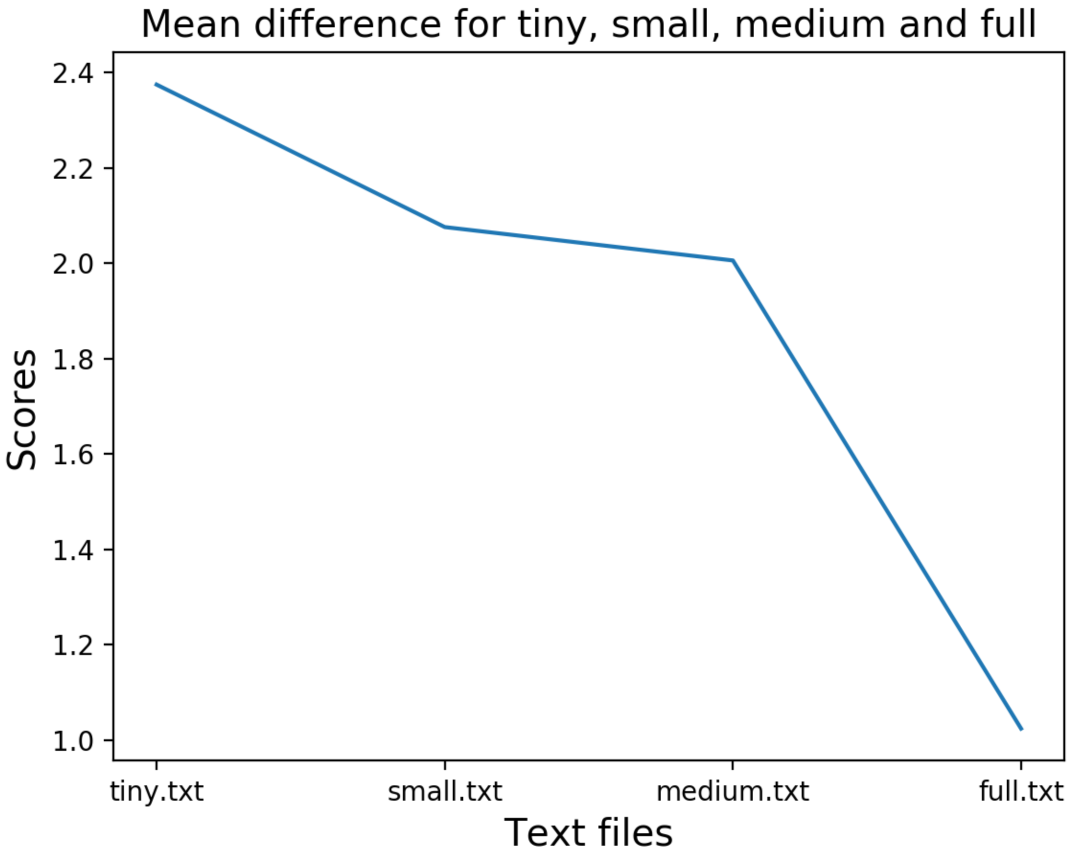
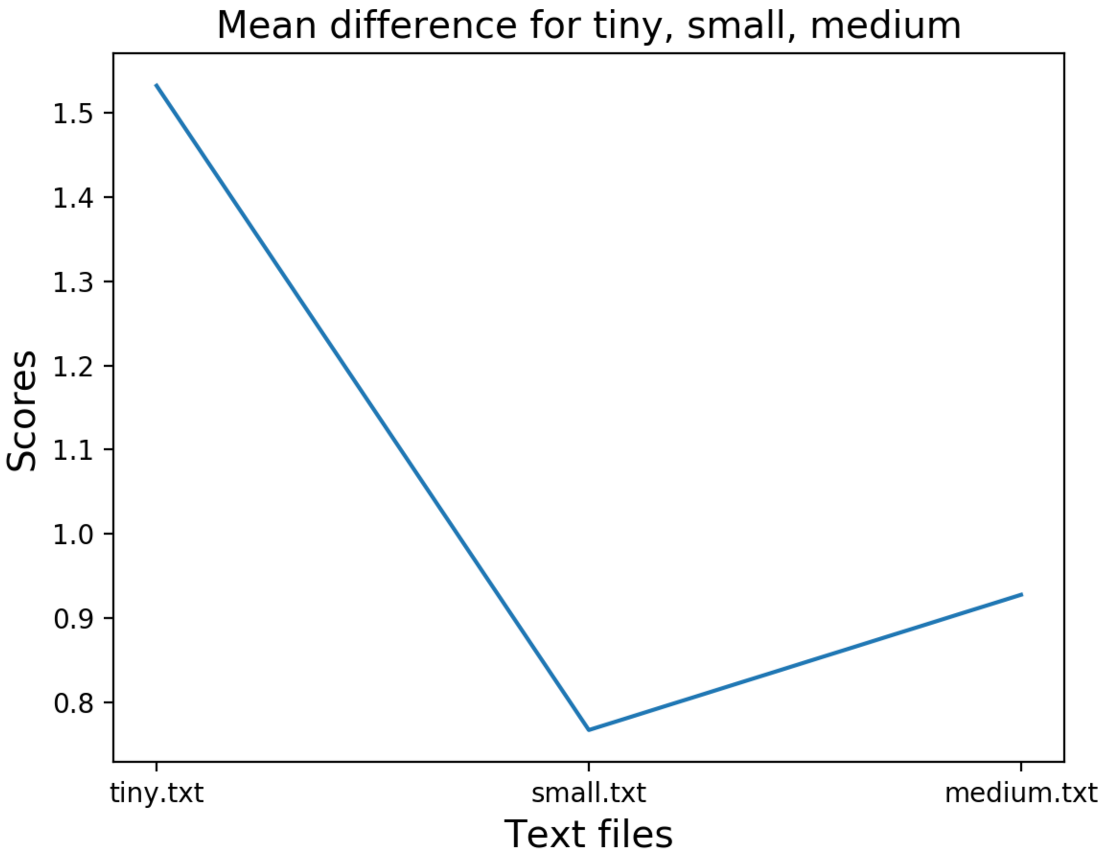
Part IV: Exploration

Question: How close are our Predicted Sentiment Scores (PSS) to the scores the reviews originally came with?

We created a function called mean_diff_pss that calculates the mean difference between the Predicted Sentiment Scores and the original scores in order to see how much the scores varied. We used the statement_pss function to get the PSS. We then subtract the PSS from the original score for each line and then divide by the total number of scores (number of lines). For tiny.txt the mean difference is 2.375, for small.txt the mean difference is 2.076, for medium.txt the mean difference is 2.006, and for full.txt the mean difference is 1.024. We found that there is a negative linear relationship between the mean difference and the size of the .txt file. The visualization shows that as the sample size increases, the mean difference decreases, indicating that the variance between the scores decreases as the file gets larger. In conclusion, the accuracy of the model is greater when working with a bigger sample.



Since we used the same file to train and test in the function discussed above, we decided to experiment by using full.txt as the training file and the other three as tests in another function called mean_diff_pss2. For tiny.txt the mean difference is 1.53, for small.txt the mean is 0.767, and for medium.txt the mean is 0.928. The visualization for this function shows that the relationship is not linear. We are making an assumption that the mean difference for medium.txt is larger than small.txt because the expected ratio of training and testing is 0.2 to 0.8. In this case, the size of medium.txt is greater than an expected testing size.


Mean difference for tiny, small, medium
Scores
tiny.txt
small.txt
medium.txt
Text files