CSE 5243: Introduction to Data Mining Assignment 1

Vaibhav Devekar  
Akshay Nikam

# Goal of the Assignment

This assignment aims at preprocessing a given rudimentary dataset to form structured datasets or feature vectors for further processing.

# Feature Vectors

We created several feature vectors from the Reuters articles:

### Data Matrix:

Data matrix is a two dimensional structured dataset, one dimension (Y) of which identifies each Reuters article while the other (X) has the extracted wordlist. Each entry (x, y) in the data matrix is the number of times the word (x) has appeared in the article (y).

### Transaction Matrix:

Transaction matrix is a two dimensional dataset which provides a list of words appeared in each article.

### Document Frequency and Inverse Document Frequency:

This dataset specifies the number of articles a word has appeared in and also the inverse of this value.

# Input Parsing and Document Frequency

We parse the input Reuters articles one by one and create a dictionary containing following items for each of them:

1. Frequency dictionary: dictionary of words appeared in the article as keys and the number of times they occur in the article as their values
2. Title: Title string of the article
3. Topics: List of topics in the article
4. Places: Places the article refers to in <PLACES> tag

As we parse through the input articles, we also record the number of articles a word appears in. This creates a **Document Frequency** map. We also compute inverse document frequency as:

Where N = total number of documents,

DF = Document Frequency

# Trimming wordlist based on thresholds

We sort the words by the frequency they appear in all the documents and remove the words that appear too many times or too few times. We identified threshold as 1% of the maximum document frequency and we remove words from both sides that are within this threshold.

# Stopwords filtering

Stopwords are the words that appear very frequently in the natural language and often do not have any interesting meaning from the knowledge discovery perspective. Such words should not be considered for data analysis. We provide a custom list of stopwords along with the submission which we use to filter the words from the Reuters articles.

# Class labels and weighted words

We identify words in <TOPICS> tags as class labels for the articles and along with words in <PLACES> tags, add them to the wordlist. We provide more weight to the words that appear in the titles as they hold more importance in that article. This is the final word list that we use in constructing the feature vectors.

# Data Matrix

From the reduced wordlist, we compute the data matrix in a two-dimensional data space where words are placed along X dimension and articles along Y. Each value is the number of times the corresponding word has appeared in the article.

# Transaction Matrix

From the data matrix generated in previous step, we can compute the transaction matrix which is a list of articles with each article specifying the list of words that appear in that article.

# Individual Contribution

For this assignment, we both worked together on all parts of assignment.