INF 551 – Spring 2019

Homework #2:

Due: March 8, Friday (end of day)
100 points

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

1. Implement a Python script "convert.py" that takes accounts.json and convert it into accounts.xml. Your program should ignore the "index" lines in the data set (e.g., {"index":{"_id":"1"}}). The output XML documents should have the following format:

```
▼<accounts>
  ▼<account number="1">
      <balance>39225</balance>
      <firstname>Amber</firstname>
      <lastname>Duke</lastname>
      <age>32</age>
      <gender>M</gender>
      <address>880 Holmes Lane</address>
      <employer>Pyrami</employer>
      <email>amberduke@pyrami.com</email>
      <city>Brogan</city>
      <state>IL</state>
    </account>
    ...
  </accounts>
```

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

2. Implement a Python program "index.py" that takes accounts.xml and creates an inverted index for the address field. The index should have an entry for each unique keyword (including the street number, e.g., 880) appearing in the addresses. The entry should list the number of accounts whose address contains the keyword. Store the index in a file "index.xml" in the XML format as follows. Store the keyword in lowercase.

```
<index>
        <entry>
                <keyword>lane</keyword>
                <accounts>
                        <number>1</number>
                        <number>70</number>
                                ...

                <accounts>

        </entry>
```

…
</index>

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

3. Implement a search program “search.py” that takes a list of keywords and return the numbers of accounts whose address contains one or more keywords in the list. Your program should utilize the index.xml created above.

   Execution format: python search.py index.xml "mill lane"

   This will return the numbers of accounts whose address has mill or lane or both (case-insensitive). For example, if your unique word "mill" has appeared in accounts [2, 3, 4, 5] and "lane" has appeared in accounts [4, 7, 8], your search.py should return **ONLY** one list that contains [2, 3, 4, 5, 7, 8] as your result (no duplicates).

Submissions

1. Name your 3 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

2. For example (Student Name : Mike James) :
   Execution format:
   - python Mike_James_convert.py accounts.json accounts.xml
   - python Mike_James_index.py accounts.xml index.xml
   - python Mike_James_search.py index.xml "mill lane"

Grading Policy:

1. Homework assignments are due at 11:59pm on the due date and should be submitted on Blackboard. Late homework will be deducted by 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 a 40 point penalty.
3. If you use non-standard python packages (except Python json and lxml packages for this homework), there will be a 30 point penalty.
4. If your convert.py and index.py take more than 5 minutes each to finish, there will be a 20 point penalty.
5. Please use Python 2.7 (installed by default on EC2) for the coursework, there will be a 20 point penalty if you use a different version.