



دانشکده مهندسی کامپیوتر

دانشگاه اصفهان

## تکلیف سوم و امتیازی بازیابی اطلاعات

دکتر محمد مهدی رضاپور

مهر والسادات نوحی

۹۹۳۶۱۳۰۶۱

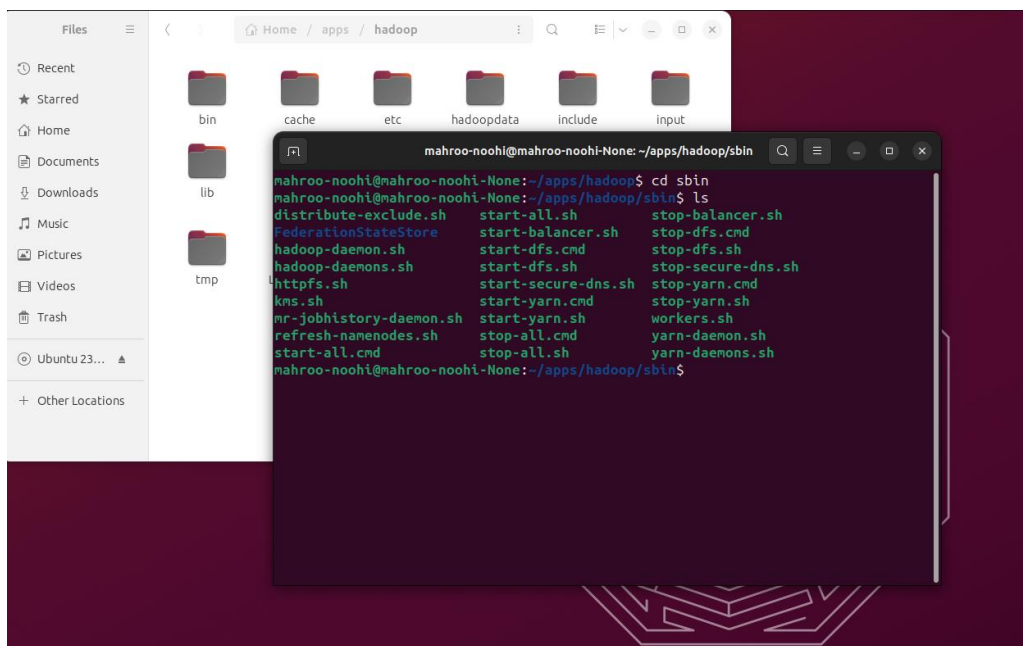
بهار ۱۴۰۳

برای ساخت جدول معکوس باید گام‌های زیر را انجام دهیم:

- ۱- نصب ماشین مجازی: از آنجایی که Hadoop برای سیستم عامل لینوکس طبق این [لینک](#) موجود بود مجبورا باید یک ماشین مجازی بالا آورده تا بتوانیم روی آن اجرا کنیم. از VmWare و همچنین نسخه ubuntu استفاده کردم.
- ۲- کانفیگ Hadoop: من از طریق این [لینک](#) سعی کردم تا Hadoop را نصب کنم. پس از اینکه با موفقیت کانفیگ و نصب گردید میتوانیم ورژن را مشاهده کنیم.

```
mahroo-noohi@mahroo-noohi-None: ~  
mahroo-noohi@mahroo-noohi-None:~$ hadoop version  
Hadoop 3.4.0  
Source code repository git@github.com:apache/hadoop.git -r bd8b77f398f626bb7791783192ee7a5dfaee760  
Compiled by root on 2024-03-04T06:35Z  
Compiled on platform linux-x86_64  
Compiled with protoc 3.21.12  
From source with checksum f7fe694a3613358b38812ae9c31114e  
This command was run using /home/mahroo-noohi/apps/hadoop/share/hadoop/common/hadoop-common-3.4.0.jar  
mahroo-noohi@mahroo-noohi-None:~$
```

