CSCD 429 Data Mining HW2 (30 Points + 10 Extra)
Due: February 11, 2015, 11:59pm

Prediction of gene/protein localization

- **Data Set Description:** This dataset was used in the 2001 kdd cup data mining competition. (http://www.cs.wisc.edu/~dpage/kddcup2001/). There were in fact two tasks in the competition with this dataset, the prediction of the "Function" attribute, and prediction of the "Localization" attribute. **Here we focus on the latter** (this is somewhat easier as genes can have many functions, but only one localization, at least in this dataset). The dataset provides a variety of details about the several genes of one particular type of organism. The main dataset (*Genes_relation.data* and *Genes_related.test*) contains row data of the following form:

  Gene ID, Essential, Class, Complex, Phenotype, Motif, Chromosome Number, Function, ***Localization****.*

  The description of data attributes was given in file *Genes_relation.names*. The first attribute is a discrete variable corresponding to the gene (there are 1243 gene values). Also the remaining 8 attributes consist of discrete variables, most of them related to the proteins coded by the gene, e.g. the "Function" attribute describes some crucial functions the respective protein is involved in, and the "Localization" is simply the part of the cell where the protein is localized. In addition to the data of the above form, there are also data files (*Interactions_relations.data* and *Interactions_relation.test*) which contain information about interactions between pairs of genes.
  - **Size**
    - Gene_relation files: 6275 examples (4346 training, 1929 test), 9 categorical attributes.
    - Interaction_relation files: 1806 records, 2 attributes (one categorical; one numerical)
  - **References**:
    - Talk overview slides about this problem and also the winner presentation in the KDD 2001 competition can be found on-line.
    - See also Answers to Questions of General Interest from Question Period 1 and Answers to Questions of General Interest from Question Period 2
  - **Task**: Perform exploratory data analysis to get a good feel for the data and prepare the data for data mining. **The task in this dataset is to make predictions on the attribute "Localization". To simplify your work, you may not use the interactions data**. Detailed knowledge of the biology should not be necessary for this application. One word of caution: **your classifier for localization should not use "function"**, since *both* fields will be withheld from the test genes when they are provided.
  - **Challenge**: This dataset is a great challenge. One issue is that there is a high proportion of missing variables in the Genes_relation data.
  - **Keys**: The keys are provided in the file keys.txt. Use this file to evaluate the accuracy of your solution.
  - **Deliverables**:
    - (20 points) All program files: in this assignment, you are required to design and implement a classification algorithm to predict gene localization. **You may choose any classification algorithms we covered in class and implement it in any programming languages**.
    -
    - (3 points) A result file in the format of **<gene ID>, <localization>** in each row
    - (7 points) A report includes
      - how to run your program;
      - the method you used to build the classifier;
      - the method you used to handle missing data;
      - the accuracy of your solution;

- **Bonus (10 points)** Use RapidMiner to tackle the same task, and compare the results generated by your own program with RapidMiner. Please document how you used it and the result generated by RapidMiner.
- Include all the files into a single .zip file and **submit your file via Canvas**.