Amanda L. Miller

Assignment #1: Language Modeling – classifying tweets

I developed a pair of unigram models, and a pair of bigram models that were trained on the hand-annotated tweets used by Waseem et al (2016), and used them to classify tweets as either illustrating ‘sexism’ or ‘nonsexism.’ I calculated the accuracy of the models against the annotator judgments. I implemented an ‘unk’ feature for the unigram model, and for both levels of the bigram model. The ‘sexist’ model was trained on the sexist tweets, and the non-sexist model was trained on non-sexist tweets. I used the data from the other set of tweets (e.g. the sexist tweets when developing the non-sexist model), to put characters that hadn’t been added to the model based on the original data set (e.g. the non-sexist tweets for the non-sexist model), into the ‘unk’ levels of the unigram dictionary, and the second level of the bigram nested dictionary. I still had to set characters that were in the test data, but had not occurred in the training data to ‘unk’ when running the classifications. I seem to have gotten similar results to those that Micha Elsner reported in his implementation of these models (see table 3). I trained using alpha=1 for the unigram model, and beta=1 for the bigram model, and then attempted to maximize the likelihood of the held out data by choosing the optimal alpha value, and the optimal beta value for the unigram and bigram models respectively. Tables 2 and 3 provide the list of values tested, and the log probability of the held out data and test data for the different models.

|  |  |  |
| --- | --- | --- |
| Alpha Value | Log prob of Held out data | Log prob of Test Data (sexist) |
| 1 |  | -341845.744584 |
| 0.9 | -167127.841317 | -341842.808044 |
| 0.5 | -167121.740813 | -341832.111477 |
| 0.2 | -167117.831611 | -341825.478666 |
| 0.1 | -167116.701739 | -341823.612763 |
| 0.05 | -167116.176834 | -341822.757099 |
| 0.01 | -167115.778175 | -341822.112999 |
| 0.005 | -167115.729745 | -341822.035131 |

Table 1. Manual Testing of alpha additive smoothing values for to attempt to maximize the held out corpus (tuning data) and testing data for the unigram language model

|  |  |  |
| --- | --- | --- |
| Beta Value | Log prob of Held out (training data) | Log prob of Test Data (sexist) |
| 1 | -365008.736463 | -365008.736463 |
| 0.9 | -175712.589051 | -359729.452999 |
| 0.5 | -163223.74321 | -334158.712256 |
| 0.2 | -149829.379006 | -306704.202712 |
| 0.1 | -143534.933135 | -293786.351199 |
| 0.05 | -139539.012221 | -285580.701585 |
| 0.01 | -135332.595002 | -276946.077813 |
| 0.005 | -134646.602043 | -275536.825052 |
| 0.001 | -134017.537516 | -274230.609021 |

Table 2. Manual Testing of Beta Smoothing Values for maximizing likelihood of held out data (tuning data), and testing data for the fitted bigram language model

Table 3 provides the accuracy of the two Ngram models that were implemented:

|  |  |  |  |
| --- | --- | --- | --- |
| **Language Model** | **Precision** | **Recall** | **f-score** |
| Fitted Unigram Model | 0.56 | 0.88 | 0.68 |
| Fitted Bigram Model | 0.48 | 1.67 | 0.748 |

Table 3. Accuracy Measures for three distinct Ngram Models

Discussion of several Tweets that were incorrectly classified:

1) The tweet ‘if you mean they have giant dicks that they're swinging around, yes.’ Which occurs in the testdata, was judged by the annotator as illustrating sexism, but my model predicted it **not** to be sexism.

2) The tweet ‘man fuck these feminazi s Kube. We don't need a black history month. Blacks don't care why solo them out’ was annotated as illustrating racism, but my fitted bigram model classified it s exhibiting sexism.

3) The tweet ‘#mainecoon #coon #cat #kitty #pet #animal #catstagram #catlover #ajka #animals #pets #pets… http://t.co/iFKVGbusza http://t.co/74UM0QKl9C’was annotated as illustrating neither sexism or racism, but my fitted bigram model classified it as illustrating ‘sexism’

The term ‘dick’ was used in isolation or as part of a morphologically complex form containing it (such as ‘dickiest’) 12 times in the training data. Only 3/ 12 cases were marked as illustrating sexism by the annotators. All three cases that were marked as illustrating ‘sexism’ involved the use of the term ‘dick’ on its own (not in a morphologically complex form). Although space characters were included as characters when building the model, this may not have been sufficient to allow the model to learn that ‘dick’ illustrates sexism, but ‘dickiest’ and ‘dickwarts’ do not.

The second example that the model got wrong is a problem with the simplified annotation system that was used in the simplified data set. We know that there was originally a ‘both’ label for tweets that were both sexist and racist,, but that the ties were broken in favor of racism. Thus, this classification as ‘sexism’ is probably correct, if we were to look back at the original annotations. I’m quite certain that at least one of the annotators would have labeled this as sexism, since it contains the word ‘feminazi.’

The third example is the most tricky. I’m not sure exactly why this one was classified as sexism. There is only one example of the word ‘kitty’ in the training data, which was used in a sexist tweet, so that alone could have caused the issue. The word ‘cat’ seems to be used more often in a non-sexist way in the training data, and the word ‘lover’ seems to be used only once in the training data in a sexist tweet. So, I think more data would have helped the model learn the best.