**Data source: https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge**

**How to run:**

* KMeans Instruction to train KMeans model
  + Download and unzip the file
  + Place the unzipped folder inside this directory
  + Install all required python libraries with pip
    - Pandas
    - Numpy
    - Scipy
    - Matplotlib
    - Nltk
    - Wordcloud
    - Stop\_words
    - Gensim
    - Pickle
    - Sklearn
    - yellowbrick
  + Run kmeans\_model\_final.ipynb
  + The run time for using 1 JSON file as raw data is about 6 hours (If you want to train the model with more data, you can change the variable number\_of\_json\_file.)
* KMeans Instruction to use Recommender System
  + Install all required python libraries with pip
    - Pandas
    - numpy
    - pickle
  + Make sure the kmeans\_model\_nlp.sav located in the same directory as recommender\_system.ipynb
  + Run recommender\_system.ipynb
  + Change the artist name and track name or recommended size if you want (NOTE: case sensitive)
* KMeans evaluation analysis
  + Redefine the variables ‘tracks\_data’, ‘test\_tracks\_data’, ‘Model’ in evaluation notebook‘recommender\_system.\_eval.ipynb’
  + Using TensorFlow API ‘tf.keras.models.load\_model’ to load the model.
  + The saved model named ‘kmeans\_mode\_nlp.sav’, can be directly imported and used to predict.
  + The saved dataset named “preprocessed\_test\_dataset.csv”, can be directly imported and used as ‘tracks\_data’ variable.
  + The saved train data set named “trial.csv”, can be directly imported and used as ‘test\_tracks\_data’ variable.
  + Then, run the notebook, ‘VG16\_fea\_reduc.ipynb’ line by line.
* Word2Vec
  + All code is saved in ‘final\_submission/code/word2vec/song2vec.ipynb’ file.
  + The preprocessed dataset for train and test has been saved as ‘train\_10000.txt’, and ‘test\_10000.txt’.
  + The ‘text\_line\_reader’ is used to load the train and test txt files.
  + The trained model has been saved as ‘w2v\_10k\_final’.
  + Genism model load API ‘models.Word2Vec.load()’ is used to load the model.
  + If you want to check the whole process of the project, you can run the notebook line by line.

**Team Contribution:**

* Chujie Cai
  + Literature review
  + Discussion and future improvements
* Fengling Zhou
  + Methodology (Kmeans)
* Mavis Tsoi
  + Experiment and Evaluation
* Haixu Ji
  + Introduction
  + Motivation
* Xiuzhong Yang
  + Methodology (word2vec)
  + Experiment and Evaluation (word2vec)