

دانشکده مهندسی کامپیوتر

دانشگاه اصفهان

**تکلیف سوم و امتیازی بازیابی اطلاعات**

دکتر محمدمهدی رضاپور

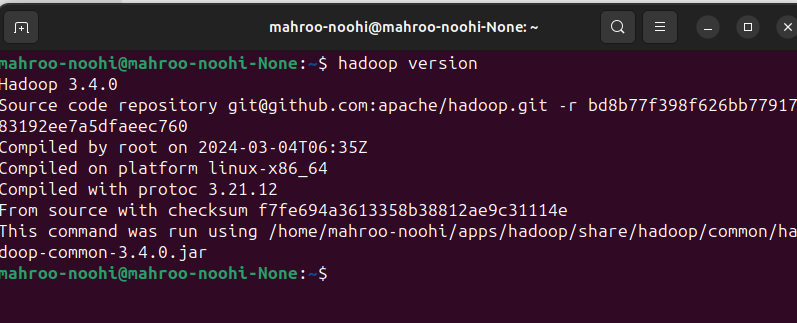
مهروالسادات نوحی

993613061

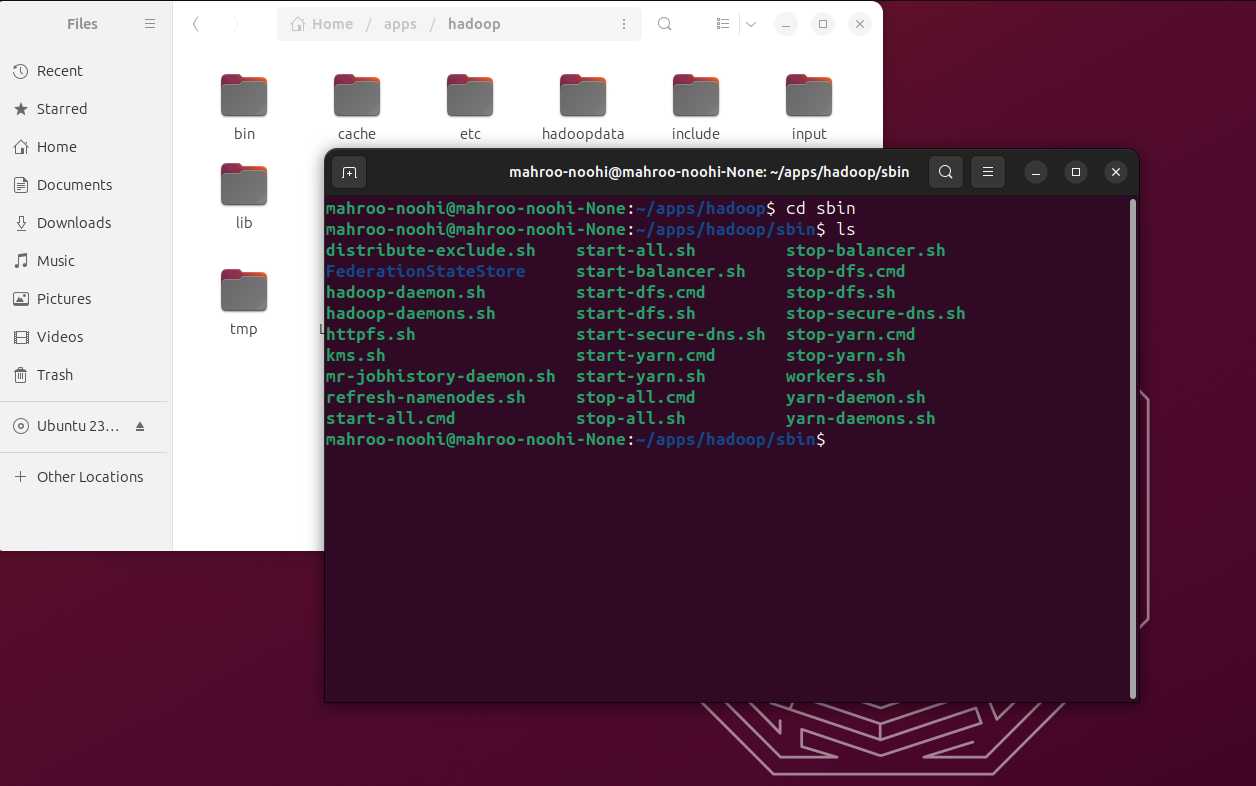
بهار 1403

برای ساخت جدول معکوس باید گام‌های زیر را انجام دهیم:

