Instructions for Wikipedia dataset Execution

CS555: Big Data Computing Lab

16th August 2021

1. Install Maven
    a. Open the terminal and type:

       sudo apt-get install maven

2. create project directory : (wiki)
    a. Browse to home directory

       cd /home/iitp

    b. Create project directory

       mkdir wiki

3. Create source files:
    a. Browse into wiki directory

       cd wiki

    b. Create new folder **src** inside wiki directory

       mkdir src

    c. Browse into src directory

       cd src

    d. Create new folder **main** inside src directory

       mkdir main

    e. Browse into **main** directory

       cd main

    f. Create new folder **java** inside main directory

       mkdir java

    g. Browse into java directory

cd java

h. Create source files within java directory
   i. **nano DriverClass.java**
   ii. **nano InvertedIndex.java**
   iii. **nano LinkGraph.java**
   iv. **nano PageRank.java**
   v. **nano PageRankSorting.java**

   paste the lines from the source code provided

   Note:- **To save file: Press- CTRL + o followed by Enter button**

   To Exit Press:- CTRL + x from nano editor

i. Browse into wiki directory

   cd /home/iitp/wiki

j. Create pom.xml file inside pagerank directory

   nano pom.xml

   paste the lines from the xml file provided

k. Compile the java source files

   mvn clean && mvn compile && maven package

4. Create input directory (inputdata) for input files

   cd /home/iitp/wiki

   mkdir inputdata

5. Copy the input file (input.txt) into inputdata folder on VM from local machine

   $ scp <path-on-local-system> <user-credentials-on-assigned-VM:path-to-copy-into-on-VM>

   $ scp /home/iitp/CS-555-Lab-2021/Lab-III/wiki/inputdata/input.txt iitp@172.16.27.166:/home/iitp/wiki/inputdata/

   paste the lines from the provided input file

6. Start all hadoop services
   a. Browse to hadoop installation sbin sub-directory

      cd /home/iitp/hadoop-2.6.0/sbin

   b. start all services

./start-all.sh

Note:- Enter password when prompted

7. Create input directories on HDFS
    a. browse to hadoop installation bin folder

        cd /home/iitp/hadoop-2.6.0/bin

    b. create directory (**wiki**)

        ./hadoop fs -mkdir /wiki

    c. create sub-directory **inputdata** within wiki directory

        ./hadoop fs -mkdir /wiki/inputdata

    d. create directory (**iprank**)

        ./hadoop fs -mkdir /wiki/iprank

    e. create directory (**prsort**)

        ./hadoop fs -mkdir /wiki/prsort

8. Copy the input text file from local directory to HDFS
    a. browse to hadoop installation bin folder

        cd /home/iitp/hadoop-2.6.0/bin

    b. Copy input file from Local-VM on to HDFS

        ./hadoop dfs -put /home/iitp/wiki/inputdata/input.txt /wiki/inputdata

9. Running the program
    a. browse to the bin directory of hadoop installation

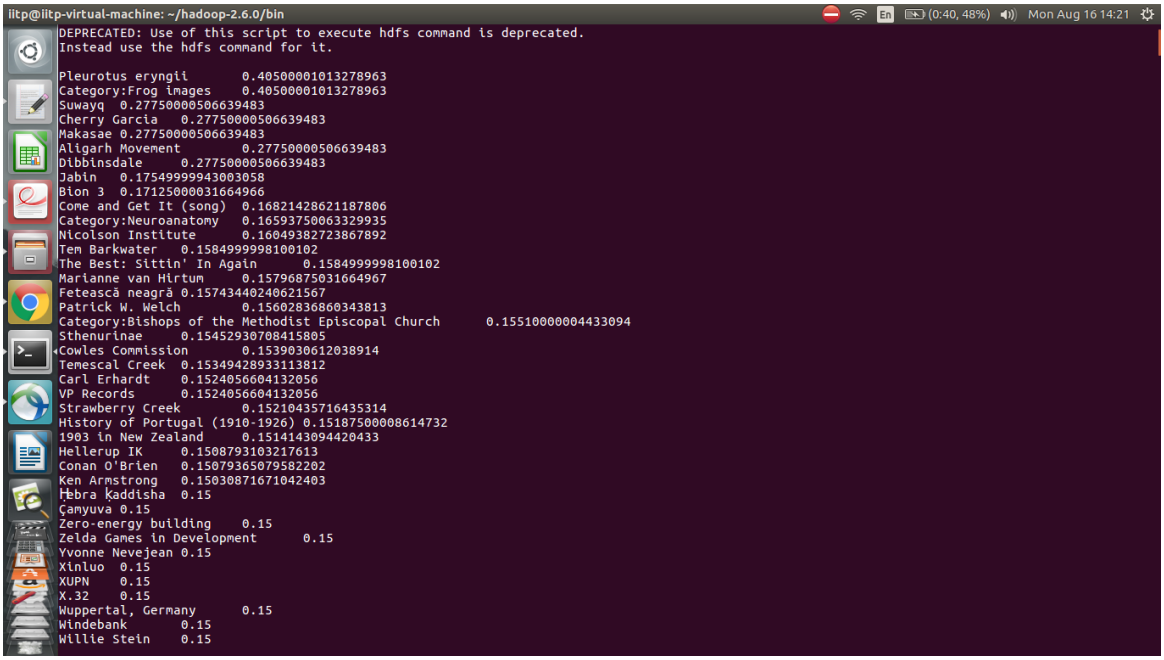
        cd /home/iitp/hadoop-2.6.0/bin

    b. Running in terminal

        ./hadoop jar /home/iitp/wiki/target/WP-1.0.0.jar DriverClass /wiki/inputdata /wiki/iprank/outputdata /wiki/outputdata /wiki/prsort/outputdata

    c. Finding outputs

        ./hadoop fs -cat /wiki/prsort/outputdata/part-r-00000


iitp@iitp-virtual-machine: ~/hadoop-2.6.0/bin
Mon Aug 16 14:21
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Pleurotus eryngii 0.40500001013278963
Category:Frog images 0.40500001013278963
Suwayq 0.27750000506639483
Cherry Garcia 0.27750000506639483
Makasae 0.27750000506639483
Aligarh Movement 0.27750000506639483
Dibbinsdale 0.27750000506639483
Jabin 0.17549999943003058
Bion 3 0.17125000031664966
Come and Get It (song) 0.16821428621187806
Category:Neuroanatomy 0.16593750063329935
Nicolson Institute 0.16049382723867892
Tem Barkwater 0.1584999998100102
The Best: Sittin' In Again 0.1584999998100102
Marianne van Hirtum 0.15796875031664967
Feteasca neagra 0.15743440240621567
Patrick W. Welch 0.15602836860343813
Category:Bishops of the Methodist Episcopal Church 0.15510000004433094
Sthenurinae 0.15452930708415805
Cowles Commission 0.1539030612038914
Temescal Creek 0.15349428933113812
Carl Erhardt 0.1524056604132056
VP Records 0.1524056604132056
Strawberry Creek 0.15210435716435314
History of Portugal (1910-1926) 0.15187500008614732
1903 in New Zealand 0.1514143094420433
Hellerup IK 0.1508793103217613
Conan O'Brien 0.15079365079582202
Ken Armstrong 0.15030871671042403
Hebra kaddisha 0.15
Çamyuva 0.15
Zero-energy building 0.15
Zelda Games in Development 0.15
Yvonne Nevejean 0.15
Xinluo 0.15
XUPN 0.15
X.32 0.15
Wuppertal, Germany 0.15
Windebank 0.15
Willie Stein 0.15