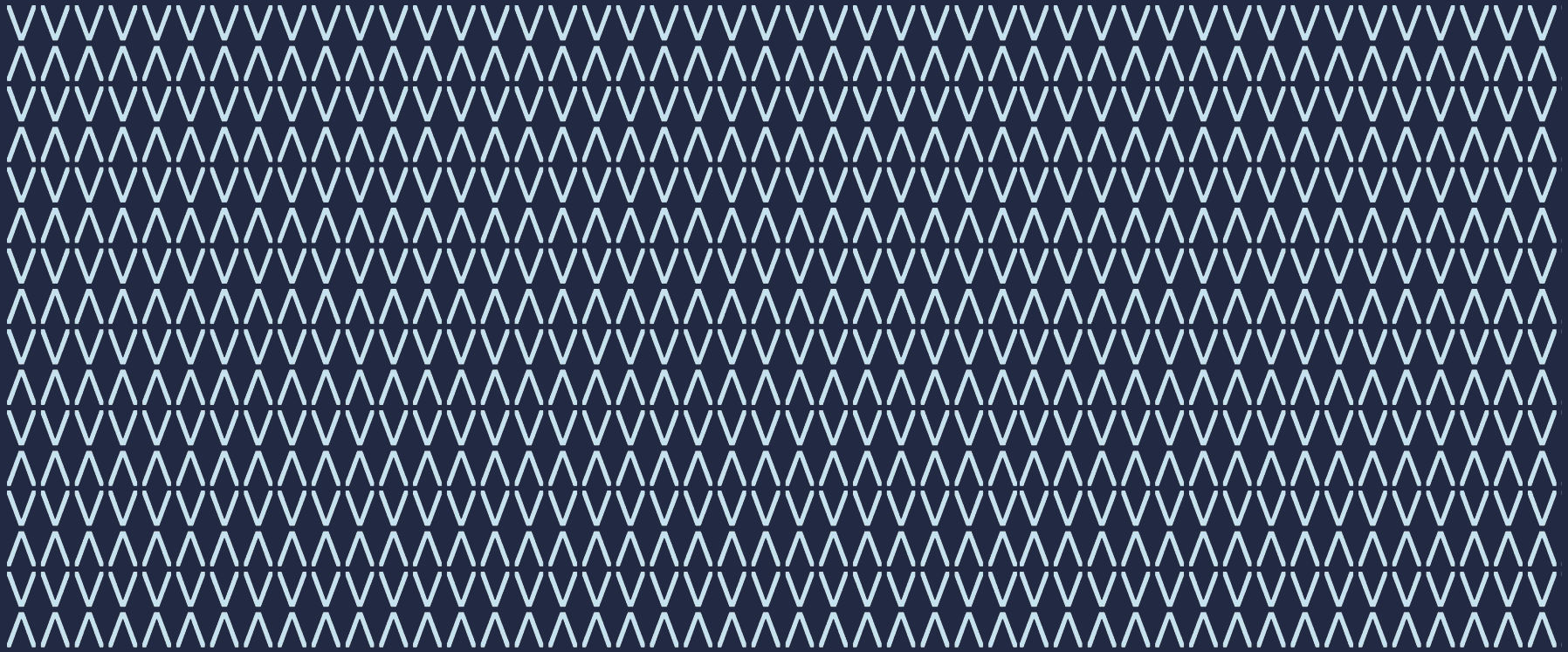
EVIDENT

# Data Science Interview Task - Beatles Lyrics analysis



# Background

Ahead of the 60th anniversary of the Beatles debut album Please Please Me being released Evident has been commissioned by a media agency to do an analysis of the lyrics of the Beatles song catalogue. To add even more excitement 5 previously unreleased songs have been discovered and will be available for fans to listen to for the first time.

They are particularly interested in the themes of the Beatles songs, and if they can be grouped together into clusters of songs. They believe this will be important so they can quickly discover what other songs the new songs should be compared to - whether it be peace, love, or any other topic.

As the media agency does some kind of celebration every few years, they have already collected all the lyrics for you and supplied them as text files. A quick look indicates that there may be duplicates in the files where there are additional live recordings that have the same lyrics, and that some files might have weird or missing data - the media company has mentioned that if anything looks amiss then it should be removed.

The media company can be very fussy, so they expect to see a small powerpoint deck outlining the results, and justifications and explanations of any methodology that is used.

# The Task

Your task is to support the media company with interesting insights identified by the Evident product manager, to group the existing songs, and to classify the new releases. This should involve:

- Remove any files that are duplicates from the set of 279 lyrics files, and remove any files that you deem to have missing informatio
- Answer **any 2** of the following 5 questions, along with a short explanation of your methodology:
    - How many of the songs feature the song name (found in the file name) in the song lyrics?
    - How many of the songs feature at least one pair of lines that rhyme?
    - Which song has the largest amount of repetition?
    - Which songs are the most similar?
    - Which song has the most mentions of peace and love?
- Derive 1 other piece of insight from the data that you find interesting, and visualise it - that could be as simple as a wordcloud of most popular songs, or you could identify previously unknown links between song title length and number of lyrics, it's up to you.
- Through the use of Machine Learning or other techniques, classify the songs into topic clusters. The number of topics is up to you, but you should be able to justify your final number of clusters, any metrics used to justify your clusters, and you should also be able to identify what the clusters of songs mean.
- You then need to predict which cluster the new 5 beatles songs belong to, and be able to explain why a song was placed in that cluster.
- Finally this analysis, with explanations, should be presented in a small slide deck that you will then present back to Evident.

# Contents + Other Details

## Contents:

The .zip file you have received contains a large number of files. It should contain:

- This PDF containing the instructions
- 279 files in a “lyrics” folder that are original lyrics by the Beatles to be analysed
- 5 files in a “new lyrics” folder, that are new songs by the Beatles soon to be released

## Other Details:

- We expect this task to take around a couple of hours, plus time to present your analysis - we hope this is a fairly fun task, so feel free to spend more or less time on the task as you wish, but we don’t expect hours of detailed analysis and prefer quality or quantity.
- Song lyrics can be unusual compared to other text - do not be worried if you end up with unusual cluster types or analysis. We are more interested in your approach to the problem than the results.
- You may do the analysis in any order you like, and you are allowed to use your topics as part of your insights.
- Your analysis should be done in Python, and your code submitted alongside your analysis for review.
- You may use any Python packages that you wish to achieve the goals of the analysis, but please leave the imports in the file so we can test your code.
- If there are any issues, questions or concerns then please feel free to reach out to Andrew (andrew@evidentinsights.com)