TASK 2

Music Genre Analysis

Rules and Guidelines

- Allowed Libraries: You can only use *numpy, pandas*, and plotting libraries like *matplotlib* or *seaborn* for coding-related tasks.
- Algorithm Implementation: If you wish to use any algorithm (e.g., PCA, clustering), you must code it using only the allowed libraries. Pre-built implementations using libraries like *Scikit-learn* are not permitted.

Task

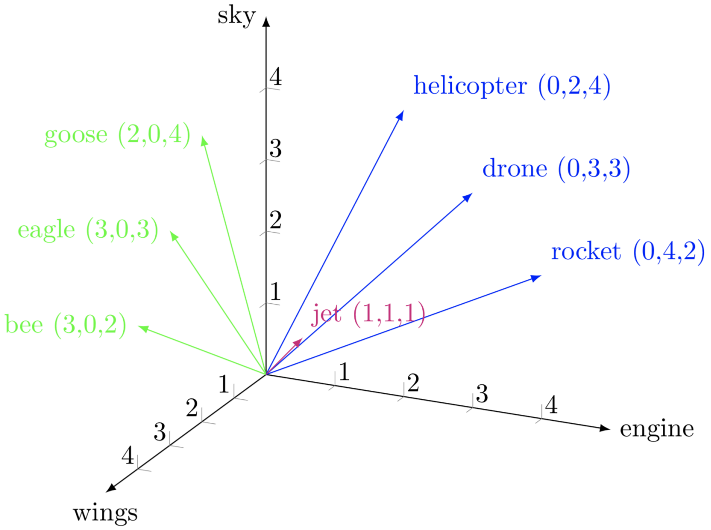
Dataset: LINK

The dataset contains songs described by **three keywords** (which mainly give insights about the instrument used, the mood of the song, and the style of the song), and each song has been labeled with its corresponding genre as a ground truth. The ground truths should not be used in any part other than the final analysis. Your task is to extract insights from the data, group similar songs together, provide actionable insights using the ground labels, and write about your findings in a clear and concise report.

<u>The following is the pipeline you are supposed to follow:</u>

1. Generate vectors for the three keywords

Core idea?



Vectorization, in the context of text analysis and NLP, refers to the process of converting textual data (like words or short phrases) into numerical vectors. These vectors can then be used for various tasks, which in our case, would be used for further analysis.

Resources: ▶ *Vectoring Words (Word Embeddings) - Computerphile*
vectorization-techniques-in-nlp-guide

What are you supposed to do?

- You are supposed to generate embeddings/vectors for the keywords given for all the songs by mainly using popular classical embedding techniques BoW(Bag of Words) or Tf-Idf.

- Compare the two methods and decide which one works best for you. Justify your choice in the report.

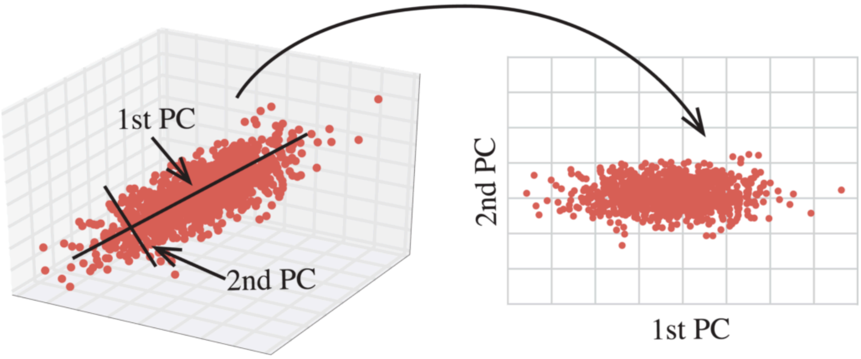
- After deciding your technique of choice, generate vectors for the keywords.

  Resources:

  1. gentle-introduction-bag-words-model

2. *text vectorization tf-idf*

2. Dimensionality reduction



Core idea?

Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much meaningful information as possible. It simplifies data visualization, reduces computational costs, and improves model performance by removing redundant or irrelevant features. Techniques like *PCA (Principal Component Analysis)* achieve this by projecting data into a lower-dimensional space.

Resources: ▶ *StatQuest: Principal Component Analysis (PCA), Step-by-Step*
▶ *StatQuest: PCA - Practical Tips*

What are you supposed to do?

- You are supposed to use these three vectors and apply *PCA (Principal Component Analysis)* on these three vectors to get **three** 2-dimensional vectors for the same.

- If you are struggling with coding PCA from scratch, a small help to make your journey easier lies within Numpy.

3. Combining the embeddings into one

Core idea?