همچنین برای اجرای این برنامه باید به صورت زیر عمل کنیم:

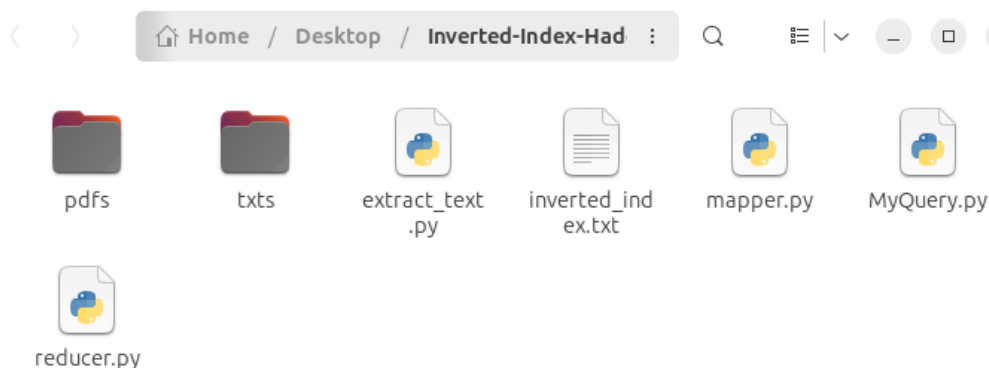


برای اجرا از دستور `./start-all.sh` استفاده میکنیم و برای غیر فعال کردن سرویس Hadoop از دستور `./stop-all.sh` استفاده میکنیم.

```
mahroo-noohi@mahroo-noohi-None:~/apps/hadoop/sbin$ ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as mahroo-noohi in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [mahroo-noohi-None]
Starting resourcemanager
Starting nodemanagers
mahroo-noohi@mahroo-noohi-None:~/apps/hadoop/sbin$ jps
13936 Jps
13145 SecondaryNameNode
12909 DataNode
12717 NameNode
13390 ResourceManager
mahroo-noohi@mahroo-noohi-None:~/apps/hadoop/sbin$
```

هر ۵ سرویس فعال و به درستی نصب شده است.

۳- ایجاد پروژه: در این گام شروع به ایجاد فایل‌های `extract_txt.py` و `reducer.py` و `mapper.py` میکنیم.



- در پوشه `pdfs` شامل ۲۰ تا مقاله ما هست که باید متن درون آن استخراج شود با اجرای فایل `extract_text.py` این کار را با کتابخانه `PyPdf2` انجام می‌دهیم. و خروجی هر فایل را متناسب با اسم فایل در پوشه `txts` قرار می‌دهیم. پس در این گام استخراج متن انجام شد.  
دستور:

`Python3 extract_text.py`

۴- ایجاد ایندکس معکوس با استفاده از پارادایم MapReduce و ابزار Hadoop Apache:

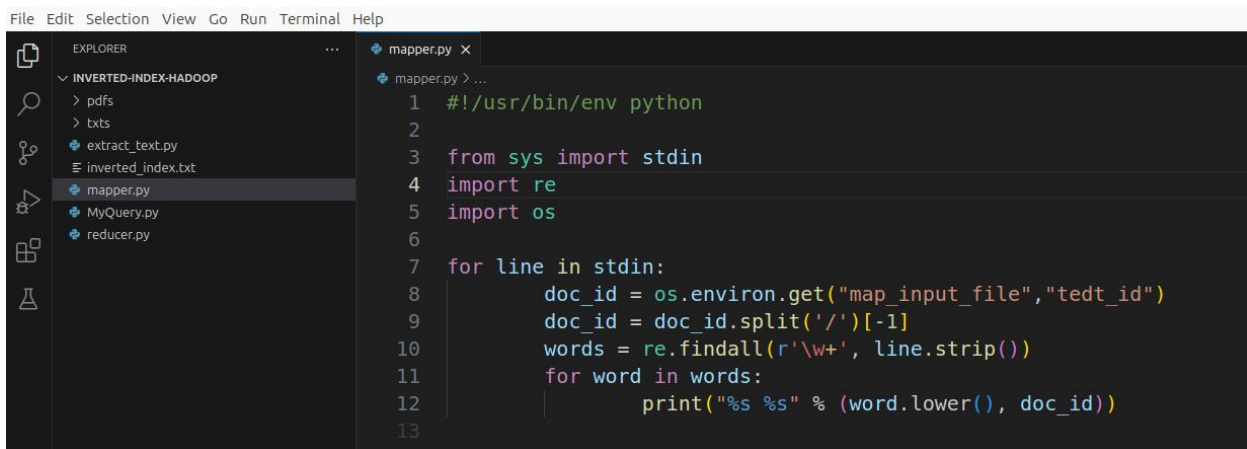
۴-۱- اول باید ۲۰ تا فایل با پسوند txt را در سیستم Hadoop مسیر دهی کنیم.

```
hadoop fs -mkdir -p /user/mahroo /pdfs
```

