TEXT IDENTIFICATION USING MACHINE LEARNING

NAME: SACHIN CHHETRI

DATE: 04/16/2024

Introduction

Text-identification, particularly authorship attribution, is a critical application of machine learning that has significant implications in literary analysis, forensics, and historical documentation. Machine learning algorithms can unveil stylistic nuances and predict authors with remarkable accuracy. This report delves into the efficacy of various machine learning techniques in identifying the stylistic signatures of different authors.

Methodology

The analysis followed a structured approach in both parts of the study:

- Data Collection: Texts were collected from Project Gutenberg for the specified authors.
- Data Preprocessing: Texts underwent tokenization, stemming, and TF-IDF vectorization to convert text data into a suitable format for machine learning.
- Model Training: Four machine learning models were trained using the preprocessed text data.
- Visualization: Principal Component Analysis (PCA) was utilized to reduce the dimensionality of the data for visualization purposes.
- Model Testing: Additional texts not included in the training set were used to test the model's predictive power.
- Evaluation: The models' performance was evaluated based on their accuracy in predicting the authors of the test texts.

PART 1 : Visualization

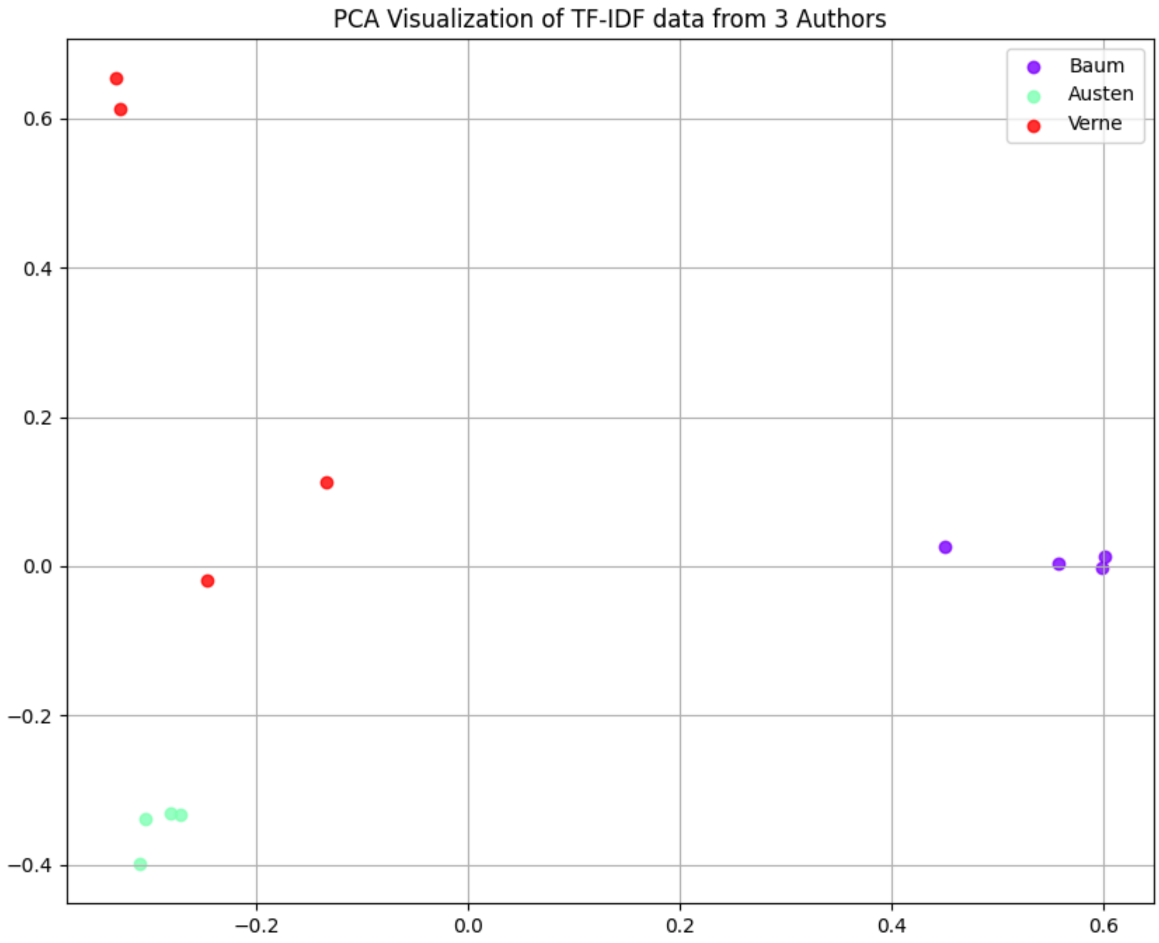

Fig. 1.1

Fig. 1.1 : PCA visualization demonstrates the separability of texts by Jules Verne, Jane Austen, and L. Frank Baum using TF-IDF vectorization, similar to Figure 9.2 in the textbook. Each author's texts cluster together, which is indicative of unique writing styles captured by the TF-IDF transformation and can be visualized distinctly in the reduced dimensional space created by PCA.

