

READ-ME

Following are instruction to compile and execute mapreduce program

a. DocWordCount.java

Compile :

```
sudo javac -cp <path-hadoop-lib> <path-hadoop-mapreduce>  
DocWordCount.java -d build -Xlint
```

Example :

```
sudo javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/*  
DocWordCount.java -d build -Xlint
```

Build JAR :

```
sudo jar -cvf DocWordCount.jar -<path-to-store build jar>  
Example : $ sudo jar -cvf DocWordCount.jar -C build/ .
```

Run:

```
sudo hadoop jar DocWordCount.jar org.myorg.DocWordCount <path of input  
files> <path to store output files>
```

Example: \$ sudo hadoop jar DocWordCount.jar org.myorg.DocWordCount
/user/cloudera/wordcount/input /user/cloudera/wordcount/output

b. TermFrequency.java

Compile :

```
sudo javac -cp <path-hadoop-lib> <path-hadoop-mapreduce>  
TermFrequency.java -d build -Xlint
```

Example :

```
sudo javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/*  
TermFrequency.java -d build -Xlint
```

Build JAR :

```
sudo jar -cvf TermFrequency.jar -<path-to-store build jar>  
Example : $ sudo jar -cvf TermFrequency.jar -C build/ .
```

Run:

```
sudo hadoop jar TermFrequency.jar org.myorg.TermFrequency <path of input  
files> <path to store output files>
```

Example: \$ sudo hadoop jar TermFrequency.jar org.myorg.TermFrequency
/user/cloudera/wordcount/input /user/cloudera/wordcount/output

c. TFIDF.java

Compile :

```
sudo javac -cp <path-hadoop-lib> <path-hadoop-mapreduce> TFIDF.java -d  
build -Xlint
```

Example :

```
sudo javac -cp /usr/lib/hadoop/*:/usr/lib/hadoop-mapreduce/*  
TFIDF.java -d build -Xlint
```

Build JAR :

```
sudo jar -cvf TFIDF.jar -<path-to-store build jar>
```

Example : \$ sudo jar -cvf TFIDF.jar -C build/ .

Run:

```
sudo hadoop jar TFIDF.jar org.myorg.TFIDF <path of input files> <path to  
store output files> <path to store output file of second map reduce>
```

Example: \$ sudo hadoop jar TFIDF.jar org.myorg.TFIDF
/user/cloudera/wordcount/input /user/cloudera/wordcount/output
/user.cloudera/wordcount/output1

Additional Instruction :

1. Delete all the output directories before executing any above jar.
2. TFIDF - program requires three input directories paths as parameter to execute
 - a. Input directory for first mapreduce.
 - b. Output directory for first mapreduce.
 - c. Output directory for second mapreduce.