COMP41450 Machine Learning

Assignment: Non-negative Matrix Factorization

Data Format

A sample data set is provided for the assignment (see *bbcnews_data.zip*), which represents a collection of news articles from the BBC website. It consists of two files:

File: bbcnews.mtx

This is the sparse term-document matrix stored in the Matrix Market format (http://math.nist.gov/MatrixMarket/), with rows representing 4058 unique terms (words) and columns representing 1400 unique documents. An entry (i,j) with value *f* in the matrix indicates that the term *i* appears in document *j* with frequency *f*. Zero values are not stored in the file. The file has the following format:

```
%%MatrixMarket matrix coordinate real general
4058 1400 161462
1 1 1.0000
1 10 1.0000
1 557 1.0000
...
```

The first line of the file is a header, which can be ignored. The second line contains 3 values. These values indicate the number of rows, number of columns, and total number of non-zero values in the matrix. All lines after this contain 3 values: the row (indexed from 1), the column (index from 1), and the term frequency *f*. So the line "1 557 1.0000" indicates that term 1 appears in document 557 once, while the line "1 595 2.0000" indicates that term 1 appears in document 595 twice.

File: bbcnews.terms

This file lists all of the 4058 terms (words) in the data set. Each line corresponds to a row in the term-document matrix, with one term per line. For example, the second line is "sales". This indicates that the values on the 2nd row of the matrix all relate to the term "sales".