Web spam filtering using topic correlation analysis

ABSTRACT

This project proposes a corpus based machine learning algorithm to classify web spam. We begin by applying topic modeling on the corpus. This is followed by determining correlation of identified topics in each document. This correlation is the crux of the system we propose. It forms the basis for classification of the documents that is performed using SVC, RF, kNN and LR. The modular structure of the process also allows performance of equivalent techniques to be compared. We use the UKWebSpam2006 dataset to evaluate the system and the results are (expected to be) promising.

INTRODUCTION:

Topic modeling has been an area of flourishing interest in information retrieval. Latent Dirichlet Allocation[1] has been found to be most effective in topic modeling on a corpus. Biro et al.[2] have used a unique method of using LDA to classify spam. Their method identifies topics as spam or not spam and further classifies documents based on the proportions of spam and non-spam topics.

However, we conjecture that a document may not necessarily be classified as spam simply by virtue of it containing a topic. Instead, co-occurence of topics in a document may be a far better indicator in this regard. For example, a document containing references to “sale” and “discounts” may not always be spam as the page may contain promotional offers by a company. Similarly, a document speaking about “discounts” and “travel” is far less likely to be spam than one about “discounts” and “biology”. Based on the findings of Gyongyi et al.[3], it is clear that spam techniques like keyword dumping, weaving and phrase stitching constitute almost all of content-based spamming techniques. Pages generated using these techniques can be reasonably expected to contain references to barely related topics.

The procedure we propose to use is described in detail in later sections. To describe briefly, we begin by applying topic modelling to identify topic proportions of each document. Then we proceed to finding topic correlations for each document. This topic correlation representation will form the feature space for the classification task that is finally performed using several different classifiers. For the training process we use the UKWebspam2006 dataset.

METHODOLOGY

The datasets require a series of transformations before they are finally passed to the classifier. We use the UKWebSpam2006 dataset for the experiments.

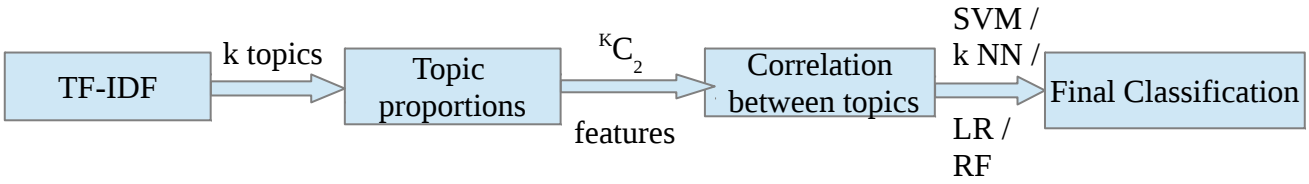
First, the tf-idf values for the body are read into a sparse matrix. This forms the corpus for training. LDA is run upon this corpus to extract topic proportions for each document. The number of topics derived is parametrised as k.

Upon obtaining topic prportions, the next step is to correlate topic proportions. For each document, we multiply every topic proportion with every other to obtain ${}^{k}C_{2}$ values. These values serve as final

features for classification.

We performed classification using the following classifiers to compare performances : k-nearest neighbor, support vector classifier, logistic regression, random forests.

The approach is summarised below



RESULTS

We modeled the corpus on 5,10,15...100 topics to compare performance and passed these features to aforementioned classifiers with different parameters. The results are in the form (precision, recall, f1-score). Refer to attached document for detailed results for each classifier.

CONCLUSION

The most promising results are yielded by high number of topics and the results are more sensitive to number of topics than to the final classifier used. As observed, the classifier often suffers from low recall. Further experiments can be aimed at addressing this issue.

REFERENCES:

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3(5):993–1022, 2003.

[2] I. Biro, J. Szabo, A. Benczur. Latent Dirichlet allocation in web spam filtering. Proceedings of the 4th international workshop on Adversarial information retrieval on the web, April 22-22, 2008, Beijing, China [doi>10.1145/1451983.1451991]

[3] Z. Gyongyi, H. Garcia-Molina. Spam – it's not just for inboxes anymore. IEEE Computer Magazine, 38(10):28–34, October 2005.