Very often, there are situations where multiple embeddings/vectors are present that give us a very segmented picture of a said document, so there are many techniques that people use to combine those embeddings into a single embedding to get a complete picture of the document. This could range from something as simple as taking averages to training a complex Feed Forward Neural Network.

You are suggested to focus on those simple techniques, as those complex methods are out of the scope of this task.

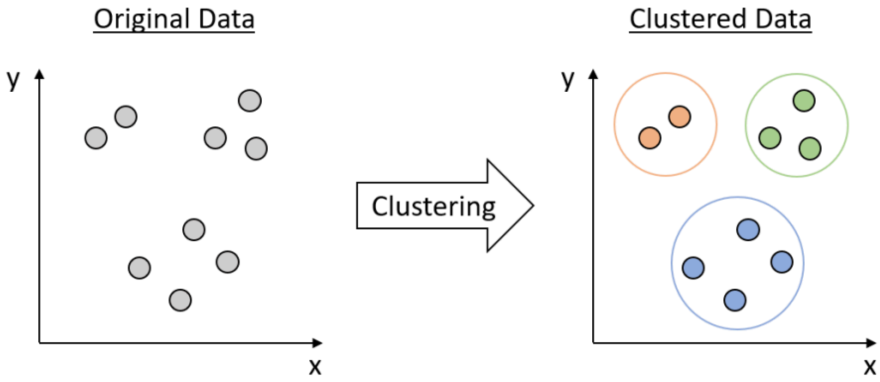
What are you supposed to do?

- You are supposed to use these three dimensionally-reduced vectors and combine them into a single embedding so that the rest of the workflow and analysis becomes easier to work with.

- There are no restrictions regarding the method one can use in this section. Some of the suggested methods could be taking an average or taking the cross-product of those embeddings.

- *Try to justify* your chosen technique in the report.

4. Clustering



Core idea?

Clustering is an unsupervised machine-learning technique that groups data points into clusters based on their similarity. The goal is to ensure that points within the same cluster are more similar to each other than to points in different clusters. It is widely used in pattern recognition, customer segmentation, and anomaly detection.

You are advised to use K-Means, which is a popular clustering algorithm that partitions data into *k* clusters.

Resources for K-Means:
▶ *StatQuest: K-means clustering*
▶ *Stanford CS229: Machine Learning | Summer 2019 | Lecture 16 - K-means…*

What are you supposed to do?

- You are supposed to use the final combined embedding you got in the previous step to carry out clustering and get the clusters to which each song belongs.

- If you are using K-Means or its variants, justify the value of k chosen, or if you are using some other clustering technique, justify why you chose those methods with the consideration of the task in hand.

- Plot a suitable curve that helps you visualize the distribution of the final set of clusters.

5. Analysis

What do we expect from the analysis part?

- What is the percentage distribution of ground truth genres in each cluster?
- Do the clusters align with the true genres, or are there significant overlaps?
- Calculate the Silhouette Score for your clustering method and interpret the result.
- Based on your understanding of the dataset, come up with a technique to assign a genre to a new song if only its keywords are present. Hence based on your ideas, assign genre to : *[piano, calm, slow], [guitar, emotional, distorted], and [synth, mellow, distorted]*

What do we expect from the report?

- The report should preferably be in *white paper format*

- It should contain clear and concise explanations of your thought process, methodology, and results.

- Try to include plots and visualization to back your results and make the report presentable.

6. Bonus

(These bonus tasks are optional but can help you stand out by showcasing creativity and deeper analytical skills. Successfully completing them will earn you extra points.)

- Try to think of some ingenious vectorizing technique and report the results. It doesn't have to be anything complex, something simple works.

Example ***(DO NOT USE THIS)****: A 26-dimensional embedding where each dimension represents a letter of the alphabet, and the values are populated by the frequency of each letter in the keywords.*

- Try to explore the dataset on your own to gain more analytical insights and report them. The analysis could range from Genre level analysis to Cluster level analysis or even if you want to analyze a specific keyword itself, the data is all yours to play with.

- Besides the Silhouette Score, which is an intrinsic metric in nature, there is another class of metrics for clusters known as extrinsic metrics. Try to explore both kinds of metrics and solidify your analysis.

(NOTE: To report the findings of Bonus tasks, add a section named Bonus in the report and write about your findings and methodology under it)

7. Final Submission Documents

The final documents that are to be submitted for this task are as follows:

- An Ipynb Notebook. The notebook should be error-free, and all cells should run when the "run all" command is clicked. All necessary graphs, tables, and markdown cells should be present to explain how the notebook works.

- A report should be submitted in PDF format. The contents of the report are discussed in section 5.

Additional Resources:

1. *clustering-overview-scikit learn*
2. *A Nearly Tight Analysis of Greedy k-means++*