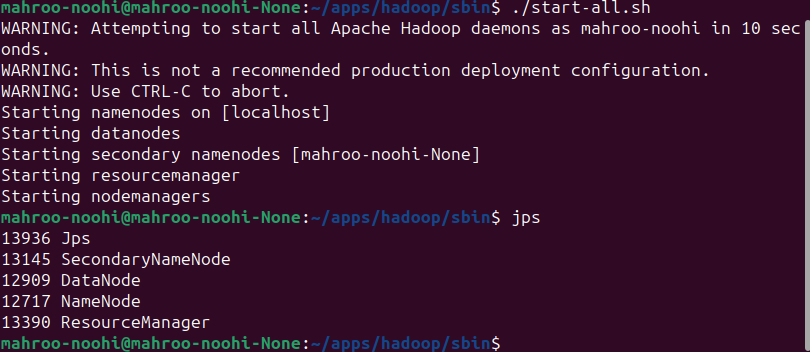
1. نصب ماشین مجازی: از آنجایی که Hadoopبرای سیستم عامل لینوکس طبق این [لینک](https://hadoop.apache.org/releases.html) موجود بود مجبورا باید یک ماشین مجازی بالا آورده تا بتوانیم روی آن اجرا کنیم. از VmWare و همچنین نسخه ubuntu استفاده کردم.
2. کانفیگ Hadoop: من از طریق این [لینک](https://www.baeldung.com/linux/hadoop-install-configure) سعی کردم تا Hadoop را نصب کنم. پس از اینکه با موفقیت کانفیگ و نصب گردید میتوانیم ورژن را مشاهده کنیم.



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

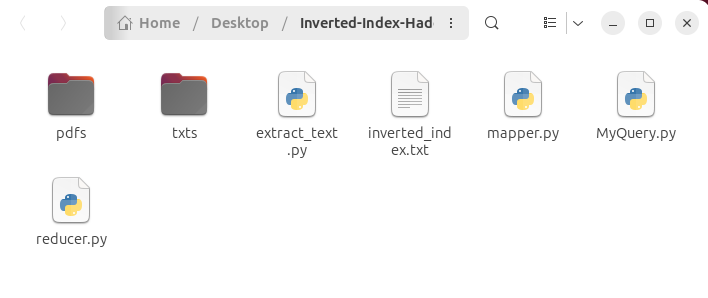


برای اجرا از دستور ./start-all.sh استفاده میکنیم و برای غیر فعال کردن سرویس Hadoop از دستور ./stop-all.sh استفاده میکنیم.



هر 5 سرویس فعال و به درستی نصب شده است.

1. ایجاد پروژه: در این گام شروع به ایجاد فایل‌های mapper.py و redeucer.py و extract\_txt.py میکنیم.



* در پوشه pdfs شامل 20 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

1. ایجاد ایندکس معکوس با استفاده از پارادایم MapReduce و ابزار Hadoop Apache:

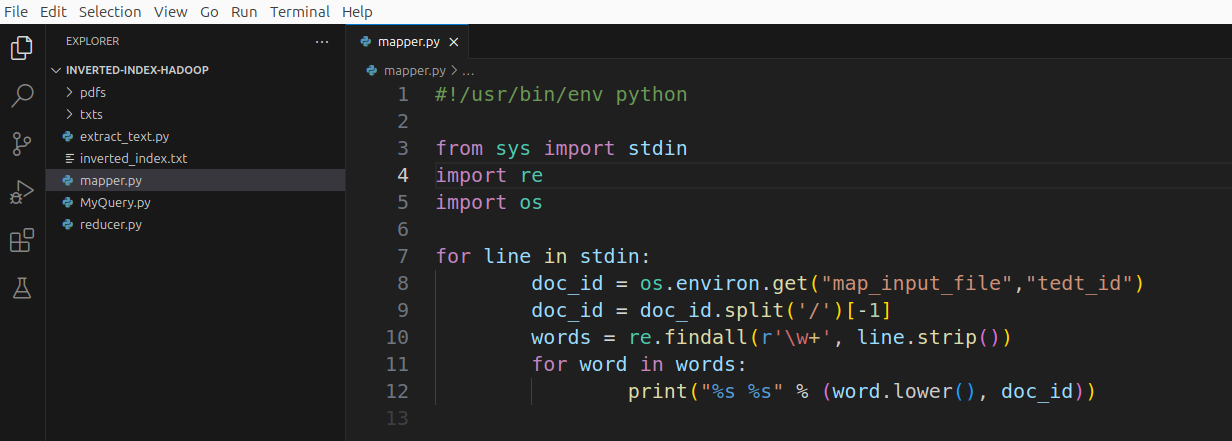
1-4- اول باید 20 تا فایل با پسوند txt را در سیستم Hadoop مسیر دهی کنیم.

