INF 551 – Spring 2017

Homework #2

Due: February 28
100 points

1. [80 points] Consider the Los Angeles International Airport (LAX) traffic data *"lax.json"*(provided to you in the same folder as this handout). This *JSON* data file lists the number of passengers who departed from or arrived at LAX terminals at the first day of each month, starting from January, 2006 to November, 2016. For example, the 5th record says there were 401,535 people arriving at Terminal 1 on 1/1/2006 through domestic flights.

   Task: Write a **Python 2.7** script *"lax.py"* that takes *"lax.json"* and a list of keywords (that specify the search condition) as the input, and outputs (prints to screen rather than writes to a file) the following statistics for the records that satisfy the search condition: **min, max, median, average, and standard deviation** (comprises only numbers exactly in this order and separated by commas). For example, min is the minimum number of passengers among all qualified records.

   The keywords in the list are separated by white spaces and may only contain keywords of 3 categories: **terminal**, **year**, and **traffic type** (departure/arrival).
   - Accepted keywords for terminals are: T1 (for terminal 1), T2, …, T6, and TBI (for Tom Bradley international terminal). Keywords are **case-insensitive**. *"t1"* should be recognized the same as *"T1"*, for example.
   - Year is a four-digit number, e.g., 2006.
   - Traffic type is either departure or arrival.

   Execution: *python lax.py lax.json "tbi t5 arrival 2015 2013"*, which returns the statistics of all flights arriving at TBI and T5 in 2013 and 2015.

   Submission: <FirstName>_<LastName>_lax.py

   Clarifications (read carefully before posting any questions on Discussion Board):
   - *"lax.json"* is the only input file that will be used to test your code in this assignment, do take advantage of this. However, you still need to take it as the first argument (*sys.argv[1]*).
   - The order of keywords is arbitrary, deliberate on their patterns and how they can be distinguished. Same as before, the input is always valid and no need to do error checks.
   - You need to consider the scenario where not keywords from all 3 categories appear. For instance, *"2015 t1"*, as a valid keyword list, should be taken to return both arrival and departure traffic at t1 in 2015. Similarly, missing keywords of year and terminal means all years and all terminals should be taken into account.

- The minimum number of keywords is 1 and no maximum, and they should be put in double quote together as one command line argument (*sys.argv[2]*). There could be multiple keywords in a same category, e.g. *"2015 2016 t1 t2"* includes traffic data in both years and at both terminals. The correct execution is ***python lax.py lax.json "2015 2016 t1 t2"*** rather than *lax.py lax.json 2015 2016 t1 t2.*
- In this assignment we only handle terminals from *T1* to *T6* and *TBI*. If no terminal keyword is given, then only and all of *T1* to *T6* and *TBI* should be considered. Data for the other terminals is ignored.
- Calculation of these 5 types of statistics, namely min, max, median, average, and standard deviation, should be implemented by yourself. **No external libraries** are allowed.

2. [20 points] Use the provided *"cds.xml"*, write XPath expressions to answer the following questions:
    a. Find all CDs not released in "UK".
    b. Final all CDs by "Bob Dylan" or "Kenny Rogers".
    c. Find all CDs with price > 10.
    d. Find the titles of CDs produced by "Polydor" and having price > 10.
    e. Find the artists of CDs with rating > 3 and released after 1990.
    f. Find the titles of CDs with the specification of their language.
    g. Find the titles of CDs with no ratings.
    h. Find the titles of CDs by artists whose name contains "Rod".

    Submission: <FirstName>_<LastName>_cds.txt (that contains all your XPath expressions)