Lab 6: Document Classification

The raw data is text, and this is first converted into a form suitable for learning by creating a dictionary of terms from all the documents in the training corpus and making a numeric attribute for each term using Weka's unsupervised attribute filter *StringToWord- Vector*. There is also the class attribute, which gives the document's label.

Data with String Attributes

The *StringToWordVector* filter assumes that the document text is stored in an attribute of type *String*—a nominal attribute without a prespecified set of values. In the filtered data, this is replaced by a fixed set of numeric attributes, and the class attribute is put at the beginning, as the first attribute.

To perform document classification, first create an ARFF file with a string attribute that holds the document's text—declared in the header of the ARFF file using *@attribute document string*, where *document* is the name of the attribute. A nominal attribute is also needed to hold the document's classification.

Exercise 1. Make an ARFF file from the labeled mini-documents in Table 1 and run *StringToWordVector* with default options on this data. How many attributes are generated? Now change the value of the option *minTermFreq* to 2. What attributes are generated now?

Exercise 2. Build a *J48* decision tree from the last version of the data you generated.

Exercise 3. Classify the new documents in Table 2 based on the decision tree generated from the documents in Table 1. To apply the same

Table 1 Training Documents

Document Text	Classification
The price of crude oil has increased significantly	yes
Demand for crude oil outstrips supply	yes
Some people do not like the flavor of olive oil	no
The food was very oily	no
Crude oil is in short supply	yes
Use a bit of cooking oil in the frying pan	no

Table 2 Test Documents

Document Text	Classification
Oil platforms extract crude oil	unknown
Canola oil is supposed to be healthy	unknown
Iraq has significant oil reserves	unknown
There are different types of cooking oil	unknown

filter to both training and test documents, use *FilteredClassifier*, specifying the *StringToWordVector* filter and *J48* as the base classifier. Create an ARFF file from Table 2, using question marks for the missing class labels. Configure *FilteredClassifier* using default options for *StringToWordVector* and *J48*, and specify your new ARFF file as the test set. Make sure that you select *Output predictions* under *More options* in the Classify panel. Look at the model and the predictions it generates, and verify that they are consistent. What are the predictions?