Exercise Sheet 06
Distributed Data Analytics
Syed Khalid Ahmed

Exercise 01: Data cleaning and text tokenization:

Flow of the program:

The program takes a single file as input, performs operations to perform data cleaning and generates an output file. This output file will be used in Exercise # 02 as an input. I have saved the output file in directories that follow a naming convention (Book-1, Book-2 , …) because the program of Exercise # 02 first reads the URL of the file in order to identify from which file the words are coming from. By using the directory names, I can identify the file.

Mapper:

In the mapper function, I am reading the file line by line and outputting it to the reducer for further processing. I am stripping the line and checking if the line is a blank space or not. If it is not then output it to the reducer.

Reducer:

In the Reducer, I have made a FileCleanser() function that removes all the punctuation marks and numbers. It also removes all the common stop words that are found in the English literature. I have created a set which contains all the possible stop words. The program uses regular expressions to remove all the punctuations and numbers. It then checks whether a word appears in the stopwords's set or not. If it is not a stopword then it is merged with the string.

Hadoop Commands Used:

Following commands were used.

To put the files in HDFS:

```
hadoop fs -put /home/khalid/Documents/*.txt /Exercise6-Data/
```

For Book#1:

```
hadoop jar /home/khalid/Downloads/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-*streaming*.jar -file /home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -mapper /home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -file /home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -reducer /home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -input /Exercise6-Data/1.txt -output /Exercise6-Solution/CleansedFiles/Book1/
```

For Book#2:

```
hadoop jar /home/khalid/Downloads/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-
*streaming*.jar -file /home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -mapper
/home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -file
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -reducer
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -input /Exercise6-Data/2.txt -
output /Exercise6-Solution/CleansedFiles/Book2/
```

For Book#3:

```
hadoop jar /home/khalid/Downloads/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-
*streaming*.jar -file /home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -mapper
/home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -file
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -reducer
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -input /Exercise6-Data/3.txt -
output /Exercise6-Solution/CleansedFiles/Book3/
```

For Book#4:

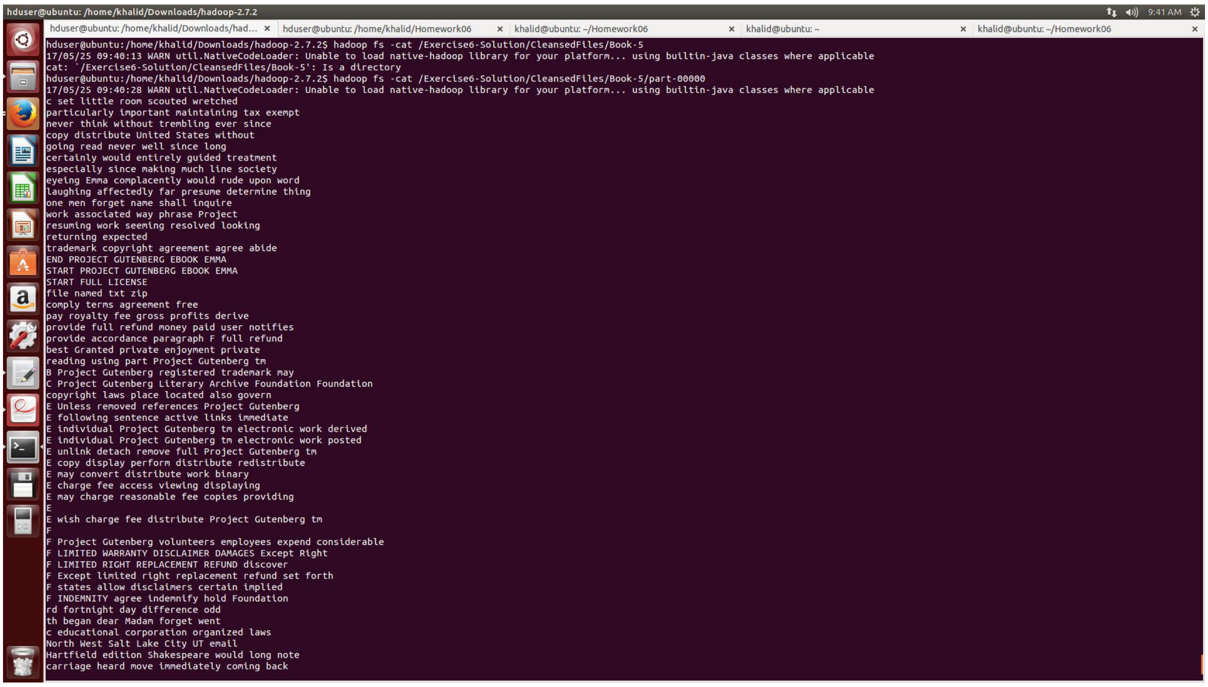
```
hadoop jar /home/khalid/Downloads/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-
*streaming*.jar -file /home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -mapper
/home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -file
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -reducer
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -input /Exercise6-Data/4.txt -
output /Exercise6-Solution/CleansedFiles/Book4/
```