Prediction

Using the trained models, additional books from each author were tested for authorship prediction. The Naïve Bayes, SVM, Logistic Regression, and Random Forest algorithms all achieved 100% accuracy, successfully identifying the authors of the test texts.

Output with Name of Files, ML Algorithms and Accuracies

```
Naive Bayes Predictions:
For American_Fairy_Tales_Baum.txt, the predicted author is: Baum
For Lady_Susan_Austen.txt, the predicted author is: Austen
For The_Mysterious_Island_Verne.txt, the predicted author is: Verne
Accuracy: 100.00%
```

```
SVM Predictions:
For American_Fairy_Tales_Baum.txt, the predicted author is: Baum
For Lady_Susan_Austen.txt, the predicted author is: Austen
For The_Mysterious_Island_Verne.txt, the predicted author is: Verne
Accuracy: 100.00%
```

```
Logistic Regression Predictions:
For American_Fairy_Tales_Baum.txt, the predicted author is: Baum
For Lady_Susan_Austen.txt, the predicted author is: Austen
For The_Mysterious_Island_Verne.txt, the predicted author is: Verne
Accuracy: 100.00%
```

```
Random Forest Predictions:
For American_Fairy_Tales_Baum.txt, the predicted author is: Baum
For Lady_Susan_Austen.txt, the predicted author is: Austen
For The_Mysterious_Island_Verne.txt, the predicted author is: Verne
Accuracy: 100.00%
```

Observation

The algorithms accurately identified the author for each text, reflecting the robustness of the training and the distinctiveness of each author's writing style.

