# Text Mining with Sentiment Analysis

## **Assignment Overview**

In this assignment, you will explore and analyze a dataset for sentiment analysis using text mining techniques. The dataset contains Amazon product reviews, and your task is to apply various text mining methods to extract valuable insights. Please follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework throughout the assignment.

The assignment involves conducting a comprehensive sentiment analysis using text mining techniques. The process is organized into five key sections, each contributing to the understanding, preparation, visualization, modeling, and advanced analysis of textual data for sentiment prediction. From initial data exploration and cleaning to the creation and evaluation of machine learning models, the assignment encompasses tasks such as data visualization, advanced techniques like topic modeling, and comparisons between different models. Additionally, insights derived from the analysis are summarized, actionable recommendations are provided, and potential avenues for future research and improvement are outlined. The assignment is designed to follow the CRISP-DM methodology, ensuring a structured and thorough approach to extracting meaningful insights from textual data in the context of sentiment analysis.

### **Dataset**

The dataset can be downloaded from BrightSpace or the following address:

[Amazon Product Reviews Dataset](https://raw.githubusercontent.com/pycaret/pycaret/master/datasets/amazon.csv)

### **Sections:**

#### **1. Data Understanding and Preparation (20 points)**

* Description:
  + Encompasses loading the dataset, performing exploratory data analysis (EDA), providing basic statistics, and preprocessing text data.
  + Aligns with the "Business Understanding" and "Data Understanding" phases of CRISP-DM.
* Tasks:
  + Load the dataset and perform exploratory data analysis (EDA).
  + Provide basic statistics about the dataset: number of records, features, data types, etc.
  + Preprocess the text data by handling missing values, if any.
  + Perform text cleaning, including lowercasing, punctuation removal, and any necessary steps.
  + Split the dataset into training (70%) and testing sets (30%).

### **2. Data Visualization and Exploration (20 points)**

* Description:
  + Involves using t-SNE to visualize documents in a two-dimensional space, interpreting and discussing the visual representation.
  + Aligns with the "Data Understanding" and "Data Visualization" phases of CRISP-DM.
* Tasks:
  + Utilize t-SNE to visualize the documents in a two-dimensional space.
  + Interpret and discuss the visual representation.

### **3. Model Building and Evaluation (20 points)**

* Description:
  + Encompasses creating machine learning models, feature selection, evaluating their performance, analyzing confusion matrices, and comparing the accuracy of different models.
  + Aligns with the "Modeling" and "Evaluation" phases of CRISP-DM.
  + Encompasses comparing the performance of different models (e.g., BERT, FastText, Bag of words, Word Embedding, …) and exploring the possibility of improving accuracy using advanced models.
* Tasks:
  + Create machine learning models (e.g., Naive Bayes, Random Forest, SVM) to predict the sentiment class.
  + Evaluate the models using appropriate metrics.
  + Analyze the confusion matrix and discuss the model's performance.
  + Reduce the model to predict only three sentiment classes.
  + Compare the accuracy and time from the reduced model with the original model.
  + Evaluate the models using appropriate metrics.
  + Analyze the confusion matrix and the AUC. Discuss the model's performance.
  + Compare the accuracy and time among different models.

### **4. Advanced Techniques and Analysis (20 points)**

* Description:
  + Involves applying advanced techniques like topic modelling, clustering, analyzing word relationships, or any interesting text-mining approach you think can be applied to this dataset.
  + Aligns with the "Modeling" and "Evaluation" phases of CRISP-DM.
* Tasks:
  + Apply topic modelling techniques (e.g., LDA) to identify topics in the dataset.
  + Utilize clustering algorithms (e.g., K-means) for visualization and analysis.
  + Analyze and identify words that are more related to negative or positive reviews.

### **5. List of Insights, Recommendations, and Future Works (20 points)**

* Description:
  + Summarizes key insights derived from the analysis, provides actionable recommendations based on the findings, and outlines potential future works for continued improvement and exploration.
* Tasks:
  + Summarize key insights from the analysis, highlighting patterns or trends.
  + Provide actionable recommendations based on the insights, addressing potential areas for improvement or further investigation.
  + Outline a list of potential future works, suggesting areas for additional research, experimentation, or enhancements.

## **Submission Guidelines**

* The deadline for the submission is Sunday, 17 December at 11 pm. There will be a penalization of 10% per each day of delay. The document was posted on Tuesday 28 Nov 2023.
* There is a Template you should use to present the assignment.
* Submit a report in Word format that includes your visualizations, and analysis for each part. Code should not be included in the document, but it can be added to a separate folder.
* The quality of the document and plots will be considered on marking each section. Also, the quality and number of references to support each statement.
* Clearly label each section corresponding to the assignment requirements.
* Submit the report and any supporting code files.

## **Note:**

* The assignment is presented individually or in groups of two.
* You are encouraged to use any software of your choice (e.g., Orange, R Studio, Python) for different parts of the assignment.
* Make sure to adhere to the CRISP-DM framework and provide clear explanations for each step.
* The assignment will be graded based on the depth of analysis, completeness, and clarity of presentation.
* The report should not exceed 8 pages, not including references, participation in the forum or cover. Visualizations, such as charts and graphs, should support your analysis.
* Submissions should be made through BrightSpace. Email submissions will not be accepted.
* Combine all assignment components into a single report document and convert it into Microsoft Word or Google Docs format. Ensure all images are embedded within the document.

References:

<https://www.datacamp.com/tutorial/text-analytics-beginners-nltk>