For Book#5:

```
hadoop jar /home/khalid/Downloads/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-
*streaming*.jar -file /home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -mapper
/home/khalid/Homework06/Exercise01/FileCleansing-mapper.py -file
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -reducer
/home/khalid/Homework06/Exercise01/FileCleansing-reducer.py -input /Exercise6-Data/5.txt -
output /Exercise6-Solution/CleansedFiles/Book5/
```

Images:

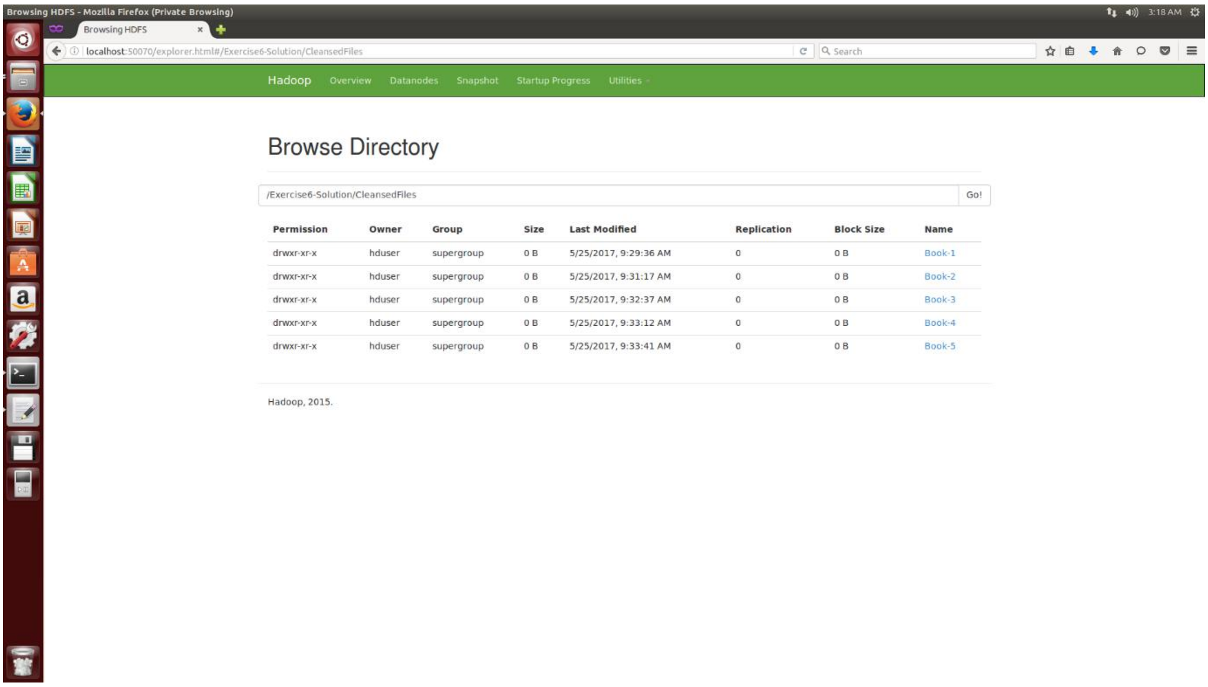
This is the output of the Hadoop cat command.



