using TF-IDF vectorization

- Perform clustering on the dataset using all 3 different types of vectorizations. The number of clusters should be 5.
  You can choose any appropriate clustering algorithm of your choice.
  Make models for each type of vectorization. We will have a total of 3 models.

- Save the Clusters Label for each model in a new CSV file named "BBCNewsArticlesClustered.csv"

- Evaluate and compare the performance of 3 models on basis of Silhouette Coefficient.

- Provide Visualizations for all 3 models. You can show scatter plots and bar graphs.

- Provide your explanation for the following questions

  1. What does Silhouette Coefficient tell us?
  2. Which algorithm you chose and why?
  3. Can you provide an appropriate name to a cluster label? If yes, then explain your observations.
  4. Which vectorization technique is the best and why?