NLP Analysis of Google Reviews for Saudi Arabian Sites

#### Problem Statement

#### Online reviews are a goldmine of customer feedback, offering invaluable insights into customer experiences and expectations. For businesses and tourist destinations in Saudi Arabia, Google reviews serve as a critical barometer of their service quality, popularity, and areas needing improvement.

#### Despite the richness of the data available in Google reviews, extracting actionable insights can be challenging due to the unstructured nature of text data and the complexity of sentiment and preference nuances. Moreover, the presence of data in various formats like JSON in the tags and ratings columns adds another layer of complexity to data processing and analysis.

#### Objective

#### The goal of this assignment is to develop a comprehensive analytical approach using Natural Language Processing (NLP) to:

#### Transform and preprocess raw review data, extracting structured information from JSON-encoded columns.

#### Perform sentiment analysis to categorise reviews as positive, neutral, or negative.

#### Employ text cleaning and NLP techniques to analyse the content of reviews, identifying common themes, keywords, and categories that emerge from customer feedback.

#### Conduct an exploratory data analysis to uncover trends and patterns related to offerings, destinations, and overall customer sentiment.

#### Dataset Provided:

* Google Reviews Dataset: A CSV file containing 10k rows of Google reviews for multiple sites in Saudi Arabia. Key columns include tags, ratings, and the review text.
* Mapping File: A supplementary JSON file providing mappings for hashkeys in the tags column to specific offerings (like accommodation, tourist attractions) and destinations (like Riyadh, Jeddah).

#### Assignment Tasks:

* **Data Preprocessing and Transformation:**
  + Extract offerings and destinations from the tags column using the provided mapping file.
  + Generate new columns from the ratings column to represent actual user ratings.
  + Derive a sentiment column using a sentiment analysis method to categorise reviews as positive, neutral, or negative.
* **Text Cleaning and NLP Analysis:**
  + Apply NLP techniques to clean the review text (removing stop words, stemming, etc.).
  + Conduct text analysis to identify common keywords and derive potential categories/themes discussed in the reviews.
* **Exploratory Data Analysis (EDA):**
  + Analyze the distribution of sentiments, offerings, destinations, and ratings.
  + Investigate patterns or correlations between different variables (e.g., relation between sentiment and ratings).

#### Deliverables:

* **Jupyter Notebook:**
  + Well-documented and modular code for each step of the process.
  + Visualisations illustrating key findings from the EDA.
* **PDF Report:**
  + Methodology: Detailed description of the approaches used for sentiment analysis, text cleaning, and data transformation.
  + Findings: Insights from the sentiment analysis, keyword/category identification, and EDA.
  + Recommendations: Strategic recommendations based on the findings, suggesting how businesses can improve based on customer feedback.

#### Evaluation Criteria:

* Accuracy and Efficiency of Data Transformation: Correct mappings and effective transformation of the columns and hashkeys.
* Effectiveness of Sentiment Analysis: Accuracy of sentiment categorization (will be tested against the real set of sentiment that we held out)
* Quality of Text Cleaning and NLP Techniques: Appropriateness and effectiveness of the methods used for text cleaning and analysis.
* Insightfulness of EDA: Depth and relevance of insights derived from the exploratory analysis.
* Clarity and Structure of Deliverables: How well the findings, methodology, and recommendations are communicated in the report