hduser@ubuntu: /home/khalid/Downloads/hadoop-2.7.2
9:41 AM
hduser@ubuntu: /home/khalid/Downloads/had...
hduser@ubuntu: /home/khalid/Homework06
khalid@ubuntu: ~/Homework06
khalid@ubuntu: ~
khalid@ubuntu: ~/Homework06

```
hduser@ubuntu:/home/khalid/Downloads/hadoop-2.7.2$ hadoop fs -cat /Exercise6-Solution/CleansedFiles/Book-5
17/05/25 09:40:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cat: `/Exercise6-Solution/CleansedFiles/Book-5': Is a directory
hduser@ubuntu:/home/khalid/Downloads/hadoop-2.7.2$ hadoop fs -cat /Exercise6-Solution/CleansedFiles/Book-5/part-00000
17/05/25 09:40:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
c set little room scouted wretched
particularly important maintaining tax exempt
never think without trembling ever since
copy distribute United States without
going read never well since long
certainly would entirely guided treatment
especially since making much line society
eyeing Emma complacently would rude upon word
laughing affectedly far presume determine thing
one men forget name shall inquire
work associated way phrase Project
resuming work seeming resolved looking
returning expected
trademark copyright agreement agree abide
END PROJECT GUTENBERG EBOOK EMMA
START PROJECT GUTENBERG EBOOK EMMA
START FULL LICENSE
file named txt zip
comply terms agreement free
pay royalty fee gross profits derive
provide full refund money paid user notifies
provide accordance paragraph F full refund
best Granted private enjoyment private
reading using part Project Gutenberg tm
B Project Gutenberg registered trademark may
C Project Gutenberg Literary Archive Foundation Foundation
copyright laws place located also govern
E Unless removed references Project Gutenberg
E following sentence active links immediate
E individual Project Gutenberg tm electronic work derived
E individual Project Gutenberg tm electronic work posted
E unlink detach remove full Project Gutenberg tm
E copy display perform distribute redistribute
E may convert distribute work binary
E charge fee access viewing displaying
E may charge reasonable fee copies providing
E
E wish charge fee distribute Project Gutenberg tm
F
F Project Gutenberg volunteers employees expend considerable
F LIMITED WARRANTY DISCLAIMER DAMAGES Except Right
F LIMITED RIGHT REPLACEMENT REFUND discover
F Except limited right replacement refund set forth
F states allow disclaimers certain implied
F INDEMNITY agree indemnify hold Foundation
rd fortnight day difference odd
th began dear Madam forget went
c educational corporation organized laws
North West Salt Lake City UT email
Hartfield edition Shakespeare would long note
carriage heard move immediately coming back
```

Directory for each book



Browsing HDFS - Mozilla Firefox (Private Browsing)
3:18 AM
Browsing HDFS
localhost:50070/explorer.html#/Exercise6-Solution/CleansedFiles
Search
Hadoop
Overview
Datanodes
Snapshot
Startup Progress
Utilities