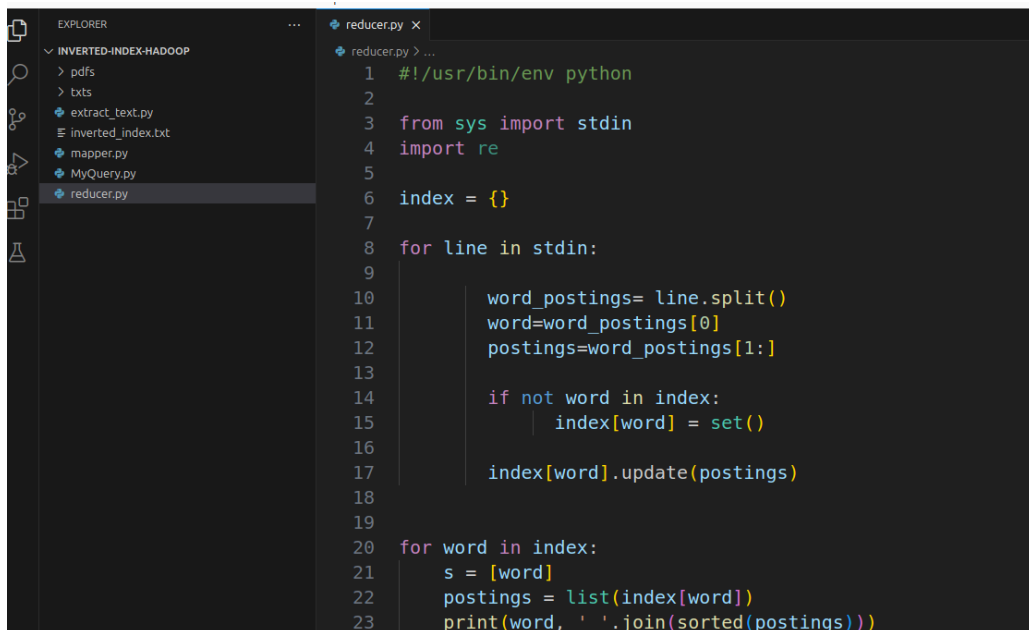
۴-۲- باید متن‌های استخراج شده را در این مسیری که ساختیم کپی کنیم.

```
hadoop fs -put /home /mahroo noohi  
/Desktop/Inverted_Index_Hadoop/txts/* /user/mahroo/pdfs
```

