1. K Nearest extension

ROW 1

Accuracy for dataset tank: 0.7475

Accuracy for dataset plant: 0.8

Accuracy for dataset perplace: 0.7825

Accuracy for dataset smsspam: 0.947841726618705

ROW 2

Accuracy for dataset tank: 0.7925

Accuracy for dataset plant: 0.7975

Accuracy for dataset perplace: 0.8575

Accuracy for dataset smsspam: 0.947841726618705

ROW 3

Accuracy for dataset tank: 0.7925

Accuracy for dataset plant: 0.7975

Accuracy for dataset perplace: 0.8575

Accuracy for dataset smsspam: 0.947841726618705

ROW 4

Accuracy for dataset tank: 0.7925

Accuracy for dataset plant: 0.7975

Accuracy for dataset perplace: 0.8575

Accuracy for dataset smsspam: 0.947841726618705

ROW 5

Accuracy for dataset tank: 0.7675

Accuracy for dataset plant: 0.785

Accuracy for dataset perplace: 0.8625

Accuracy for dataset smsspam: 0.947841726618705

ROW 6

Accuracy for dataset tank: 0.7975

Accuracy for dataset plant: 0.775

Accuracy for dataset perplace: 0.86

Accuracy for dataset smsspam: 0.947841726618705

ROW 7

Accuracy for dataset tank: 0.6525

Accuracy for dataset plant: 0.625

Accuracy for dataset perplace: 0.6675

Accuracy for dataset smsspam: 0.8812949640287769

ROW 8

Accuracy for dataset tank: 0.7875

Accuracy for dataset plant: 0.7375

Accuracy for dataset perplace: 0.7425

Accuracy for dataset smsspam: 0.8812949640287769

ROW 9

Accuracy for dataset tank: 0.7875

Accuracy for dataset plant: 0.7375

Accuracy for dataset perplace: 0.7425

Accuracy for dataset smsspam: 0.8812949640287769

ROW 10

Accuracy for dataset tank: 0.7875

Accuracy for dataset plant: 0.7375

Accuracy for dataset perplace: 0.7425

Accuracy for dataset smsspam: 0.8812949640287769

ROW 11

Accuracy for dataset tank: 0.735

Accuracy for dataset plant: 0.735

Accuracy for dataset perplace: 0.74

Accuracy for dataset smsspam: 0.8812949640287769

ROW 12

Accuracy for dataset tank: 0.775

Accuracy for dataset plant: 0.71

Accuracy for dataset perplace: 0.745

Accuracy for dataset smsspam: 0.8812949640287769

2. Bayes extension

ROW 1

Accuracy for dataset tank: 0.7975

Accuracy for dataset plant: 0.9475

Accuracy for dataset perplace: 0.82

Accuracy for dataset smsspam: 0.9640287769784173

ROW 2

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.935

Accuracy for dataset perplace: 0.87

ROW 3

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.935

Accuracy for dataset perplace: 0.87

Accuracy for dataset smsspam: 0.9640287769784173

ROW 4

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.935

Accuracy for dataset perplace: 0.87

Accuracy for dataset smsspam: 0.9640287769784173

ROW 5

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.9475

Accuracy for dataset perplace: 0.8725

Accuracy for dataset smsspam: 0.9640287769784173

ROW 6

Accuracy for dataset tank: 0.89

Accuracy for dataset plant: 0.94

Accuracy for dataset perplace: 0.8675

Accuracy for dataset smsspam: 0.9640287769784173

ROW 7

Accuracy for dataset tank: 0.7975

Accuracy for dataset plant: 0.9475

Accuracy for dataset perplace: 0.82

ROW 8

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.935

Accuracy for dataset perplace: 0.87

Accuracy for dataset smsspam: 0.9640287769784173

ROW 9

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.935

Accuracy for dataset perplace: 0.87

Accuracy for dataset smsspam: 0.9640287769784173

ROW 10

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.935

Accuracy for dataset perplace: 0.87

Accuracy for dataset smsspam: 0.9640287769784173

ROW 11

Accuracy for dataset tank: 0.875

Accuracy for dataset plant: 0.9475

Accuracy for dataset perplace: 0.8725

Accuracy for dataset smsspam: 0.9640287769784173

ROW 12

Accuracy for dataset tank: 0.89

Accuracy for dataset plant: 0.94

Accuracy for dataset perplace: 0.8675

3. Explanation

Based on the results, the performance of the two extensions, K Nearest Neighbors (KNN) and Naive Bayes.

K Nearest Neighbors (KNN) Extension:

1. Functions added are:
    a. functions are for evaluating the accuracy of a document classification system using similarity measures.
    b. The `calculate_similarity_values` function computes similarity values between development documents and training documents using specified similarity functions.
    c. Then, `count_label_occurrences` tallies occurrences of labels in the k most similar documents.
    d. Finally, `evaluate_predictions` assesses the accuracy of predictions based on these similarities.
    e. Overall, the system evaluates how well the training set predicts the labels of development documents.
2. The accuracy of the KNN extension across all datasets ranges from 0.6525 to 0.7975.
3. The average accuracy of KNN extension across all datasets is approximately 0.7658.

Bayes Extension:

1. Functions added:
    a. These functions are for computing term frequency sums, log likelihood ratios, and evaluating document classifications.
    b. `compute_term_frequency_sum` sums up term frequencies in document vectors.
    c. `compute_log_likelihood` calculates the log likelihood ratios of term frequencies between two classes.
    d. `evaluate_dev_docs` utilizes similarity functions to classify documents and assesses accuracy based on predicted labels against actual labels.
    e. Overall, these functions aid in text classification tasks by quantifying term frequencies and evaluating classification performance.
2. The accuracy of the Bayes extension across all datasets ranges from 0.7975 to 0.9640.
3. The average accuracy of Bayes extension across all datasets is approximately 0.8917.

Analysis:

- I used stemming and remove stop words for both but it didn't have much impact on accuracy

- The Bayes extension consistently outperforms the KNN extension across all datasets.
- The Bayes extension shows higher accuracy values compared to the KNN extension in every case.
- The average accuracy of the Bayes extension is significantly higher than that of the KNN extension.

Bayes extension performed significantly better than the KNN extension across all datasets. Therefore, the Bayes extension would be preferred for the given classification tasks due to its higher accuracy.