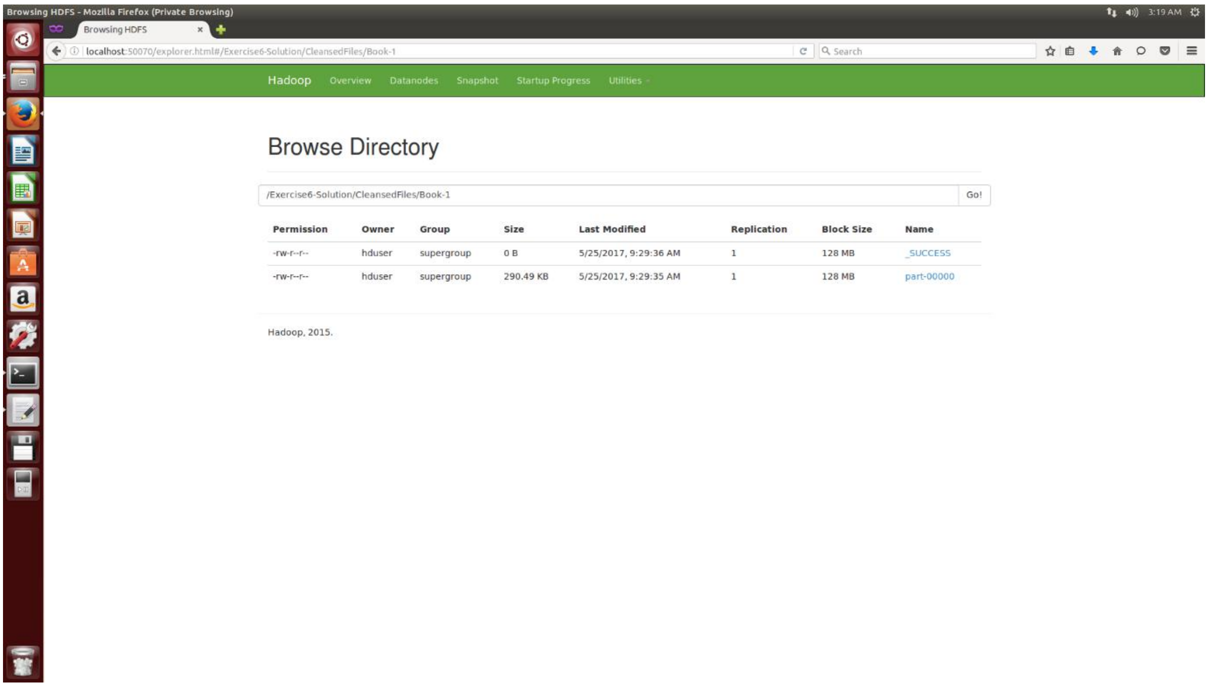
Browse Directory

/Exercise6-Solution/CleansedFiles

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hduser	supergroup	0 B	5/25/2017, 9:29:36 AM	0	0 B	Book-1
drwxr-xr-x	hduser	supergroup	0 B	5/25/2017, 9:31:17 AM	0	0 B	Book-2
drwxr-xr-x	hduser	supergroup	0 B	5/25/2017, 9:32:37 AM	0	0 B	Book-3
drwxr-xr-x	hduser	supergroup	0 B	5/25/2017, 9:33:12 AM	0	0 B	Book-4
drwxr-xr-x	hduser	supergroup	0 B	5/25/2017, 9:33:41 AM	0	0 B	Book-5

Hadoop, 2015.

Each Book directory has part-00000 as the output file.


Browsing HDFS - Mozilla Firefox (Private Browsing)
3:19 AM
Browsing HDFS
localhost:50070/explorer.html#/Exercise6-Solution/CleansedFiles/Book-1
Search
Hadoop
Overview
Datanodes
Snapshot
Startup Progress
Utilities

Browse Directory

/Exercise6-Solution/CleansedFiles/Book-1 Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hduser	supergroup	0 B	5/25/2017, 9:29:36 AM	1	128 MB	_SUCCESS
-rw-r--r--	hduser	supergroup	290.49 KB	5/25/2017, 9:29:35 AM	1	128 MB	part-00000

Hadoop, 2015.

Exercise 02: TFIDF scores of words:

Flow of the program:

The program takes the all of the input files from the previous program, performs operations to compute TF , IDF and TF*IDF scores and generates an output file.

Mapper:

In the mapper function, I am reading the files line by line and outputting it to the reducer for further processing. Since I have to find the id of the file from which the word is coming, so I have used:

```
input_file = os.environ['map_input_file']
```

to get the URL of the file that is currently being used by the mapper. It the returns a URL string from which I extract the directory name, perform a split and then use the number to identify the file. For example, my directory name is Book-2. I can extract 2 from this directory name which will tell me that this is file # 2.

The program then reads line from the file and generates and generates a key/value pair in which key is the word and value is the file id.

Reducer:

In the Reducer, the first function Compute() function counts the number of words in each file and the word counts for a specific file. The second function TFIDF() uses the information obtained in the first function to calculate the TF, IDF and TF*IDF scores for each word of a particular file.

The output consists of four columns for word, TF, IDF and TF*IDF scores respectively.

Hadoop Commands Used:

