**Name: Megha Bista**

**Project Name: Taylor Swift’s Song classification and analyzation**

**1. Introduction:**

Text mining also known as text analysis is the process of transforming unstructured texts into structured data for easy analysis. It uses natural language processing (NLP) which allows machines to understand and process human language automatically by classifying texts by sentiments, topics, and intent. Text mining has allowed large and complex datasets to be analyzed in simple yet effective ways. Text mining can be used in several different fields, but it is especially beneficial in the field of business due to its scalability, real-time analysis, and its consistent criteria.

**2. Dataset:**

The dataset that I will be using for this project is from Kaggle. It was created by [PromptCloud](https://www.promptcloud.com/) using the API exposed by Genius.com. The format of this dataset is a CSV file. It contains the album name, track title, track number, lyric text, the line number of the lyric in the track, and the year of release of the album.

**3. Data Preparation:**

The first process of text mining is gathering the data. Then the next process is preparing your data using several natural language processing techniques such as tokenization, parsing, lemmatization, stemming, and stop removal. These techniques aid in building the input of the machine-learning model. Then the final step is to use text mining methods such as text classification and text extraction.

**4. Exploratory Data Analysis**

After the data has been prepared, the next step is to perform exploratory data analysis (EDA).EDA involves visualizing and summarizing the data to gain insights into the relationships between the variables. For this dataset, EDA can involve creating pair plots to visualize distributions of single variables as well as relationships between two variables. Ngram, Bigram, and Trigram analysis can be done to see the comparisons in the lyrics from different years. Scatterplots can also be created to visualize linguistic variation between categories.

**5. Modeling**

The next step is to select modeling techniques. Three algorithms that I will be using for this data are k-nearest neighbor (SVMs), Naive Bayes, and DL methods. The goal of modeling is to analyze, predict, and categorize different aspects of the lyrics from different years. The modeling process involves splitting the data into training and testing sets, training the model on the training set, and then evaluating the model's performance on the testing set.

**6. Evaluation**

Model evaluation is an important step in text mining. This involves calculating performance metrics, such as accuracy, precision, recall, and F1 score to determine how well the model performs. For this dataset, I will be using ROUGE metrics because they calculate the lengths and number of sequences overlapping between the original text and the extracted text. It needs to be defined manually.