۵- اجرای کدهای mapper, reducer:



```
File Edit Selection View Go Run Terminal Help  
EXPLORER  
INVERTED-INDEX-HADOOP  
  pdfs  
  txts  
  extract_text.py  
  inverted_index.txt  
  mapper.py  
  MyQuery.py  
  reducer.py  
mapper.py x  
mapper.py > ...  
1  #!/usr/bin/env python  
2  
3  from sys import stdin  
4  import re  
5  import os  
6  
7  for line in stdin:  
8      doc_id = os.environ.get("map_input_file", "tedt_id")  
9      doc_id = doc_id.split('/')[1]  
10     words = re.findall(r'\w+', line.strip())  
11     for word in words:  
12         print("%s %s" % (word.lower(), doc_id))  
13
```



```
EXPLORER  
INVERTED-INDEX-HADOOP  
  pdfs  
  txts  
  extract_text.py  
  inverted_index.txt  
  mapper.py  
  MyQuery.py  
  reducer.py  
reducer.py x  
reducer.py > ...  
1  #!/usr/bin/env python  
2  
3  from sys import stdin  
4  import re  
5  
6  index = {}  
7  
8  for line in stdin:  
9  
10     word_postings= line.split()  
11     word=word_postings[0]  
12     postings=word_postings[1:]  
13  
14     if not word in index:  
15         index[word] = set()  
16  
17     index[word].update(postings)  
18  
19  
20 for word in index:  
21     s = [word]  
22     postings = list(index[word])  
23     print(word, ' '.join(sorted(postings)))
```

با دستور زیر میتوانیم با Hadoop اجرا کنیم:

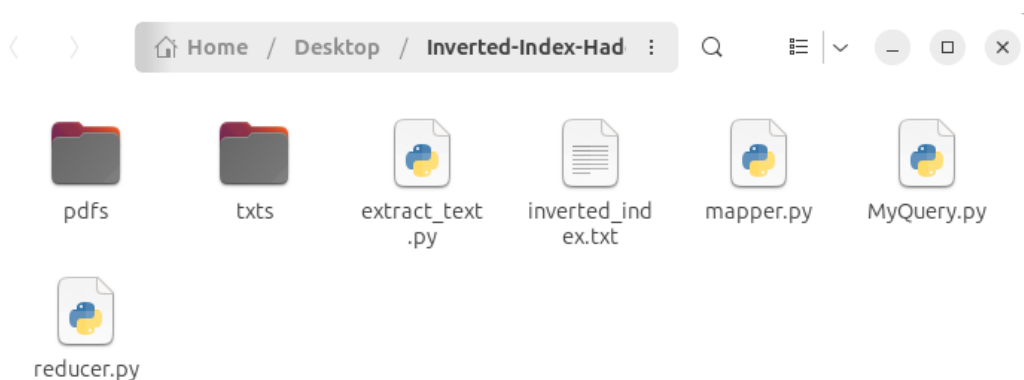
```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar  
-mapper "/home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/mapper.py"  
-reducer "/home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/reducer.py"  
-input "/user/mahroo/pdfs/*"  
-output "/user/mahroo/output"
```

در هنگام اجرا به صورت زیر است:

```
2024-04-17 20:18:40,553 INFO mapreduce.Job: Job job_1713372133609_0002 running in uber mode : false  
2024-04-17 20:18:40,554 INFO mapreduce.Job: map 0% reduce 0%  
2024-04-17 20:18:59,323 INFO mapreduce.Job: map 5% reduce 0%  
2024-04-17 20:19:00,355 INFO mapreduce.Job: map 30% reduce 0%  
2024-04-17 20:19:24,415 INFO mapreduce.Job: map 45% reduce 0%  
2024-04-17 20:19:25,428 INFO mapreduce.Job: map 60% reduce 0%  
2024-04-17 20:19:47,258 INFO mapreduce.Job: map 65% reduce 0%  
2024-04-17 20:19:49,336 INFO mapreduce.Job: map 80% reduce 0%  
2024-04-17 20:19:50,356 INFO mapreduce.Job: map 85% reduce 0%  
2024-04-17 20:19:58,536 INFO mapreduce.Job: map 85% reduce 28%  
2024-04-17 20:20:00,589 INFO mapreduce.Job: map 90% reduce 28%  
2024-04-17 20:20:01,605 INFO mapreduce.Job: map 100% reduce 28%  
2024-04-17 20:20:03,625 INFO mapreduce.Job: map 100% reduce 100%  
2024-04-17 20:20:04,648 INFO mapreduce.Job: Job job_1713372133609_0002 completed successfully
```

