Tutorial: Text Mining

Raymond J. Mooney
University of Texas at Austin

ABSTRACT:

Most data mining methods assume that the data to be mined is represented in a structured relational database. However, in many applications, available electronic information is in the form of unstructured natural-language documents rather than structured databases. This tutorial will review machine learning methods for text mining. First, we will review standard classification and clustering methods for text which assume a vector-space or "bag of words" representation of documents that ignores the order of words in text. We will discuss naive Bayes, Rocchio, nearest neighbor, and SVMs for classifying texts and hierarchical agglomerative, spherical k-means and Expectation Maximization (EM) methods for clustering texts. Next we will review information extraction (IE) methods that use sequence information to identify entities and relations in documents. We will discuss hidden Markov models (HMMs) and conditional random fields (CRFs) for sequence labeling and IE. We will motivate the methods discussed with applications in spam filtering, information retrieval, recommendation systems, and bioinformatics.