NewsPet

Michael Fulker, Anthony Hauber and **Tyson Williams**
COM S 472 : Principles of Artificial Intelligence
Department of Computer Science
Iowa State University, Ames, IA 50011
{calculo, thauber, tyson}@cs.iastate.edu

Abstract

For companies that do work updating data based on news stories, many manhours are wasted by reading irrelevant articles. We aim to enhance productivity of this work by creating and analyzing NewsPet: an automated news-feed categorizer which utilizes a Naive Bayes text classifier.

Introduction

//TODO

Architecture

//TODO

Database

//TODO

Implementation

Web Interface

//TODO

Trainer Service

The trainer service checks a message queue (see Message Queue subsection) to get training items (each of which consists of a classifier ID, a desired category ID, and a document's text). A batch of training items of the same classifier ID are collected, and a thread is spawned to process them. If multiple batches have the same classifier ID (for example, if new items are added to the message queue after the thread was spawned), the subsequent ones must wait for the first to finish.

The actual training is done with `NaiveBayesTrainer` objects from the Mallet framework (McCallum 2002). Each training thread acquires a `NaiveBayesTrainer` object (either deserialized from database, or instantiated for initial training), passes the training items through a filter such that they are converted to objects the Mallet framework is compatible with, runs these instances through the `ClassifierTrainer`, and persists the `NaiveBayesTrainer` back to the database.

Reader/Categorizer Service

//TODO

Message Queue

Both the Trainer Service and the Reader Service use the same Message Queue module (with different configuration and interpreting of messages). //TODO

Analysis

For testing, an instance of NewsPet was batch-trained for the following categories, using data from Reuters-21578 corpus (Lewis 1999):

1. //TODO
2. //TODO

//TODO

Results

//TODO

Conclusion

//TODO

References

Lewis, D. D. 1999. Reuters-21578 corpus, distribution 1.0. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.