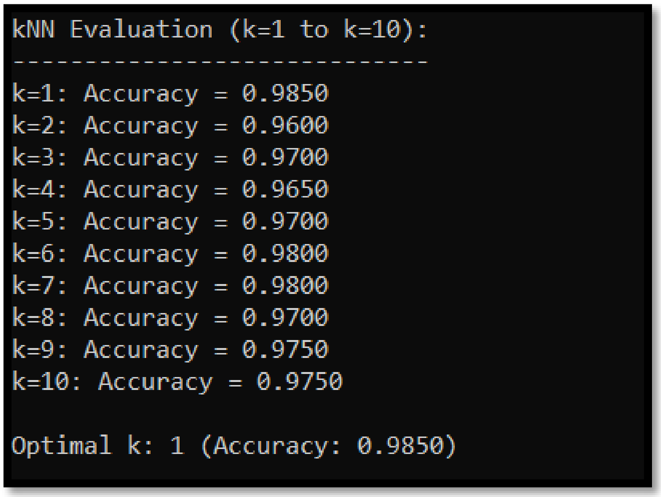
CS 7800 Information Retrieval

Assignment-2

1- NAME: Bodhu Shravya UID: U01122915 EMAIL: bodhu.2@wright.edu
2- NAME: Lohitha Donuri UID: U01125638 EMAIL: donuri.3@wright.edu
3- NAME: Niharika Kanugovi UID: U01108474 EMAIL: Kanugovi.2@wright.edu

Running KNN on different K values to find the optimal K:

To find the optimal K value we have run the knn model from ka values 1 to 10 and we have got the below results.



```
kNN Evaluation (k=1 to k=10):
-----------------------------
k=1: Accuracy = 0.9850
k=2: Accuracy = 0.9600
k=3: Accuracy = 0.9700
k=4: Accuracy = 0.9650
k=5: Accuracy = 0.9700
k=6: Accuracy = 0.9800
k=7: Accuracy = 0.9800
k=8: Accuracy = 0.9700
k=9: Accuracy = 0.9750
k=10: Accuracy = 0.9750

Optimal k: 1 (Accuracy: 0.9850)
```

kNN Performance Analysis

- **Optimal k Selection**:
  - Tested k=1 to k=10; k=1 achieved the highest accuracy (98.5%).
  - Rationale: Smaller k values (like k=1) work well when the data has clear clusters and low noise, as the nearest neighbor dominates the prediction.
- **Trade-offs**:
  - **Pros**: High accuracy (98.5%) suggests the dataset has well-separated classes.
  - **Cons**: k=1 is sensitive to outliers/noise and has a risk of overfitting.

Experimental Results:

Classifier Performance Summary:

Metric	**KNN(K=1)**	**SVM(Linear)**
Accuracy	98.5	98
Confusion Matrix	98 1 2 99	98 1 3 98
False Positives	1	1
False Negatives	2	3

Comparative Analysis:

KNN Performance(K=1):

Strengths:

- By exploiting local data patterns, we achieved the maximum accuracy (98.5%).
- Fewest total mistakes (3 vs. SVM's 4), indicating strong class separation.
- Cosine similarity was effective at capturing document relationships in high-dimensional space.

Weaknesses:

- The reliance on a single nearest neighbor increases the risk of overfitting.
- In real-world deployment, loud or confusing documents may provide a challenge.

SVM Performance:

Strengths:

- Near-perfect accuracy (98.0%), with improved theoretical generalization.
- Margin maximization protects outliers.
- High-dimensional text features were handled efficiently by the linear kernel.

Weaknesses:

- False negatives are slightly greater (3 vs2 in kNN), indicating stronger categorization boundaries.

Conclusion:

Both classifiers performed exceptionally on the dataset, with:

- kNN (k=1) achieved a little greater accuracy (98.5%), showing that the classes are well separated in the feature space. Its success indicates that document similarity (by cosine distance) is a valid metric for this purpose.
- SVM followed closely (98.0%), demonstrating that a linear decision boundary effectively distinguishes between Hockey and Windows documents.

The smallest difference in accuracy (0.5%) indicates that any classifier is appropriate for this dataset. The decision between them could depend on:

- kNN (k=1): Preferred if raw accuracy is important and the data is known to be clean.
- SVM: Preferred if robustness to possible noise is important.

This experiment validates that both methods, when correctly designed, are extremely effective for binary text classification tasks with well-preprocessed