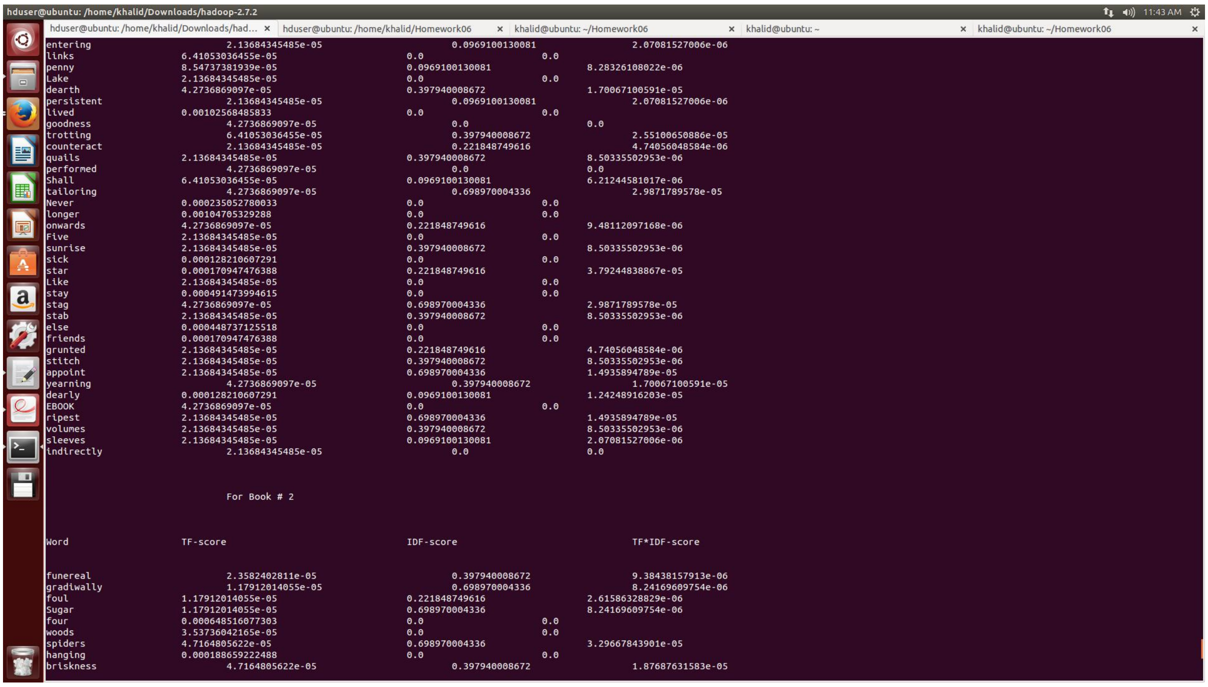
Following command was used

```
hadoop jar /home/khalid/Downloads/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-*streaming*.jar -file /home/khalid/Homework06/Exercise02/TF-IDF-mapper.py -mapper /home/khalid/Homework06/Exercise02/TF-IDF-mapper.py -file /home/khalid/Homework06/Exercise02/TF-IDF-reducer.py -reducer /home/khalid/Homework06/Exercise02/TF-IDF-reducer.py -input /Exercise6-Solution/CleansedFiles/Book-*/part-00000 -output /Exercise6-Solutions/Output/
```

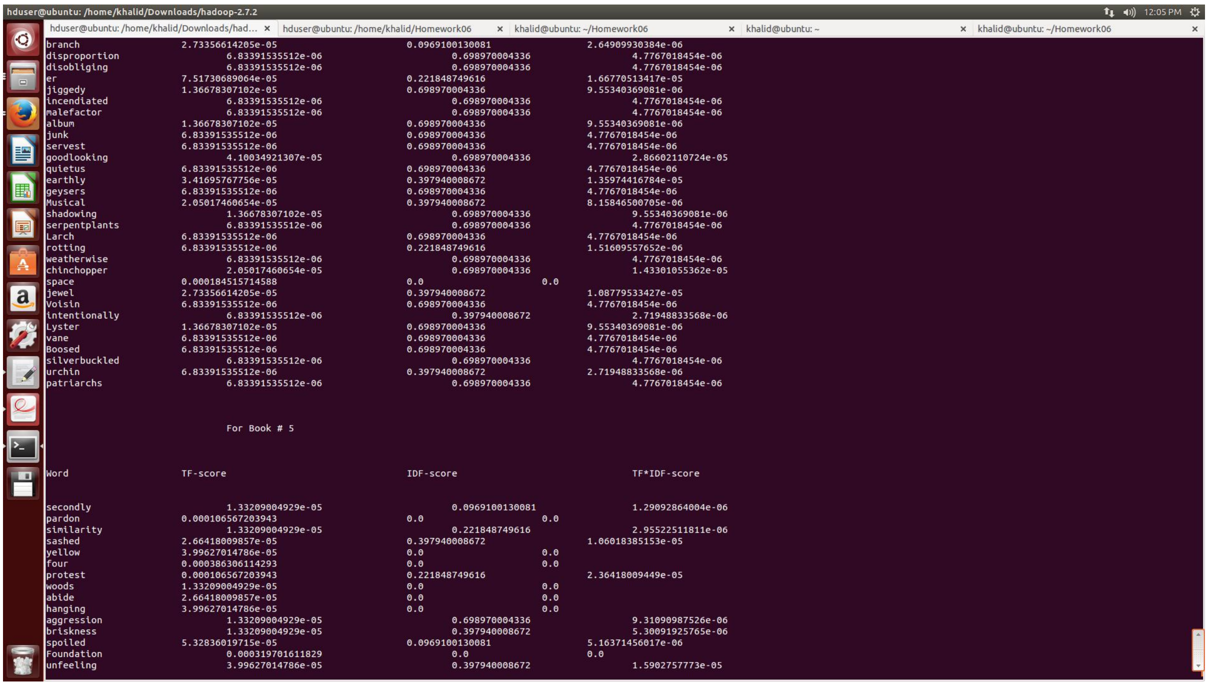
In the above command, Book-* means that it will get all the directories that match this wildcard and output their part-00000 files.

