Linyi Chen
11791 HW3 Report

1. Table *

Features	Classifier	Accuracy	Precision	Recall	F-measure
Count base	MNB	0.06276	0.04850	0.06276	0.02791
Count base	SVM	0.07327	0.07740	0.07327	0.06182
Count base	MLP	0.07678	0.06384	0.07678	0.06040
Tf-idf based	MNB	0.06754	0.04531	0.06754	0.02786
Tf-idf based	SVM	0.13030	0.06905	0.13030	0.08310
Tf-idf based	MLP	0.10513	0.08522	0.10513	0.08583

* I only trained model with 5000 training examples, since processing the entire 1.2G training file far exceeded the CPU load of my laptop(The fan actually went crazily loud even when I trained the model with 5000 examples).

2. Broad insight

Feature-wise:

I noticed that for all model selections, tf-idf based features produced better predictions on accuracy and recall than count based features. The improvement for SVM model on accuracy is the greatest - about 77.8%. However, for precision, only MLP's precision went up with tf-idf based features. Similarly, not all model performs better on F-measure with tf-idf based features - MNB went down by about 0.18% with tf-idf based features. At the same time, there are noticeable improvement on F-measures of SVM and MLP model. To sum up, I think Tf-idf based features outperform count based features, since the four measures mostly went up when implemented with tf-idf based features.

Model-wise:

I observed that in terms of accuracy, precision, recall and F-measure, regardless of use of features, MLP and SVM models both outperform MNB significantly. When comparing the performance of SVM and MLP models on count based features, SVM is better than MLP with tf-idf features on regarding precision and F-measure, but is worse than MLP with accuracy and recall. In addition, When comparing the performance of SVM and MLP models on count based features, SVM is better than MLP with tf-idf features regarding accuracy and recall, but is worse than MLP with precision and F-measure. These differences between SVM and MLP are not as significant as the difference between them and MNB. I think overall MLP and SVM models perform equally well, and are both significantly better than MNB.

Combining all of the above, SVM and MLP with tf-idf features are the best combinations among all. Models trained with tf-ids features are generally better than those trained with count based features.

3. Relative Performance Analysis

There are in total 3139 dev examples, I counted those that have “all ones”(all models predict correctly, easy examples), “mix of zero and ones” (some are correct and some are not) and “all zeros” (all models get it wrong, so hard for all), which are 10, 728 and 2401, respectively.

I found that MNB models made the most mistakes, which correspond to their relatively lower accuracy discussed in the previous section. I then also took a closer look at the examples from each category by printing out the answer type, answer and query. I found that there are not noticeable feature or pattern within each category. For instance, below are the ten examples all models predict correctly:

```
reference,answer type,answer,query
111111,,server,weblogic11g -- the 11.x july 2009 release of weblogic server a java-ee application @placeholder suite from oracle .
111111,,web,frontpage -- frontpage is a design tool for @placeholder publishing which was part of microsoft office .
111111,,server,weblogic11g -- the 11.x july 2009 release of weblogic @placeholder a java-ee application server suite from oracle .
111111,,database,ddl -- data definition language is a subset of sql to manipulate structural elements of a @placeholder not the content of tables .
111111,,ios,uinavigationbar -- the uinavigationbar class implements a control for navigating hierarchical content in @placeholder .
111111,,java,jsr250 -- jsr 250 : common annotations for the @placeholder platform .
111111,,web,evaporate.js -- evaporatejs is a javascript library for directly uploading files from a @placeholder browser to aws s3 using s3 s multipart upload .
111111,,s,dynamics-ax-2009 -- microsoft-dynamics ax is one of microsoft @placeholder enterprise resource planning software products .
111111,,api,tweetstream -- tweetstream provides access to twitter s streaming @placeholder in python or ruby .
111111,,java,gottox -- gottox refers to gottox socket.io-java-client which is a socket.io client implementation in @placeholder
```

There is no answer type for all examples, and the position of the answer in the query, as well as the answer itself are all different. It is also true for examples that all models predict incorrectly and those that only some of the models get correct answer. After taking a closer look into the code, I found that the query itself is not included into the training, maybe adding the query into the features may help with the performance. In terms of feature performance, I removed stopwords in extracting tf-idf features, and converted all words into lowercase. I think this might be one of the reasons tf-idf based features are better. In terms of model performance, I think MNB perform the worst since it is the least complex and there is no parameter to tune.