Our team, Caffeinated Quantum Squadron, intends to create a dataset for tweet sentiment analysis using Apache Airflow and a DAG integrated into a pipeline. Constructing such a pipeline involves a series of steps, including data collection, data cleaning, data classification, data storage, and data uploading.

- **Data Collection**: We plan to use Twint, a Pythonic library that allows us to collect Twitter data without having to obtain Twitter API credentials.
- **Data Cleaning**: After collecting the data, we will remove irrelevant information within the tweet text, such as Twitter handles, hashtags, URL links, HTML references, and punctuations, and apply stemming to reduce the dimensions of the data.
- **Data Classification**: We will label the tweets with their corresponding sentiment using the pre-trained TextBlob sentiment analysis model, which uses a Naïve Bayes classifier trained on a large corpus of labeled data to predict the sentiment of a given text. Specifically, the sentiment scores range from -1 to +1 and will be mapped to one of the three sentiment labels: positive, negative, neutral. Each tweet will be fed into TextBlob model, and the new classified information will be added within the existing DataFrame.
- **Data Storage**: The resulting labeled and preprocessed tweet data will be stored in a PostgreSQL database management system within Airflow.
- **Data Uploading**: The final task within the pipeline will upload the resulting dataset from PostgreSQL to Kaggle using the Kaggle API.

To create an organized and effective pipeline, we break down each of the steps listed above into individual tasks and arrange them in a specific order within the pipeline.