INF 551 – Fall 2018

Homework #2:

Due: October 7, Sunday (end of day)
100 points

In this homework, we ask to take the same data set "prize.json" as in homework #1, convert it into XML document, build an inverted index for "motivation" field, and use the index to answer search questions. Specific tasks are as follows.

1. Implement a Python script "convert.py" that takes prize.json and convert it into prize.xml. The output XML documents should have the following format:

```
<prizes>
  <physics>
    <laurate id="941" year="2017">
      <firstname>Rainer</firstname>
      <surname>Weiss</surname>
      <motivation>
        for decisive contributions to the LIGO detector and the observation of gravitational waves
      </motivation>
      <share>2</share>
    </laurate>
    <laurate id="942" year="2017">
      <firstname>Barry C.</firstname>
      <surname>Barish</surname>
      <motivation>
        for decisive contributions to the LIGO detector and the observation of gravitational waves
      </motivation>
      <share>4</share>
    </laurate>
    ...
  </physics>
  <chemistry>...</chemistry>
  ...
</prizes>
```

   Execution format: python convert.py prize.json prize.xml

2. Implement a Python program "index.py" that takes prize.xml and creates an inverted index for the motivation field. Store the index in a file "index.xml" in the XML format as follows.

```
<index>
        <entry>
                <keyword>ligo</keyword>
                <ids><id>941</id><id>942</id></ids>
        </entry>
        …
</index>
```

   Execution format: python index.py prize.xml index.xml

3. Implement a search program "search.py" that takes a list of keywords (which may contain multiple tokens in a list) and return a list of ids of laurates whose motivation field contain one or more keywords in the list. Your program should utilize the index.xml created above.

Execution format: python search.py index.xml “ligo waves”
Sample output: [941,942,943]

In this case, “ligo wave” represent a list contains keywords “ligo” and “waves”.

You may use Python json and lxml packages for this homework.
You’d better test your python code on EC2 before submission.

Submissions: Name your 2 scripts as below and submit to Blackboard by the due time. **DO NOT** place them in a folder or zip file.

- <FirstName>_<LastName>_convert.py
- <FirstName>_<LastName>_index.py
- <FirstName>_<LastName>_search.py “ligo waves”

Eg:
Student name: Mike James
Execution format:
python Mike_James_convert.py prize.json prize.xml
python Mike_James_convert.py prize.xml index.xml
python Mike_James_convert.py index.xml “ligo waves”

Note: Please use Python 2.7 (installed by default on EC2) for the coursework.

Grading Policy:

1. Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. Late homework will be deducted 10% of its points for every 24 hours that it is late. No credit will be given after 72 hours of its due time.
2. If your python code cannot be run with the commands as above, there will be 40 % penalty based on the points you get.
3. If you use non-standard python packages (except Python json and lxml packages for this homework), there will be 30 points penalty.