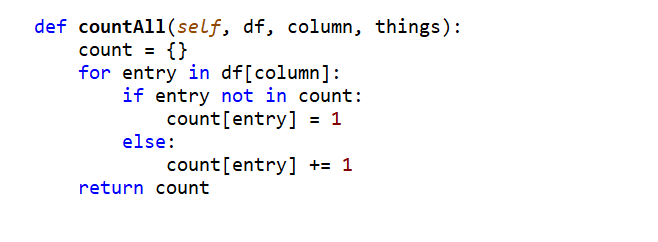
Data Prepossessing

In order to further the accuracy of our models’ predictions we looked at the dataset in which it was trained. The dataset used to train the previous models consisted of 34886 movie records with 186 unique genres. To explore this further a function was created to count the number of movie records for each genre, this function is shown below:



A sample of the top 16 genres in descending order of number of movie entries are shown below:

|  |  |
| --- | --- |
| Movie Genre | Movie count |
| drama | 9397 |
| comedy | 6468 |
| unknown | 6083 |
| action | 1355 |
| horror | 1209 |
| thriller | 1197 |
|  | 1148 |
| romance | 1006 |
| western | 874 |
| crime | 583 |
| adventure | 564 |
| musical | 491 |
| film | 465 |
| science | 445 |
| war | 376 |
| mystery | 315 |

It was evident from these results that the dataset was highly unbalanced, 70% of the total movie records were in the top five genres. This was causing our model to overfit and consistently predict the top 3 genres, drama, comedy and unknown. On top of the model’s imbalances it also had a large number of miscellaneous and empty movie genres. These were acting as noise in our dataset and lowering its prediction accuracy.

To correct these issues, we decided to remove these miscellaneous genres and balance the number of movies for each genre. In order to order to balance the dataset we found 2 additional datasets with the same format as the original and all 3 datasets were combined.

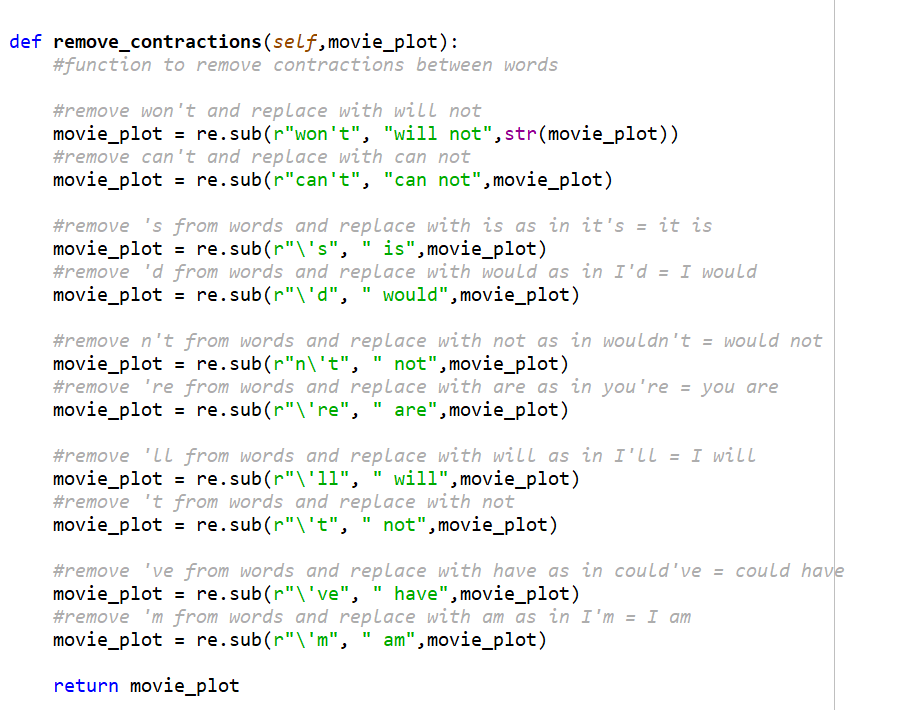
The 2 additional datasets we chose can be found here:

<https://github.com/ishmeetkohli/imdbGenreClassification/tree/master/data>

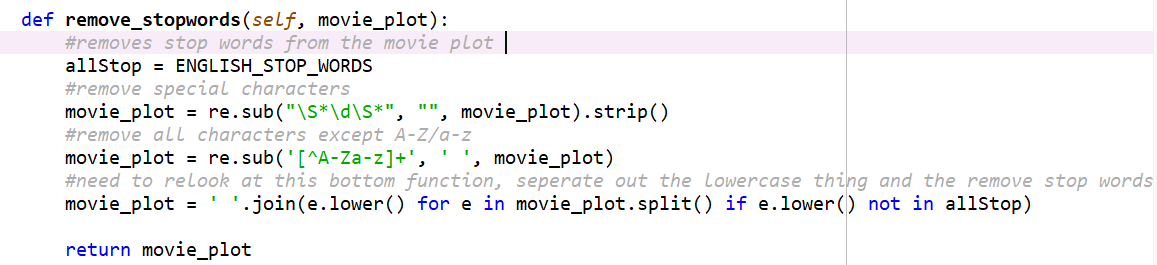
<https://github.com/shashankvmaiya/Movie-Genre-Multi-Label-Text-Classification/tree/master/Data>

The collective dataset was then filtered, any genres with less than 500 movies were removed and all other genres had their movie count limited to 1000. After the dataset was balanced, we set about standardising the movie plots by removing any special characters or numbers, we also removed any contractions between words and ensured all words were in lower class. The functions we wrote to standardise the movie plots is shown below:

Function to remove contractions:



Function to remove stop-words, special characters, numbers and to turn all words lowercase



After training the model on this new dataset the accuracy improved by 15% and was no longer overfitting on any 1 genre.