Course: INT 93S
Quarter: Summer 2017
Assigned: 5 July 2017
Due: 9:00 11 July 2017

Lab 3: Analysis and Visualization of Data

Collaboration Guidelines: For lab assignments, code can be written together (with your lab partner) and turned in separately. Written analysis for the lab, while it can be discussed, should be written separately.

Lab Computer Setup

The lab computers have a python distribution called Enthought Python Distribution.
https://www.enthought.com/products/epd/

1. To access it search for "Canopy" in the Windows Search Box
2. You can add packages:
    a. Click on "Package Manager"
    b. Click on "Available"
    c. Search for package name and select it from the list
    d. Click "Install"
    e. This lab uses numpy, scipy, matplotlib and bokeh.
3. Use Canopy's Editor to access these libraries. (they will not show up in IDL)

Turn-in

1. Create a readme file titled "lastname_firstname_readme.txt" which includes:
    a. First name and last name of you and your partner
    b. How much of the lab you finished in class
    c. Any references you used
    d. Anything you would like the professor and TA to know.
2. Put the files for Part 1, 2, 3 and 4 in a folder named "lastname_firstname_lab3"
3. Zip the folder and upload it to Gauchospace

Part 1: Analyze and plot time series data using Matplotlib library *(6 points)*
All plots and visualizations in this lab must include a proper title, labels, and a legend.

1. Import the csv file "harryPotter-film-novel.csv" and format it as a numpy array. *(1pt)*
2. Using matplotlib, plot the data, and save the plot as "lab3_part1_graph1.png" *(1pt)*
3. Using numpy, create a correlation matrix with the data, graph it with matplotlib, and save it as "lab3_part1_graph2.png" *(1pt)*
4. Analyze the two columns that have the greatest correlation coefficient: *(2pts)*
    a. Using scipy, calculate the Pearson's Correlation Coefficient and p-value of the two columns.
    b. Plot the two columns using matplotlib. Save the plot as "lab3_part1_graph3.png"

    c. Normalize the two columns using numpy and plot them using matplotlib. Save the plot as “lab3_part1_graph4.png”
5. In a txt file “lab3_part1.txt” answer the following questions. *(1pt)*
    a. Which two columns did you choose?
    b. What was the Pearson's Correlation Coefficient and p-value of the two columns?
    c. Do you think that this correlation is valid? Why?
    d. What conclusions can you draw from this data?

Part 2: Visualize time series data using Bokeh library *(4 points)*

1. Plot the data from Part 1 (2 columns minimum) using the bokeh library, using data visualization concepts that support the data. *(2pt)*
2. Include a form of interactivity. ie: hovering over or clicking on data gives you more information about the data. *(1pt)*
3. In a txt file “lab3_part2.txt” answer the following question. *(1pt)*
    a. Explain what were you trying to show with this visualization.
    b. List 3 design choices you made and explain how they support the data.

Part 3: Extra credit *(.5 pts)*

1. Modify the code to use the pytrends library: https://github.com/GeneralMills/pytrends
2. Have the code prompt you for a username and password for the API. http://anh.cs.luc.edu/python/hands-on/3.1/handsonHtml/io.html
3. Query the library for 5+ data streams.
4. Run the same analysis as Part 1, producing and saving the same graphs.
5. In a text file named “lab3_part3.txt”, answer the questions from Part 1 Step 7 about the data.

Part 4: Extra credit *(.5 pts)*

1. Import a dataset that you are potentially using for your research project, your partner’s research project, or another dataset that you are interested in.
2. Chart a portion of it using the bokeh library a meaningful way.
3. Write a paragraph which answers these questions:
    a. What dataset did you choose?
    b. What design choices did you make and how does it support the data?
    c. What conclusions can you draw from this data?