HW1 - Python Basics

29/10/2017

Submission date: 14/11/17

- Submission is individual or in pairs.
- Submit a ZIP file containing your files named with 9 digit of your ID. Submission Example: '200567989.zip'.
- Submit your HW solution as an 'Jupyter/IPython notebook', after you zip your files extract them into a new folder and make sure there are no runtime errors.
- Use python version 2.7.

1 Warm up

1. Summarize the following packages capabilities and usage in your own words (in 2-3 lines, Hebrew/English), then pick one function from each package that you think is useful and explain its usage.
    - numpy
    - scipy
    - matplotlib
    - pandas
    - scikit-learn
    - NLTK

2 Setting the environment

1. Set up python environment. Download Anaconda, version 3.6.

3 Basic Python 0

1. Draw N random iid samples from a 1D Gaussian distribution with $\mu = 0, \sigma^2 = 1$(You can use numpy.random library).
2. Calculate the empirical mean of the random samples.
3. Plot the difference between the empirical mean and the true mean as a function of N.
4. On the same figure, add (you can use error bar function) the standard deviation of the empirical mean.

   The standard deviation of the empirical mean can be calculated in a straight forward way, and is given by:

   $$\sigma_{\hat{\mu}} = \sqrt{\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)} = \frac{\sigma}{\sqrt{N}}$$

   Where σ is the standard deviation of the iid Gaussian distribution, and $\sigma_{\hat{\mu}}$ is the standard deviation of the empirical mean.

4 Basic Python 1

1. Load the ‘airports.txt’ file that is found in the course website, and print the first line.
2. Print all airport names of a specific country you chose.
3. Write a program that counts the number of airports in each country.

5 Basic Python 2

1. Download the routes data (https://openflights.org/data.html). This database stores every unique flight route that OpenFlights knows about. Take a moment to look at the fields available in the routes data (listed on the OpenFlights page).
2. Use both data sources and calculate the length of each route. Plot a histogram showing the distribution of distances flown, following these guidelines:

- Read the airports file (airports.txt) and build a dictionary mapping the unique airport ID to the geographical coordinates (latitude and longitude.) This allows you to look up the location of each airport by its ID.
- Read the routes file (routes.dat) and get the IDs of the source and destination airports. Look up the latitude and longitude based on the ID. Using those coordinates, calculate the length of the route (using geo_distance.py) and append it to a list of all route lengths.
- Plot a histogram based on the route lengths, to show the distribution of different flight distances.