### **Building a Sentiment Analysis using IMDB Movie Reviews**

#### **Introduction**

Creating a vocabulary index is crucial in Natural Language Processing (NLP) for analyzing text data. This article outlines a method to generate such an index using the IMDB movie reviews dataset for sentiment analysis, focusing on data loading, text preprocessing, and vocabulary creation.

#### **Loading the Dataset**

We begin by loading the IMDB dataset using the Pandas library. After installing the dataset from Kaggle and checking the format, we know that the file is named IMDB\_Dataset.csv, with a column labelled review.

| df = pd.read\_csv('IMDB\_Dataset.csv') # Load the IMDB dataset |
| --- |

#### **Preprocessing the Text**

Next, we preprocess the reviews to standardize the text. This process involves:

1. Converting text to lowercase.
2. Removing punctuation.
3. Tokenizing the text into individual words.

The preprocessing function is defined as follows:

| def preprocess\_text(text):  text = text.lower() # Convert to lowercase  text = text.translate(str.maketrans('', '', string.punctuation)) # Remove punctuation  return text.split() # Tokenize into words |
| --- |

#### **Combining Reviews into a Single List**

We combine all processed reviews into a single list of words using the following code:

| all\_words = []  for review in df['review']: # Assuming the reviews are in a column named 'review'  all\_words.extend(preprocess\_text(review)) |
| --- |

#### **Creating a Vocabulary**

Using the Counter class, we count the occurrences of each word and create a mapping from each word to a unique index:

| word\_counts = Counter(all\_words) # Count occurrences of each word  word\_to\_index = {word: idx + 1 for idx, (word, \_) in enumerate(word\_counts.items())} # Create mapping |
| --- |

We reserve index 0 for padding purposes.

#### **Saving the Vocabulary Index**

Finally, we save the vocabulary index to a CSV file for easy access:

| word\_index\_df = pd.DataFrame(list(word\_to\_index.items()), columns=['Word', 'Index']) # Convert to DataFrame  word\_index\_df.to\_csv('word\_index.csv', index=False) # Save to CSV |
| --- |

#### **Conclusion**

This process demonstrates the creation of a vocabulary index from the IMDB movie reviews dataset. The resulting word\_index.csv file is essential for effective text representation in NLP tasks. This methodology can be applied to various datasets, providing a solid foundation for further text analysis and machine learning exploration.