# 1. Spam Filter: Naïve Bayes Learning

Implement a spam filter using Naïve Bayes learning on Enron emails dataset. You can download it from the link[1]. Download only Enron1, and Enron6. Then, unzip the files and create a folder; each text file is an email; spam and ham go in separate folders.

You should implement the Naïve Bayes learning algorithm described in class, that means you cannot use any Naïve Bayes package already implemented, i.e. *sklearn*, *nltk*, and apply it to learn a classifier that separate spam from ham email as accurately possible. You must write a preprocessor that converts an email message into a vector of feature values (words), calculating probabilities, and so on.

Split dataset 70% and 30% for training/testing. Train two different models; first trained from Enron1, and second from Enron6. Report confusion matrix and your prediction accuracy for each test sets with both models.

# 2. Anomaly Detection: Nelson Rules for Control Chart

Western Electric Company Rules (WECO) have been widely used for Shewhart control charts in order to increase the sensitivity of detecting assignable causes of process change.

A 1984 update by Lloyd S. Nelson to the popular Western Electric Rules, in order to make the probability of detecting an out of control condition by chance approximately equal across all tests. The Western Electric Rules have probabilities that have more variability from one another (some are more likely than others to occur by chance). There are eight different Nelson Rules that are explained on below.

Implement a Python script that checks anomaly in daily credit card expenditure dataset according to average weekly expenses (The expenditures are in today's dollar, and 7 days=1 week). This script uses the first 52 weeks in the training (for calculating $\bar{x}$ and $\sigma$), so rest for testing.

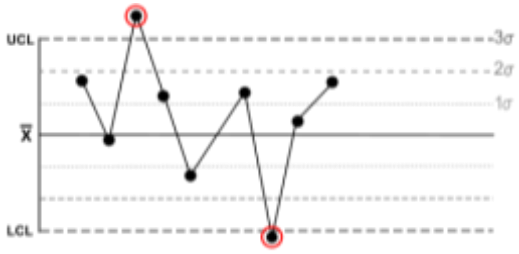
Generate a table as csv file that should be as follows:

| Week | Rule 1 | Rule 2 | Rule 3 | Rule 4 | Rule 5 | Rule 6 | Rule 7 | Rule 8 |
|---|---|---|---|---|---|---|---|---|
| 52 | | | | | | | | |
| … | | | | | | | | |
| 156 | | | | | | | | |

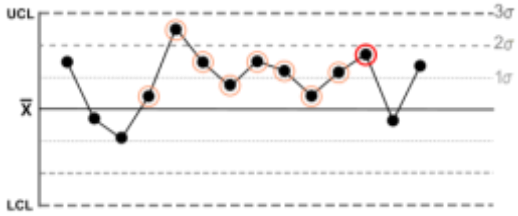
The code should fill the cell by 1, if there is an anomaly for the respected week and rule; otherwise, the value of the cell should be 0.

[1] *http://www2.aueb.gr/users/ion/data/enron-spam*

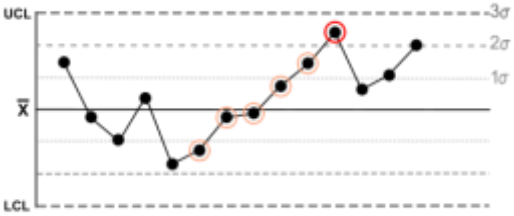
Rule 1: One point is more than 3 standard deviations from the mean (outlier)



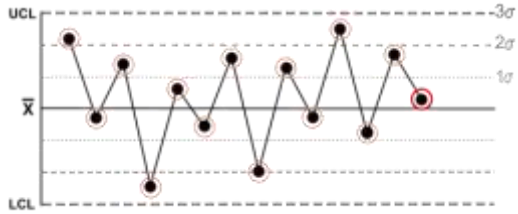

Rule 2: Nine (or more) points in a row are on the same side of the mean (shift)



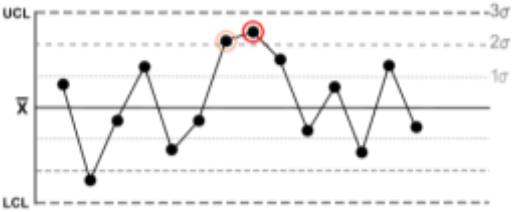

Rule 3: Six (or more) points in a row are continually increasing (or decreasing) (trend)



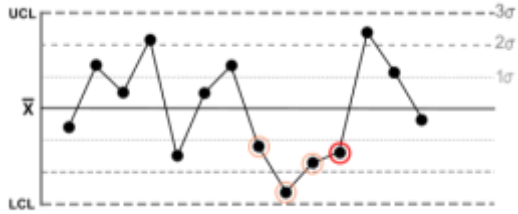

Rule 4: Fourteen (or more) points in a row alternate in direction, increasing then decreasing (bimodal, 2 or more factors in data set)



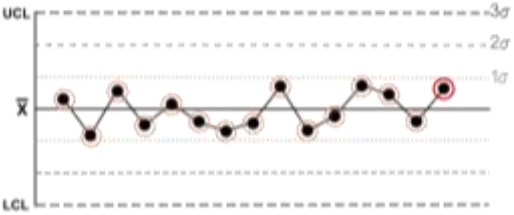

Rule 5: Two (or three) out of three points in a row are more than 2 standard deviations from the mean in the same direction (shift)




Rule 6: Four (or five) out of five points in a row are more than 1 standard deviation from the mean in the same direction (shift or trend)




Rule 7: Fifteen points in a row are all within 1 standard deviation of the mean on either side of the mean (reduced variation or measurement issue)




Rule 8: Eight points in a row exist with none within 1 standard deviation of the mean and the points are in both directions from the mean (bimodal, 2 or more factors in data set)