Math 444 – Data Mining Spring 2017

Exercise 5

1. The following problem applies ideas from text mining to interpretation of genetic codes.

   The information needed for a living cell to function is encoded in its DNA. The code is written by using four nucleotides that are represented by the letters A, C, G and T. A distinctive message in the genomic sequence is a piece of text, called *a gene*. Within a gene, the biological information is encoded by means of sequences of three letters, called *codons*. The codons are read and they express themselves, e.g., in proteins that the cell itself is able to produce, based on its genetic code. in the following, we interpret the genetic sequence as a collection of documents and apply simple data mining ideas to see if there are distinct document types hidden in the genetic code.

   Download the attached data file `ccrescentus.fa`. This is a Fasta-format file containing a fragment of the genomic sequence of the bacterium *Caulobacter Crescentus*. If you are more interested what you are looking at, see the web page http://caulo.stanford.edu/caulo, or for more comprehensive information, the article Nierman WC et al. (2001) *Complete genome sequence of Caulobacter crescentus* Proc Natl Acad Sci U S A 98(7): 4136-41.

   (a) Download then the Matlab program `LoadSeq.m`, a file that reads Fasta-files and converts them to a Matlab string. Type

   ```
    genseq = LoadSeq('ccrescentus.fa');
   ```

   and check what you have got in your workspace.

   (b) The variable `geneseq` is a long sequence of single letters. We now break the sequence in single documents, each document being of length L. Furthermore, we interpret that groups of k letters represent a word. More precisely:

   Suppose that we decide to consider each letter a word. In this case, the number of different words in our dictionary would be four (A, C, T or G). Likewise, if we decide that $k = 2$, the words consist of two letters (AA, AC, AT etc.), the number of words in the dictionary is $4^2 = 16$. More generally, the number of words of k letters is 4^k.

   Once we have decided what L and k are, we can form a term-document matrix. Download the attached Matlab program `CalcFreq.m`. This program splits the gene sequence in documents of desired length, and creates the term-document matrix of the genetic code:

   ```
   A = CalcFreq(genseq,k,L);
   A = A';
   ```

   produces the term-document matrix A by breaking the code to L letter documents and interpreting all k-letter sequences as words (Check the dimensions, making sure that you have the right number of words in your dictionary.)

   (c) The question that we ask now is: Does the data have a special structure with some particular word length?

   To answer this question, we visualize the data. Fix your L, e.g., $L = 300$, and using different values of k, e.g., $k = 1, 2, \ldots, 9$, compute the corresponding term-document matrix A, center your data, and plot a scatter plot of the two first principal components of your data. Does the choice $k = 3$ corresponding to the codon length give a qualitatively different plot?

   Test if the result is sensitive to the choice of L, by running your program with a couple of different values of L. Test also if the result is sensitive to where you start a word, by removing a couple of letters from the beginning of the genetic code and rerunning your program.

(d) For $k = 3$, run a clustering algorithm, plotting the PCA scatter plots by using different colors for different clusters.

(e) Compute the NMF of your term-document matrix (The non-centered version), with the number of feature vectors equal to the number of clusters. Then plot the PCA components of the feature vectors on top of the PCA plot of your data. Do the feature vectors represent the clusters?

(f) Using the k-medoids algorithm to find a medoid of each cluster. Then compute the cosine of the angle between the term-document vectors and the medoids. In particular, if you would use one of the medoid vectors as a query, how well would you identify the corresponding cluster?