Images:

Output of the program


hduser@ubuntu: /home/khalid/Downloads/hadoop-2.7.2
11:43 AM
hduser@ubuntu: /home/khalid/Downloads/had...
hduser@ubuntu: /home/khalid/Homework06
khalid@ubuntu: ~/Homework06
khalid@ubuntu: ~
khalid@ubuntu: ~/Homework06

```
entering                        2.13684345485e-05                       0.0969100130081                 2.07081527006e-06
links                   6.41053036455e-05                       0.0                     0.0
penny                   8.54737381939e-05                       0.0969100130081                 8.28326108022e-06
Lake                    2.13684345485e-05                       0.0                     0.0
dearth                  4.2736869097e-05                        0.397940008672                  1.70067100591e-05
persistent                      2.13684345485e-05                       0.0969100130081                 2.07081527006e-06
lived                   0.00102568485833                        0.0                     0.0
goodness                        4.2736869097e-05                        0.0                     0.0
trotting                        6.41053036455e-05                       0.397940008672                  2.55100650886e-05
counteract                      2.13684345485e-05                       0.221848749616                  4.74056048584e-06
quails                  2.13684345485e-05                       0.397940008672                  8.50335502953e-06
performed                       4.2736869097e-05                        0.0                     0.0
Shall                   6.41053036455e-05                       0.0969100130081                 6.21244581017e-06
tailoring                       4.2736869097e-05                        0.698970004336                  2.9871789578e-05
Never                   0.000235052780033                       0.0                     0.0
longer                  0.00104705329288                        0.0                     0.0
onwards                 4.2736869097e-05                        0.221848749616                  9.48112097168e-06
Five                    2.13684345485e-05                       0.0                     0.0
sunrise                 2.13684345485e-05                       0.397940008672                  8.50335502953e-06
sick                    0.000128210607291                       0.0                     0.0
star                    0.000170947476388                       0.221848749616                  3.79244838867e-05
Like                    2.13684345485e-05                       0.0                     0.0
stay                    0.000491473994615                       0.0                     0.0
stag                    4.2736869097e-05                        0.698970004336                  2.9871789578e-05
stab                    2.13684345485e-05                       0.397940008672                  8.50335502953e-06
else                    0.000448737125518                       0.0                     0.0
friends                 0.000170947476388                       0.0                     0.0
grunted                 2.13684345485e-05                       0.221848749616                  4.74056048584e-06
stitch                  2.13684345485e-05                       0.397940008672                  8.50335502953e-06
appoint                 2.13684345485e-05                       0.698970004336                  1.4935894789e-05
yearning                        4.2736869097e-05                        0.397940008672                  1.70067100591e-05
dearly                  0.000128210607291                       0.0969100130081                 1.24248916203e-05
EBOOK                   4.2736869097e-05                        0.0                     0.0
ripest                  2.13684345485e-05                       0.698970004336                  1.4935894789e-05
volumes                 2.13684345485e-05                       0.397940008672                  8.50335502953e-06
sleeves                 2.13684345485e-05                       0.0969100130081                 2.07081527006e-06
indirectly                      2.13684345485e-05                       0.0                     0.0



                                For Book # 2



Word                    TF-score                                IDF-score                               TF*IDF-score


funereal                        2.3582402811e-05                        0.397940008672                  9.38438157913e-06
gradiwally                      1.17912014055e-05                       0.698970004336                  8.24169609754e-06
foul                    1.17912014055e-05                       0.221848749616                  2.61586328829e-06
Sugar                   1.17912014055e-05                       0.698970004336                  8.24169609754e-06
four                    0.000648516077303                       0.0                     0.0
woods                   3.53736042165e-05                       0.0                     0.0
spiders                 4.7164805622e-05                        0.698970004336                  3.29667843901e-05
hanging                 0.000188659222488                       0.0                     0.0
briskness                       4.7164805622e-05                        0.397940008672                  1.87687631583e-05
```



