Spring 2024

BEM 106: Homework #6 Description

Goal: The goal of this homework is to get you started with clustering and text analysis using Python's scikit-learn package.

Part 1 - Clustering

The dataset (hw-6-dataset.xlsx) contains various features of 42,305 songs on Spotify. You can find more information the features here:

https://developer.spotify.com/documentation/web-api/reference/get-several-audio-features

Your task is to reverse-engineer the Spotify "genre" classification scheme.

1. Using K-means/Bisecting K-means/PCA/feature selection, experiment with (at least) 2 ways of deriving the "genre" clusters.
2. For each trial, describe how you chose the features and which clustering method you used.
3. Finally, report the Rand score of each trial between your assigned clusters and the true clusters. The Rand score is a measure of cluster similarity, see rand_score in the scikit-learn package.

(PS: There are 15 Spotify-provided genres.)

Part 2 – Sentiment in Text

The dataset (hw-6-dataset2.json) contains Amazon reviews and ratings for roughly 800,000 patio and lawn equipment products.

Your tasks are the following:

1. Figure out how to load this .json file into a Python dataframe.
2. Using either AFINN or VADAR, obtain the "sentiment" for each review.
3. Plot the distribution (as a density plot) of your sentiment scores for each Amazon rating. That is, create 5 separate density plots for 1-star, 2-star, etc. reviews. Do different ratings have different distributions of sentiment? Is it what you would expect, e.g. 1-star reviews have a distribution of sentiment scores that is has a lower mean than 2-star reviews?