Linguistic Dumpster Diving:
Geographical Classification of Arabic Text Using Words People Commonly Throw Away

Ron Zacharski[1], Ahmed Abdelali[2], Stephen Helmreich[2], and James Cowie[2]
University of Mary Washington[1], New Mexico State University[2]

In many text analysis tasks it is common to remove frequently occurring words as part of the pre-processing step prior to analysis. Frequent words are removed for two reasons: first, because they are unlikely to contribute in any meaningful way to the results; and, second, removing them can greatly reduce the amount of computation required for the analysis task. In the literature on information retrieval and text classification, such words have been called *noise in the system, fluff words,* and *non-significant words.* While the removal of frequent words is correct for many text analysis tasks, it is not correct for all tasks. There are many analysis tasks where frequent words play a crucial role. To cite just one example, Mosteller and Wallace in their seminal book on stylometrics noted that the frequencies of various function words could distinguish the writings of Alexander Hamilton and James Madison. We use a similar frequent word technique to geographically classify Arabic news stories. In representing a document, we throw away all content words and retain only the most frequent words. In this way, we represent each document by a vector of common word frequencies. In our study we used a collection of 4,167 Arabic documents from 5 newspapers (representing Egypt, Sudan, Libya, Syria, and the U.K.). We then train on this data using a sequential minimal optimization algorithm, and evaluate the approach using 10-fold cross-validation. Depending on the number of frequent words, results range from 92% classification accuracy to 99.8%.