Shreayan Chaudhary
February 9, 2023
Homework #1b Information Retrieval 601.666

Models I experimented with:

1. Logistic Regression
2. Gaussian Naive Bayes
3. Support Vector Classifier
4. Decision Trees
5. Random Forest
6. K Nearest Neighbours
7. XGBoost
8. Multi Layer Perceptron (MLP)

Algorithm:

1. First I created a dictionary of the Parts Of Speech from the given data
2. Then I created a dictionary mapping each word to its POS
3. Created feature for POS counts in the line
4. Created other features described below
5. Split the data into train and dev splits
6. Train the models on train split and evaluate on test split
7. Quantify the performance and see where model is working well and not working well
8. Try to come up with new features for places where model is not working well
9. Repeat steps 6-8

Features that worked wel:

1. Length of the text
2. Length of the text after trimming whitespaces
3. Number of words in the text
4. If the text starts with a quotation (> : @) (to detect quotes)
5. If the text has the words “wrote”, “writes”, “said”, “told”, “tells” (to detect quotes)
6. Number of spaces in the text (to detect blank lines)
7. Count of (| _ - / \ +) in the text (to detect figures)
8. Ratio of uppercase characters to all the characters
9. Ratio of numeric characters to all the characters
10. If any "headline" keyword is found (From, Subject, Date, Organization, References)
11. Number of non-word chars (special chars) in the text
12. If regex for email found in the text
13. If regex for phone number found in the text
14. If length of words in current line matches previous line (to detect tables)
15. One feature for each POS containing the count of words with that POS in text

Some features that did not work well:

1. Total number of characters in the previous line.
2. Total number of characters in the next line.
3. If the line starts with a numeric character.

Performance and Metrics:

The following table shows the accuracy when the model is trained on the training set and evaluated on the development set: All accuracies are evaluated on the dev split.

Model	Accuracy on lines	Accuracy on segments
Logistic Regression	85.73	75.64
Gaussian Naive Bayes	55.18	61.53
Support Vector Classifier	73.87	50
Decision Trees	92.24	83.33
Random Forest	**95.15**	**90.46**
Multi Layer Perceptron	94.9	86.51
XGBoost	94.73	84.61
K Nearest Neighbours	80.60	62.82

Result:
We can see that the Random Forest is working the best for the given problem. It gives an accuracy of 95.15% for lines and 90.46% for segments on the development data, which means it is not overfitting and is able to generalize well on new data.

Some areas that could be improved:
It is not able to properly predict the items. I have added a feature for checking bullet points (1., 2., etc. but sentences stretching multiple lines are not able to be learned by the model. It is giving a bad performance and instead classifying them as a PTEXT.