۶- ایندکس ساخته شده از فایل سیستم Hadoop به محل پروژه کپی می‌کنیم:

```
hadoop fs -copyToLocal /user/mahroo /pdfs-output/part-00000 /home  
/mahroo_noohi/Desktop/Inverted_Index_Hadoop/inverted_index.txt
```



۷- با داشتن فایل `inverted_index` می‌توان برنامه را تست کرد و کوئری‌های بولینی را وارد کرد. برای این کار الزم است تا فایل `MyQuery.py` را اجرا کنیم.

### Python3 MyQuery.py

بریم چند تا تست ببینیم:

کوئری ساده:

```
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: clear
Documents satisfying the query 'clear': ['3.txt', '2.txt', '6.txt', '17.txt', '10.txt']
○ mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$
```

```
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: mahroo
No documents found for the query 'mahroo'
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: goodbye
No documents found for the query 'goodbye'
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: goodbye
No documents found for the query 'goodby'
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: homework
['homework']
No documents found for the query 'homework'
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: method
Documents satisfying the query 'method': ['12.txt', '1.txt', '4.txt', '20.txt', '17.txt', '6.txt', '15.txt', '5.txt', '13.txt', '11.txt']
○ mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$
```

کوئری بولینی:

```
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: this
Documents satisfying the query 'this': ['13.txt', '1.txt', '19.txt', '17.txt', '7.txt', '18.txt', '15.txt', '6.txt', '10.txt', '12.txt', '14.txt', '3.txt', '9.txt', '11.txt', '20.txt', '4.txt', '8.txt', '2.txt', '5.txt', '16.txt']
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: good
Documents satisfying the query 'good': ['2.txt', '1.txt', '4.txt', '3.txt', '10.txt', '16.txt', '20.txt', '9.txt']
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: this not good
['this', 'not', 'good']
Documents satisfying the query 'this not good': ['13.txt', '19.txt', '7.txt', '17.txt', '6.txt', '18.txt', '12.txt', '15.txt', '11.txt', '5.txt', '14.txt', '8.txt']
○ mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$
```

```
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: hello
Documents satisfying the query 'hello': ['2.txt', '7.txt']
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: good
Documents satisfying the query 'good': ['1.txt', '4.txt', '3.txt', '9.txt', '20.txt', '16.txt', '2.txt', '10.txt']
● mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$ /bin/python3.12 /home/mahroo-noohi/Desktop/Inverted-Index-Hadoop/MyQuery.py
Enter a query: hello and good
['hello', 'and', 'good']
Documents satisfying the query 'hello and good': ['2.txt']
○ mahroo-noohi@mahroo-noohi-None:~/Desktop/Inverted-Index-Hadoop$
```

محتوای جدول معکوس به صورت زیر می‌باشد.

hecd 1.txt  
hector 19.txt  
heewoo 12.txt  
heidelberg 10.txt 20.txt 8.txt  
height 11.txt  
heightened 1.txt  
heins 16.txt  
heinshiener 20.txt  
heir 11.txt 3.txt  
heitfield 4.txt  
held 1.txt 12.txt 7.txt  
helens 11.txt  
heliocentric 20.txt  
helium 15.txt  
heliyon 6.txt  
helland 1.txt  
hellendoorn 16.txt  
hellenic 13.txt  
hellmann 14.txt  
hello 2.txt 7.txt  
helmut 13.txt  
help 1.txt 10.txt 11.txt 16.txt 2.txt 5.txt 7.txt  
helped 1.txt  
helpful 1.txt 10.txt 16.txt  
helping 16.txt  
helps 1.txt 10.txt 7.txt  
helvajian 20.txt  
hemmer 8.txt  
hence 1.txt 13.txt 14.txt 18.txt 3.txt 4.txt 9.txt

Q hello