# **Project Documentation:**

# **NLP Pipeline for Amazon Food Reviews**

## **Overview**

The purpose of this project is to create a robust Natural Language Processing (NLP) pipeline for analyzing Amazon food reviews. The pipeline incorporates three text representation techniques: Bag of Words, TF-IDF, and Word2Vec. The ultimate goal is to extract meaningful insights from the reviews, allowing for a comprehensive understanding of customer sentiments and preferences.

## **Objectives**

* Data Preprocessing:
  + Conduct Exploratory Data Analysis (EDA) to gain insights into the structure of the text data.
  + Perform text preprocessing to clean and prepare the Amazon food review data for analysis.
  + Further EDA to understand the distribution and characteristics of the preprocessed data.
  + Feature Engineering/Extraction to enhance the quality of the dataset.
* Feature Extraction Techniques:
  + Bag of Words:
    - Tokenize text and create a vocabulary.
    - Vectorize reviews by counting word occurrences.
  + TF-IDF:
    - Implement the standard TF-IDF formula.
    - Highlight important words in each review.
  + Word2Vec:
    - Utilize the gensim library to train a Word2Vec model on the dataset.
    - Generate word embeddings for each review.
* Analysis and Comparison:
  + Evaluate and compare the effectiveness of each model in extracting insights from Amazon food reviews.

## **Tools and Technologies**

* Programming Language: Python
* Libraries:
  + nltk for text preprocessing
  + numpy and pandas for data manipulation
  + gensim for Word2Vec
  + sklearn for Bag of Words and TF-IDF models

## **Dataset**

* Source: Amazon Food Reviews Dataset (from kaggle)

## **Implementation Details**

### **1. Data Preprocessing**

Steps:

a. EDA for text preprocessing:

- Explored the distribution of review lengths, word frequencies, etc.

b. Text Preprocessing:

- Handled missing data, removed duplicates, and performed other necessary data cleaning steps.

c. EDA for understanding the data:

- Further analyzed the preprocessed data to identify patterns and trends.

d. Feature Engineering/Feature Extraction:

- Enhanced the dataset by creating new features or extracting relevant information.

### **2. Feature Extraction Techniques**

#### Bag of Words:

* Tokenized text and created a vocabulary.
* Vectorized reviews by counting word occurrences.

#### TF-IDF:

* Implemented the standard TF-IDF formula.
* Highlighted important words in each review.

#### Word2Vec:

* Used gensim to train a Word2Vec model on the dataset.
* Generated word embeddings for each review.

## **Reporting and Findings**

### **Bag of Words**

* Strengths:
  + Simple and interpretable model.
  + Efficient for large datasets.
* Limitations:
  + Ignores word order and context.

### **TF-IDF**

* Strengths:
  + Considers word importance in the corpus.
  + Provides a measure of the uniqueness of words.
* Limitations:
  + May not capture semantic relationships.

### **Word2Vec**

* Strengths:
  + Captures semantic relationships and word context.
  + Produces dense word embeddings.
* Limitations:
  + Requires a large amount of training data.

## **Conclusion:**

This documentation aims to provide a clear and comprehensive understanding of the NLP pipeline's implementation and outcomes, facilitating further research and development in the field of sentiment analysis and customer feedback interpretation.