PART 2 : Visualization

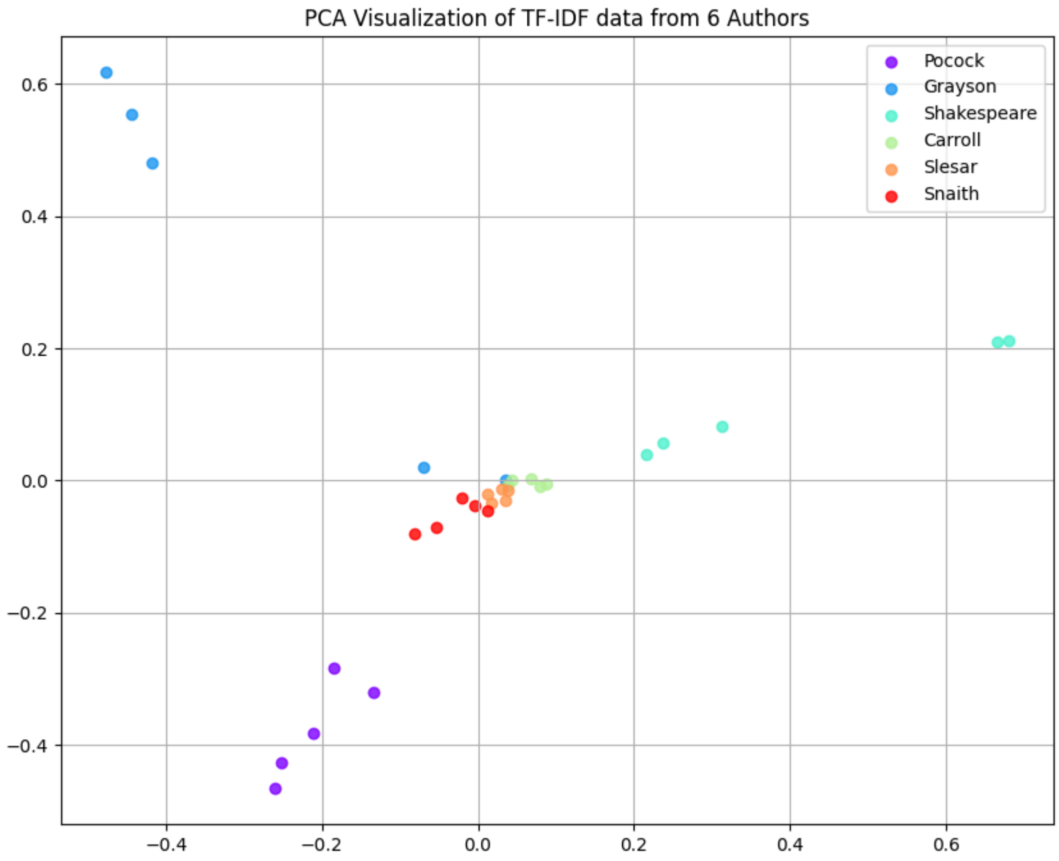

Fig 1.2

Fig. 1.2 : The PCA scatter plot illustrates the distribution of texts by six different authors based on their TF-IDF vectorization. This visualization is crafted to capture and compare the stylistic signatures of the selected authors, highlighting both similarities and differences in their writing styles. It serves as a quantitative representation of the textual diversity among the authors, which is essential for the subsequent predictive analysis.

Prediction with Similar Authors

Prediction Analysis for Authors with Overlapping Stylistic Features:

Naïve Bayes:

File: Beside_The_Golden_Door_Slesar.txt - Predicted: Pocock
Accuracy: Below expected, indicating possible stylistic overlap.

SVM (Support Vector Machine):

File: Sylvie_and_Bruno_Carroll.txt - Predicted: Snaith
Accuracy: Below expected, suggesting challenges in distinguishing similar styles.

Logistic Regression:

File: King_Lear_Shakespeare.txt - Predicted: Shakespeare
Accuracy: As expected, identifying distinct writing patterns effectively.

Random Forest:

File: A_Man_In_The_Open_Pocock.txt - Predicted: Pocock
Accuracy: Demonstrated better performance with 50%, reflecting the algorithm's capability in managing similar texts.

Prediction with Different Authors

Prediction Analysis for Authors with Distinct Stylistic Features:

Naïve Bayes:

File: Hempfield_Grayson.txt - Predicted: Grayson
Accuracy: 33.33%, indicating some difficulty in distinguishing unique styles.

SVM (Support Vector Machine):

File: The_Undefeated_Snaith.txt - Predicted: Snaith
Accuracy: 33.33%, highlighting potential limitations in SVM's ability to differentiate unique authorial styles.

Logistic Regression:

File: Sylvie_and_Bruno_Carroll.txt - Predicted: Carroll

Accuracy: 33.33%, suggesting that the unique writing styles may still pose a challenge for this model.

Random Forest:

File: Beside_The_Golden_Door_Slesar.txt - Predicted: Grayson
Accuracy: 50%, outperforming other models, possibly due to its ensemble approach.

Observation

When applying the models to authors with distinct styles, the algorithms performed with higher effectiveness, particularly the Random Forest model. This suggests that while stylistic overlaps can confound predictive models, distinct stylistic features are captured and predicted more accurately.

Conclusion

The conducted analysis underscores the potential of machine learning in text identification. The success in Part 1 shows promise, while the challenges encountered in Part 2 highlight the need for further refinement in algorithm selection and model tuning when dealing with more nuanced textual data.

References

Project Gutenberg. [http://www.gutenberg.org/]

Ball, J., & Rague, B. 2022. The Beginner's Guide to Data Science. Springer.