UE19CS322 Big Data Assignment 0

Word Count using MapReduce

This is the zero'th assignment for the UE19CS322 Big Data Course at PES University. The assignment consists of a single task and focuses on running a MapReduce job to count the number of words in a text file. This is an **ungraded but mandatory** assignment used to test your installation of Hadoop and does not involve writing any code.

The link to all the files required for the assignment can be found here.

Assignment Objectives and Outcomes

1. This assignment will help the student install and setup Hadoop.
2. At the end of this assignment, the student will be able to view their VM configuration and setup Hadoop for further assignments.

Ethical practices

Although ungraded, this assignment is **mandatory, not advisory**. All students are expected to perform this assignment. Not only will this help you to verify your Hadoop installation but it also will **setup your team's submission profile for later assignments**. It additionally helps us confirm if your virtual machine configuration is suitable to run all the assignments.

Please do not run the script repeatedly. All requests are logged and **frequent requests in attempt to sabotage the system will result in your team being blacklisted** on the portal.

The Dataset

The dataset is the book "Alice in Wonderland" by Lewis Caroll obtained from Project Gutenberg

Software/Languages to be used:

1. Python `3.8.x`
2. Hadoop `v3.2.2` only

Marks

This assignment is **ungraded**

Tasks Overview:

1. Clone the repository and obtain all the required files
2. Run the script
3. Verify your installation on this portal.

Submission Date

16th September, 11:59 PM

Task Specifications

Problem Statement

Find the number of occurrences of a given word

Description

Find the number of occurrences of the word "alice"

Running the Script

1. Ensure that Hadoop is running

```
$HADOOP_HOME/sbin/start-all.sh
```

You should see the following running processes (in any order) when you run `jps`. **If you do not see all these processes, your Hadoop installation is incorrect**

```
DataNode
SecondaryNameNode
Jps
ResourceManager
NameNode
NodeManager
```

2. Install `curl`

```
sudo apt install curl -y
```

3. Clone the repository and navigate into the directory

```
git clone https://github.com/Cloud-Computing-Big-Data/UE19CS322-A0
cd UE19CS322-A0/
```

4. Give the script executable access

```
chmod +x *.pyc
```

5. Run the file with **your Team ID as a command line argument**

```
python3 script.pyc BD_1_2_3_4
```

6. After execution of the complete script, you *should* see the following message on your terminal. Visit the portal and view the results of the execution.

```
Starting Hadoop Installation verification...
Verification concluded. Submission has been made to the portal.
Please check the portal for the results.
```

Note that your results will take a few minutes to show up on the portal. Please be patient and do not submit repeatedly.