Justification of Decisions

Tokenization (Implemented):

- ***Decision***: The custom `tokenize` function uses regular expressions to break text into tokens, including words, numbers, and punctuation
- ***Justification***: This approach is chosen for its simplicity and effectiveness in handling a wide range of textual data. It ensures that important features like contractions and punctuation marks, which can carry significant meaning, are not discarded
- ***Outcome***: Training and Test Accuracy improved from 0.42 and 0.41, to 0.87 and 0.89

Smoothing (Implemented):

- ***Decision***: Laplace smoothing is applied in the `predict` method.
- ***Justification***: This technique addresses the issue of zero probability for unseen words in the training data. It adds a small, constant value (1 in this case) to the count of each word, preventing the probability of any event from being zero

Feature Selection (Implemented):

- ***Decision***: All words in the text are used as features without removing stop words.
- ***Justification***: The decision to keep all words, including stop words, since they can provide meaningful context that aids in classification

Model Parameters (Implemented):

- ***Decision***: The model uses class priors and word counts as its primary parameters.
- ***Justification***: Class priors provide a baseline probability of each class, and word counts are essential for calculating the likelihood of a text belonging to a particular class- These parameters are necessary for a Naive Bayes classifier. In addition, logarithms of class priors are used instead of raw probabilities to prevent underflow problems that can occur when multiplying many small probabilities together

Handling Head and Tail Entities (Did not implement):

- ***Decision***: The current implementation does not explicitly handle head and tail entities.
- ***Justification***: The effectiveness of the Naive Bayes classifier is already high without needing to explicitly using head or tail entities. This can mainly be attributed to its ability to learn and generalize from statistical patterns in the text. It capitalizes on the distribution and frequency of words, and looks at contextual clues around potential entities to make predictions. Hence, it enables the classifier to accurately predict classes even without specific entity identification. This also suggests that the overall linguistic patterns in the training dataset is sufficient for each class, i.e. the training set is representative enough. Therefore explicit entity recognition is less critical for this classification accuracy.

Model Accuracy

	Training (Cross-validated)	Test
Model Accuracy	**0.87**	**0.89**

Confusion Matrix

	Characters	Director	Performer	Publisher
Characters	**83**	**9**	**3**	**8**
Director	**4**	**85**	**2**	**3**
Performer	**7**	**3**	**91**	**2**
Publisher	**2**	**0**	**1**	**97**

Publisher

	True Publisher	True Not
System Publisher	**97**	**13**
System Not	**3**	**287**

Publisher recall: ***0.97***
Publisher precision: ***0.8818181818181818***

Director

	True Director	True Not
System Director	**85**	**12**
System Not	**9**	**294**

Director recall: ***0.9042553191489362***
Director precision: ***0.8762886597938144***

Performer

	True Performer	True Not
System Performer	**91**	**6**
System Not	**12**	**291**

Performer recall: ***0.883495145631068***
Performer precision: ***0.9381443298969072***

Characters:

	True Characters	True Not
System Characters	**83**	**13**
System Not	**20**	**284**

Characters recall: ***0.8058252427184466***
Characters precision: ***0.8645833333333334***

Pooled:

	True Yes	True No
System Yes	**356**	**44**
System No	**44**	**1156**

Micro Average Precision: ***0.89***
Micro Average Recall: ***0.89***

Macro Average Precision: ***0.8902086262105592***
Macro Average Recall: ***0.8908939268746128***

Error Analysis:

The errors in our output stem from how the same words can be used in multiple different contexts

> *Let's take row 1334 as an example, “First Blood is a 1972 novel by David Morrell which was adapted into the 1982 film starring Sylvester Stallone as John Rambo.” Here it is supposed to be tagged as performer, but is predicted to be director. The attribute of film can confuse the classifier as it can be used in the context of director too*
>
> *In row 1923 “Covers ranged from a tongue - in - cheek excerpt of Avril Lavigne 's "" Sk8er Boi "" to the Louis Armstrong classic "" What a Wonderful World "" ”. Here it is predicted to be a character instead of performer. This happens due to the similar words both character and performer share, as a performer and character can be part of a movie or play,leading to miss-classifications.*
>
> *Similarly in row 241 "" Crime on the Waterfront "" and the resulting 1953 Waterfront Crime Commission provided Elia Kazan with the factual background for his 1954 film "" On The Waterfront . ""“”. Here the tag was supposed to be director, but was predicted as character. This can be because of the word film, as a director directs a film, a film contains characters within itself.*

The confusion matrix supports these observations, showing overlaps between 'characters' and 'director', 'performer', and 'publisher', mainly due to shared domain-specific language and the challenge of role identification. Named entities, like names of individuals or specific titles, also seem to disproportionately influence the classification, leading to a higher rate of misclassification. These issues point to the need for enhanced feature engineering, such as incorporating contextual information or named entity recognition, to better differentiate between classes in complex or ambiguous contexts.

Advanced text preprocessing, like lemmatization, could also help in focusing on the core meaning of words, thereby reducing lexical variability and improving classification accuracy. Overall, these enhancements could mitigate the inaccuracies arising from the shared use of words across different labels."