IT-Discovery at TREC 2010 Legal

Aron Culotta,
Andy Liu,
Mark Cordover
IT-Discovery
Washington, D.C.
{aron,andy,mark}@it.com

Bennett Borden
Williams Mullen
Washington, D.C.
bborden@williamsmullen.com

Sam Strickland
Strickland Review
Washington, D.C.
sstrickland@stricklandereview.com

ABSTRACT

IT-Discovery participated in both the Learning Task (Topics 201-207) and the Interactive Task (Topics 301, 303). For the Learning Task, we used an optimized Naive Bayes classifier, which obtained 90-97% cross-validation accuracy on the provided seed sets for each topic. For the Interactive Task, we used the same classifier trained with one round of active learning. The annotator averaged 36.5 hours per topic, resulting in a cross-validation classification accuracy of 90%.

1. LEARNING TASK

In the Learning Task, we are provided with *seed sets* of documents for each topic. We train a standard document classifier to assign labels to all remaining documents.

1.1 Classifier

We implement a multinomial Naive Bayes classifier [1] due to its ease of implementation and superior scalability. To overcome a number of known problems with Naive Bayes, we implement many of the suggestions described in Rennie et al. [2], including transformations based on term frequency, document frequency, and document length. With these transformations, Naive Bayes has been found to be competitive with more computationally intensive classifiers, such as support vector machines.

We compute several types of features for the classifier:

- Unigrams, bigrams, trigrams from message body
- Unigrams from the subject
- Time of day
- Correspondent names
- Correspondent type (one-to-one, has-cc, has bcc)
- Document length (five bins)
- Whether the email has an attachment or url
- Whether the email contains words repeated an unusually large number of times

We implemented two additional optimizations:

- **Feature Selection:** We sort all features by their Information Gain, then include the number of features that optimizes cross-validation accuracy. We optimize this number separately for each topic.
- **Class Imbalance:** To mitigate class imbalance problems, we up-sample the minority class until there are an equal number of positive and negative instances.

1.2 Active Learning

Although our final submission to the Learning Track did not use active learning, we did perform initial experiments that were surprisingly ineffective. After a few rounds of additional document labeling, the classifier's accuracy did not appear to be improved on the seed set. There are several possible causes of this, which we will discuss in our final submission.

2. INTERACTIVE TASK

For the Interactive Task, we use the same classifier as the Learning Task, trained using one round of active learning. First, the annotators labeled 28,787 documents using IT-Discovery's search tool. After training the classifier on these documents, we assigned a relevance probability to each document. We then sampled an additional 987 documents according to a stratified sample of relevance probabilities: most relevant (0.85-1.0), relevant (0.7-0.85), less relevant (0.6-0.7), and hardly relevant (0.5-0.6). We then re-trained the classifier on all annotated documents and re-classified all documents to produce the final results.

References

[1] R. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley and Sons, 1973.

[2] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.