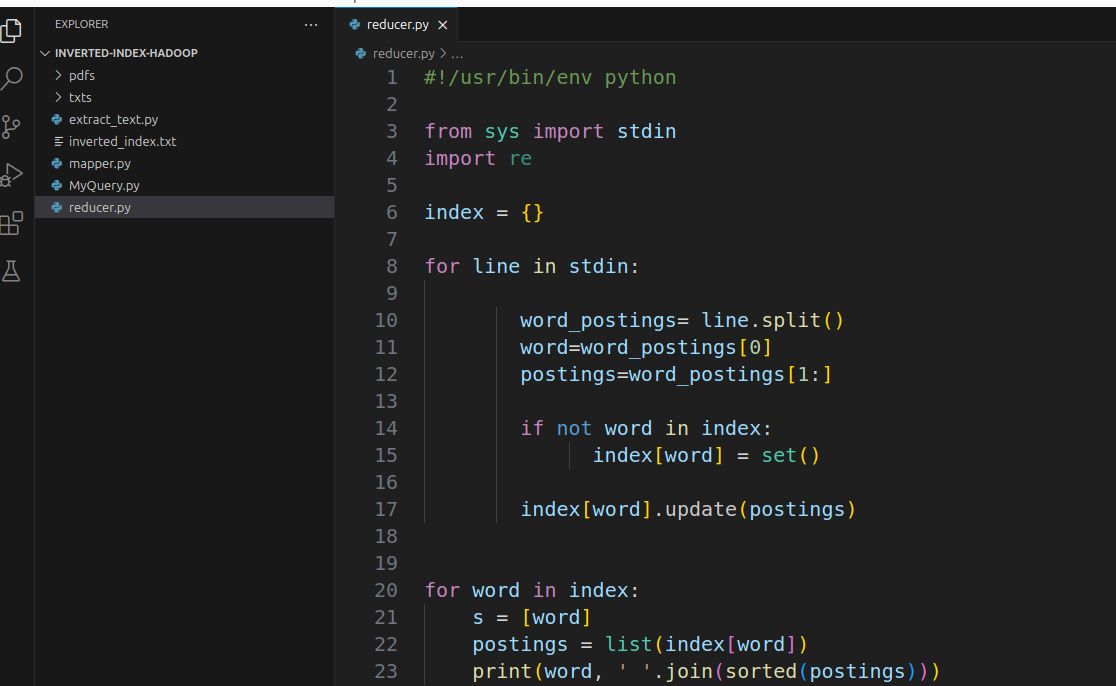
hadoop fs -mkdir -p /user/mahroo /pdfs

2-4- باید متن‌های استخراج شده را در این مسیری که ساختیم کپی کنیم.

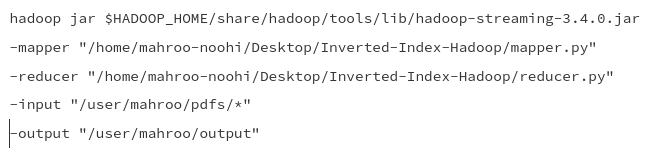
hadoop fs -put /home /mahroo\_noohi /Desktop/Inverted\_Index\_Hadoop/txts/\* /user/mahroo/pdfs

1. اجرای کدهای mapper ,reducer:

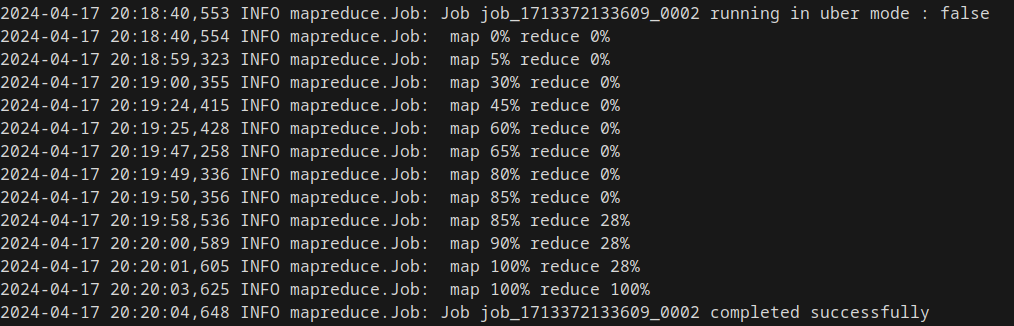




با دستور زیر میتوانیم با Hadoop اجرا کنیم:

