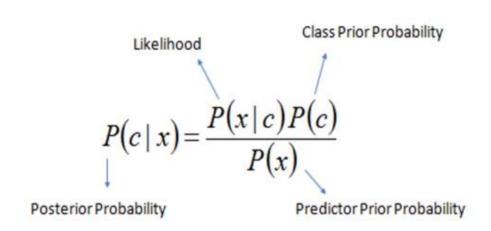
Assignment 2 Part 2

Implementing Naive Bayes Classifier using Spark MapReduce

Naïve-Bayes Classifier:

Naive Bayes Classifier



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Algorithm Description

- 1. Data Loading and Preprocessing:
 - Load the movie reviews dataset into an RDD and remove the header row.
 - Split the data into training (70%) and testing (30%) datasets.
 - Transform the training data into tuples of (label, text) and clean the text by removing punctuation and stop words, then apply stemming to reduce words to their base form.
- 2. Calculate Prior Probabilities:
 - Compute the prior probability for both "positive" and "negative" labels based on the proportion of each in the training dataset.

3. Word Counting:

- Create RDDs for positive and negative reviews. For each label, generate a list of (label, word) pairs.
- Count the total number of words and distinct words in both positive and negative reviews.
- Cache the word counts for positive and negative reviews to optimize performance.
- 4. Conditional Probability Calculation with Laplace Smoothing:
 - Use Laplace smoothing to calculate the conditional probabilities for each word appearing in positive or negative reviews.
 - Calculate the logarithm of the smoothed probabilities for numerical stability.

5. Review Classification:

- For each review in the test dataset, calculate the log-probabilities for both positive and negative sentiments based on the words present in the review.
- Classify the review as "positive" or "negative" based on the higher logprobability.

6. Evaluation Metrics:

- Calculate true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- Compute accuracy, precision, recall, and F-measure based on the counts of TP, TN, FP, and FN.

7. Output Results:

Prior Probabilities:

Prior Positive Probability	0.4988999057062034
Prior Negative Probability	0.5010715204160356

Metrics:

Accuracy	0.85
Precision	0.87
Recall	0.83
F-measure	0.85