hduser@ubuntu: /home/khalid/Downloads/hadoop-2.7.2
12:05 PM
hduser@ubuntu: /home/khalid/Downloads/had...
hduser@ubuntu: /home/khalid/Homework06
khalid@ubuntu: ~/Homework06
khalid@ubuntu: ~
khalid@ubuntu: ~/Homework06

```
branch                  2.73356614205e-05                       0.0969100130081                 2.64909930384e-06
disproportion                   6.83391535512e-06                       0.698970004336                  4.7767018454e-06
disobliging                     6.83391535512e-06                       0.698970004336                  4.7767018454e-06
er                      7.51730689064e-05                       0.221848749616                  1.66770513417e-05
jiggedy                 1.36678307102e-05                       0.698970004336                  9.55340369081e-06
incendiated                     6.83391535512e-06                       0.698970004336                  4.7767018454e-06
malefactor                      6.83391535512e-06                       0.698970004336                  4.7767018454e-06
album                   1.36678307102e-05                       0.698970004336                  9.55340369081e-06
junk                    6.83391535512e-06                       0.698970004336                  4.7767018454e-06
servest                 6.83391535512e-06                       0.698970004336                  4.7767018454e-06
goodlooking                     4.10034921307e-05                       0.698970004336                  2.86602110724e-05
quietus                 6.83391535512e-06                       0.698970004336                  4.7767018454e-06
earthly                 3.41695767756e-05                       0.397940008672                  1.35974416784e-05
geysers                 6.83391535512e-06                       0.698970004336                  4.7767018454e-06
Musical                 2.05017460654e-05                       0.397940008672                  8.15846500705e-06
shadowing                       1.36678307102e-05                       0.698970004336                  9.55340369081e-06
serpentplants                   6.83391535512e-06                       0.698970004336                  4.7767018454e-06
Larch                   6.83391535512e-06                       0.698970004336                  4.7767018454e-06
rotting                 6.83391535512e-06                       0.221848749616                  1.51609557652e-06
weatherwise                     6.83391535512e-06                       0.698970004336                  4.7767018454e-06
chinchopper                     2.05017460654e-05                       0.698970004336                  1.43301055362e-05
space                   0.000184515714588                       0.0                     0.0
jewel                   2.73356614205e-05                       0.397940008672                  1.08779533427e-05
Voisin                  6.83391535512e-06                       0.698970004336                  4.7767018454e-06
intentionally                   6.83391535512e-06                       0.397940008672                  2.71948833568e-06
Lyster                  1.36678307102e-05                       0.698970004336                  9.55340369081e-06
vane                    6.83391535512e-06                       0.698970004336                  4.7767018454e-06
Boosed                  6.83391535512e-06                       0.698970004336                  4.7767018454e-06
silverbuckled                   6.83391535512e-06                       0.698970004336                  4.7767018454e-06
urchin                  6.83391535512e-06                       0.397940008672                  2.71948833568e-06
patriarchs                      6.83391535512e-06                       0.698970004336                  4.7767018454e-06



                                For Book # 5



Word                    TF-score                                IDF-score                               TF*IDF-score


secondly                        1.33209004929e-05                       0.0969100130081                 1.29092864004e-06
pardon                  0.000106567203943                       0.0                     0.0
similarity                      1.33209004929e-05                       0.221848749616                  2.95522511811e-06
sashed                  2.66418009857e-05                       0.397940008672                  1.06018385153e-05
yellow                  3.99627014786e-05                       0.0                     0.0
four                    0.000386306114293                       0.0                     0.0
protest                 0.000106567203943                       0.221848749616                  2.36418009449e-05
woods                   1.33209004929e-05                       0.0                     0.0
abide                   2.66418009857e-05                       0.0                     0.0
hanging                 3.99627014786e-05                       0.0                     0.0
aggression                      1.33209004929e-05                       0.698970004336                  9.31090987526e-06
briskness                       1.33209004929e-05                       0.397940008672                  5.30091925765e-06
spoiled                 5.32836019715e-05                       0.0969100130081                 5.16371456017e-06
Foundation                      0.000319701611829                       0.0                     0.0
unfeeling                       3.99627014786e-05                       0.397940008672                  1.5902757773e-05
```