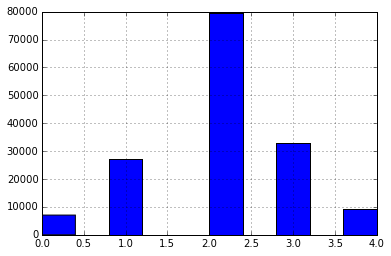
I will be predicting sentiment labels for Rotten Tomatoes reviews/phrases. I’ve looked at the distribution of the sentiment labels (0-4) and found that the average label is 2.063578
(neutral).



The sentiment labels are:

0 - negative  
1 - somewhat negative  
2 - neutral  
3 - somewhat positive  
4 - positive

The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. Each Sentence has been parsed into many phrases by the Stanford parser. Each phrase has a PhraseId. Each sentence has a SentenceId. Phrases that are repeated (such as short/common words) are only included once in the data.

Train.tsv file:

* contains the phrases and their associated sentiment labels.
* SentenceId to track which phrases belong to a single sentence.
* Counts = 156,060 PhraseID and 8544 SentenceIDs

Test.tsv file:

* contains just phrases
* Need to predict sentiment labels

Since I am predicting sentiment analysis, I plan to use some algorithm or machine learning method that we will touch on later in class.