Data Mining Exercise 3: 06.02.2020

1. Let us consider file bloodp.xls (import). File contains two columns that are systolic (sbp) and diastolic (dbp) blood pressure values in mmHg. Replace zero values using mean value of variable in question. Replace also missing values using the mean. Correct the erroneous values using the following rules. sbp must be greater than 80. Values below must be multiplied by 10. dbp must be over 40. Values below must be multiplied by 10. sbp over 300 or dbp over 160 are impossible: remove.
2. Build a linear regression model that can be used in predicting dbp by sdb and a constant term. You can do this by creating an observation matrix **O**=[**1 sdb dbp**] where first column is filled with ones, second and third columns are systolic and diastolic blood pressure values from task 1. select **y**=**dbp** and **X**=[**1 sdb**]. Coefficients for your model are $\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ where T is transpose operation and (-1) means inverse matrix. In this task we do not consider statistical confidence of the parameter **b**. Matlab has function regress() that is normally used in linear regression.
3. In text mining tasks we might be interested in finding a document that is as close as possible to a reference document. Let us consider that we have our set of keywords S={$word_1$, $word_2$,…,$word_k$}. Frequencies of keyword occurences in our refrerence document is **Fo**= [15, 7, 6, 11, 4] and number of words N_w=500. In addition, we have two other documents that has respective values of $\mathbf{Fo_1}$=[1, 4, 3, 3, 6] N_{w1}=200, $\mathbf{Fo_2}$=[20, 1, 5, 16, 9] And N_{w2}=210. Calculate the cosine distance between our reference document and the two other documents. Normalize first the word occurrences with respective word count. Cosine distance between normalized documents **x** and **y** is calculated using formula. (Numerator is a dot product)

$$d(\mathbf{x},\mathbf{y})=1-\frac{\mathbf{x}.\mathbf{y}}{|x||\mathbf{y}|}$$

4. Load in the image file numbers.jpg. Use script dm2.m and convert image into binary form. Calculate the hamming distances between 5^{th} column of the image between all other columns. What is the greatest distance?
5. Calculate binary correlation of image columns. What columns have the greatest correlation?
6. Show that cosine measure is not a metric.