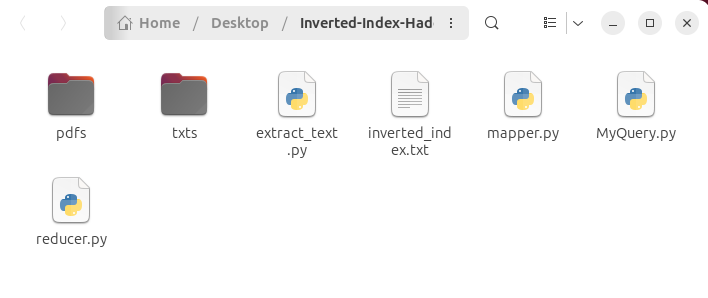


در هنگام اجرا به صورت زیر است:



1. 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

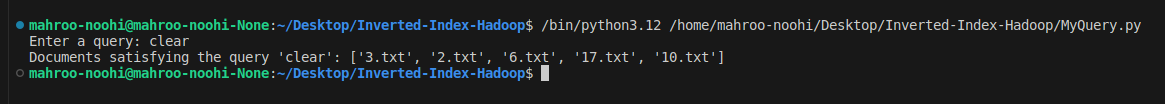


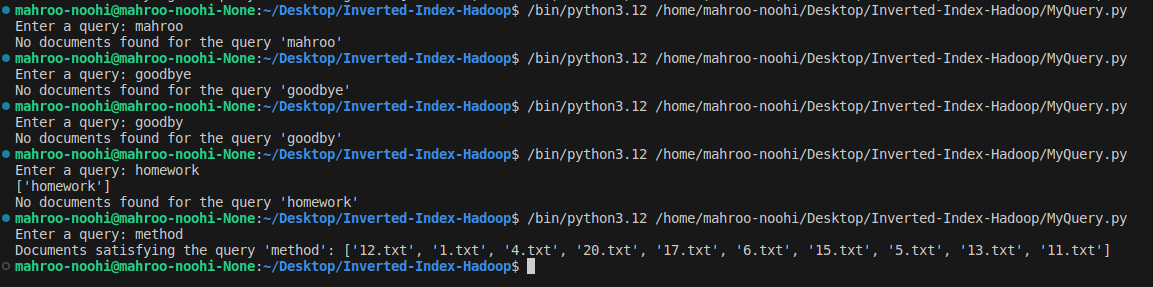
1. با داشتن فایلinverted\_index می‌توان برنامه را تست کرد و کوئری‌های بولینی را وارد کرد. برای این کار الزم است تا فایل MyQuery.py را اجرا کنیم.

Python3 MyQuery.py

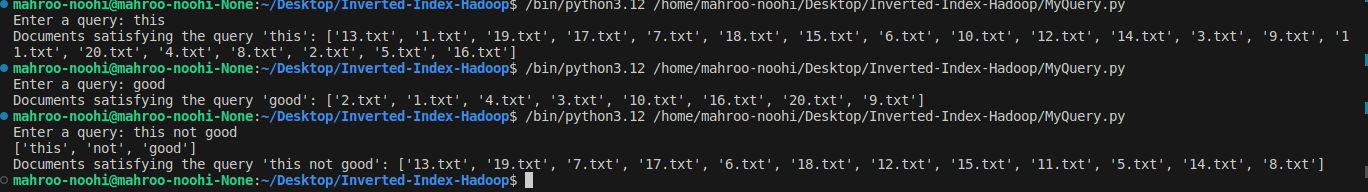
بریم چند تا تست ببیینم:

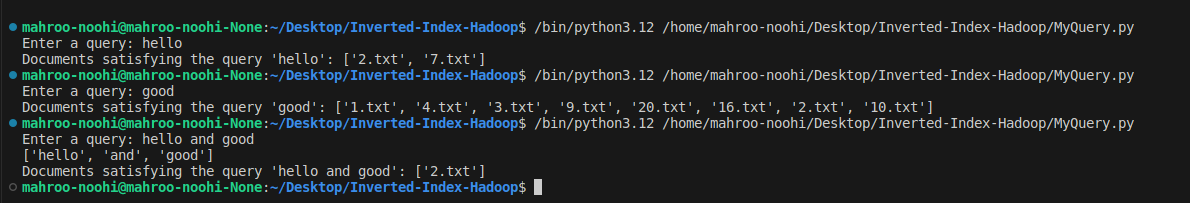
کوعری ساده:





کوعری بولینی:





محتوای جدول معکوس به صورت زیر می‌باشد.

