DSCI 551 – HW2 (Spring 2021)

Searching XML collections using Python & lxml

100 points, Due 2/28

In this homework, you will be provided a collection of XML documents put together in an input directory. Your task is to build an inverted index for searching the text contents/values of elements and values of attributes over the documents. Libraries permitted in this homework are: os, sys, re, xml, and lxml only. Specially, there are two tasks:

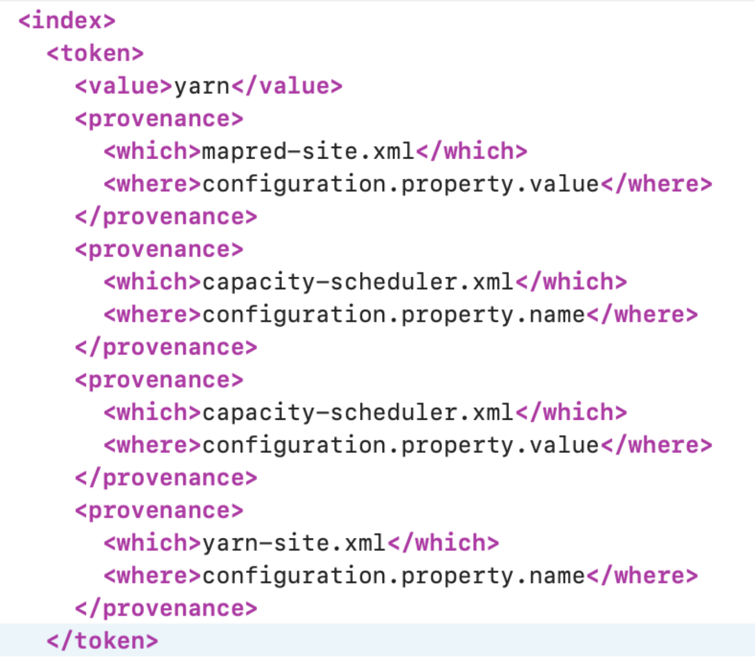
1. Indexing: write a Python script "index.py" that take an "input" directory with a collection of XML documents, then output an index file also in XML format. The index should contain for each token, the token's value, and all of this token's provenance. 'which' tag should include the name of the XML file that has the token, and 'where' tag should include the path to find the token in relevant file.

   You can assume values are tokenized by white spaces and punctuation characters. For example: 'yarn.nodemanager.aux-services' can be tokenized to ['yarn', 'nodemanager', 'aux', 'services']. (You don't need to consider about numbers like '1.0'.)

   Execution format:

   ```
   Python index.py input index.xml
   ```

   Example of output file:



```
<index>
  <token>
    <value>yarn</value>
    <provenance>
      <which>mapred-site.xml</which>
      <where>configuration.property.value</where>
    </provenance>
    <provenance>
      <which>capacity-scheduler.xml</which>
      <where>configuration.property.name</where>
    </provenance>
    <provenance>
      <which>capacity-scheduler.xml</which>
      <where>configuration.property.value</where>
    </provenance>
    <provenance>
      <which>yarn-site.xml</which>
      <where>configuration.property.name</where>
    </provenance>
  </token>
```

2. Searching: write a Python script "search.py" that takes the same "input" directory, the index file you created in part 1, and a list of keywords (separated by white spaces). It should return all elements whose text content or attribute contains at least one of the keywords. For each element, indicate which XML it comes from. If no such keywords in the file, return “No such tokens”.

   Sample execution format:

   ```
   Python search.py index.xml input "hdfs replication"
   ```

   Should return all elements whose values contain hdfs or replication or both.

   Example output:

   ```
   Element: <value>hdfs://localhost:9000</value>
   File: core-site.xml
   Element: <name>dfs.replication</name>
   File: hdfs-site.xml
   ```

Submission:

1. Please include 2 python scripts and 1 output files below in a folder: index.py, search.py, index.xml. Please make sure all your scripts’ names are correct and do not add directories Q1, Q2…
2. Scripts that take long run time (e.g. more than 2 minutes) will lead to deductions of points.
3. Please use Python 3.8 and make sure your scripts can work on ec2 for all the homework.
4. Name your zip file : Firstname_Lastname_hw2.zip.