HW2 (HDFS and XML)

DSCI 551, Fall 2022

100 points

Due: September 30, Friday, 11:59pm

Write a Python script, named xml2tsv.py, which takes an XML file of HDFS file system image and convert it to TSV format. The TSV file is similar to that generated by the "hdfs oiv" command, but only has the following five columns:

Path	ModificationTime	BlocksCount	FileSize	Permission
/	9/5/2022 22:59	0	0	drwxr-xr-x
/user	9/7/2022 1:18	0	0	drwxr-xr-x
/user/john	9/6/2022 20:05	0	0	drwxr-xr-x
/user/john/a	9/6/2022 20:07	0	0	drwxr-xr-x
/user/john/WordCount.java	9/6/2022 20:05	1	3269	-rw-r--r--
/user/mary	9/7/2022 1:19	0	0	drwxr-xr-x
/user/mary/a	9/7/2022 1:20	0	0	drwxr-xr-x
/user/mary/a/hello.txt	9/7/2022 1:20	1	12	-rw-r--r--

Note the path column should list all directories and files in the system, and the modification time and permission should have the above format. Note that d in permission means directory.

Sample execution format:

```
python3 fsimage70.xml fsimage70.tsv
```

where fsimage70.xml is the input XML file and fsimage70.tsv is the output TSV file.

Note that you should not assume the first inode in the INodeSection is for root and the first directory element in INodeDirectorySection is for the root directory. Instead, find the inumber of root by searching for inode with **empty name element**.

Requirements: you should use **xpath** function of lxml in your code to retrieve elements and values.

Submission: your xml2tsv.py script.

Permitted libraries: lxml, pandas, datetime