Assignment 3, STA5206

Provide all the code and outputs along with your answers. Preferably in notebooks that can be loaded. Placing answers of text in ’markdown’ is advisable.

Descriptive answers are of importance. Please do not be ’laconic’/’brief’ in your descriptions. What is meant that although your code works, the plots are correct, and you provide an accurate comment; adding some ’insight’ for the reader as to the conclusions and outcomes of your analysis are considered noteworthy and earn points. If there are 4 conclusions to be derived they should all be stated. The check list for your grade will be:

- does the code work?
- is the code succinct and easily readable?
- is the code commented to explain what is being done?
- does the output follow the code so that it is easily seen what code produces which outputs? (please do not put all the code in a segment with a series of outputs, that will result in points deducted)
- Do the conclusions and insight comments follow the inputs?
- Do the conclusions for each answer provided point out all the interesting features of the analysis?
- Does the answer discuss possible ambiguities in the results?

Effectively each segment of answers requires code/outputs/discussion

Quite often students lose points for incomplete discussions. You must display an understanding of the outputs and the value generated. Another place where you may lose points is on quality of the code and figures. If you can make the figures look better and be more clearly presented, then do so. If you need to standardise the data, then do so, because you will not be directed to do so when it is expected.

Question 1 (10 pnts)

You are provided with 2 synthetically generated json files, ’emails.json’ and ’emails_new.json’. These can be imported and parsed with JuliaLang via

$usingJSON; emails\_parsed = JSON.parsefile(”emails.json”);$

- Import the 2 datasets and print the first email element of each parsed JSON, and also print how many unique words exist in each string. (you may need to ’split’ the string into individual words) (2pnts)
- Produce a vector of all the unique words from all the emails which includes both files and also the length. (1pnt)
- Produce a function which will take in an email (string), and return a ’word vector’ of the normalized counts. This word vector is produced by making a dictionary with keys as all the possible words with values set to zero. Then for each word in the email seen the count of the word increments the value of the word key. Then normalize that vector. (2pnts)
- Produce a matrix for all the word vectors from the emails.json and then another matrix for the emails_new.json. Each word vector comes from a single email string and the matrix will have these vectors appended to each other. (1pnt)
- Fit a PCA model to the data for the emails.json data and then visualize the values of the eigen values. (1pnts)
- Visualize the transfomed data from the 2 largest eigenvalue eigenvectors and project onto a scatter plot. (1pnt)
- Project the points of the emails_new.json data onto the previous scatter plot and identify the email which is an outlier. (2pnts)

Question 2 (10 pnts)

The data for this question comes from `https://github.com/vincentarelbundock/Rdatasets/blob/master/csv/datasets/EuStockMarkets.csv` which can be imported from R datasets directly as ’EuStockMarkets’. Each row is a day of market values.

- Load the data of EuStockMarkets.csv (1pnt)
- Fit a PCA model to the normalized dataset (2pnts)

- Plot the transformed data on the 2 largest eigenvalues on a scatter plot (3pnts)
- Plot the transformed data on a scatter plot where the color of the marker corresponds to the rownumber of the datapoint (2pnts)
- Plot the projection along each PCA component on a separate 'radial' (polar) plot. Make the labels on the circumfrence the months of the year. In Julia the basic plot can be achieved 'angles = row_numbers .* (2pi / 365)', 'plot(angles, component1projections, proj=:polar)'. (2pnts)