**Analyzing the web with Spark on your local machine**

**Made by Yunyan Wu**

**Part A: Configure your environment**

Please note that this handout only support Linux and Mac OS X at the moment. In addition, You need a github account to have access to clone all repositories needed.

1. Download and Install Docker For Mac following the link:

<https://docs.docker.com/docker-for-mac/install/>

1. Download and Install Git following the link:

<https://help.github.com/articles/set-up-git/>

1. Download and Install GitHub Desktop following the link:

<https://help.github.com/articles/set-up-git/>

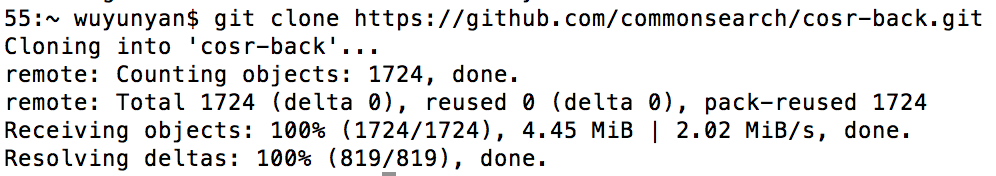
**Part B: Install cosr-back and cosr-ops repositories on your local machine**

1. Install cosr-back on your local machine
2. To clone this repository from GitHub, go to your local workspace directory and run:

git clone https://github.com/commonsearch/cosr-back.git

cd cosr-back

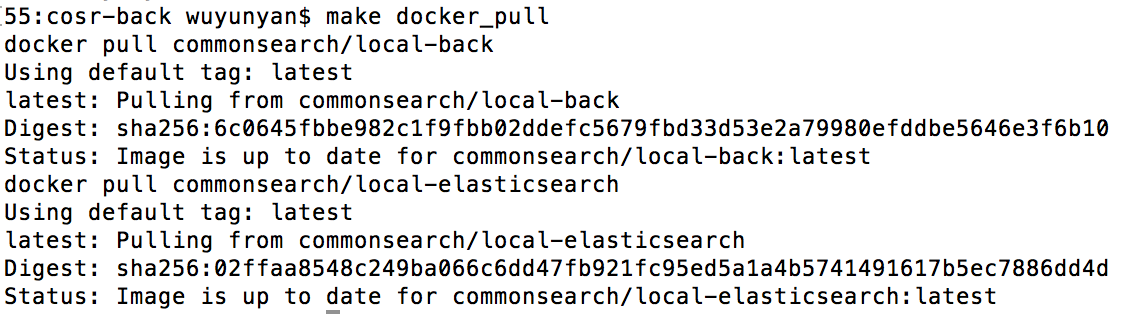
If succeed, you will get something like the following:



1. Next, there are a few Docker images available that contain all the dependencies you will need to run cosr-back. To download them from the Docker Hub, just run:

make docker\_pull

If succeed,you will get:



1. Install cosr-ops on your local machine
2. To clone this repository from GitHub, go to your local workspace directory and run:

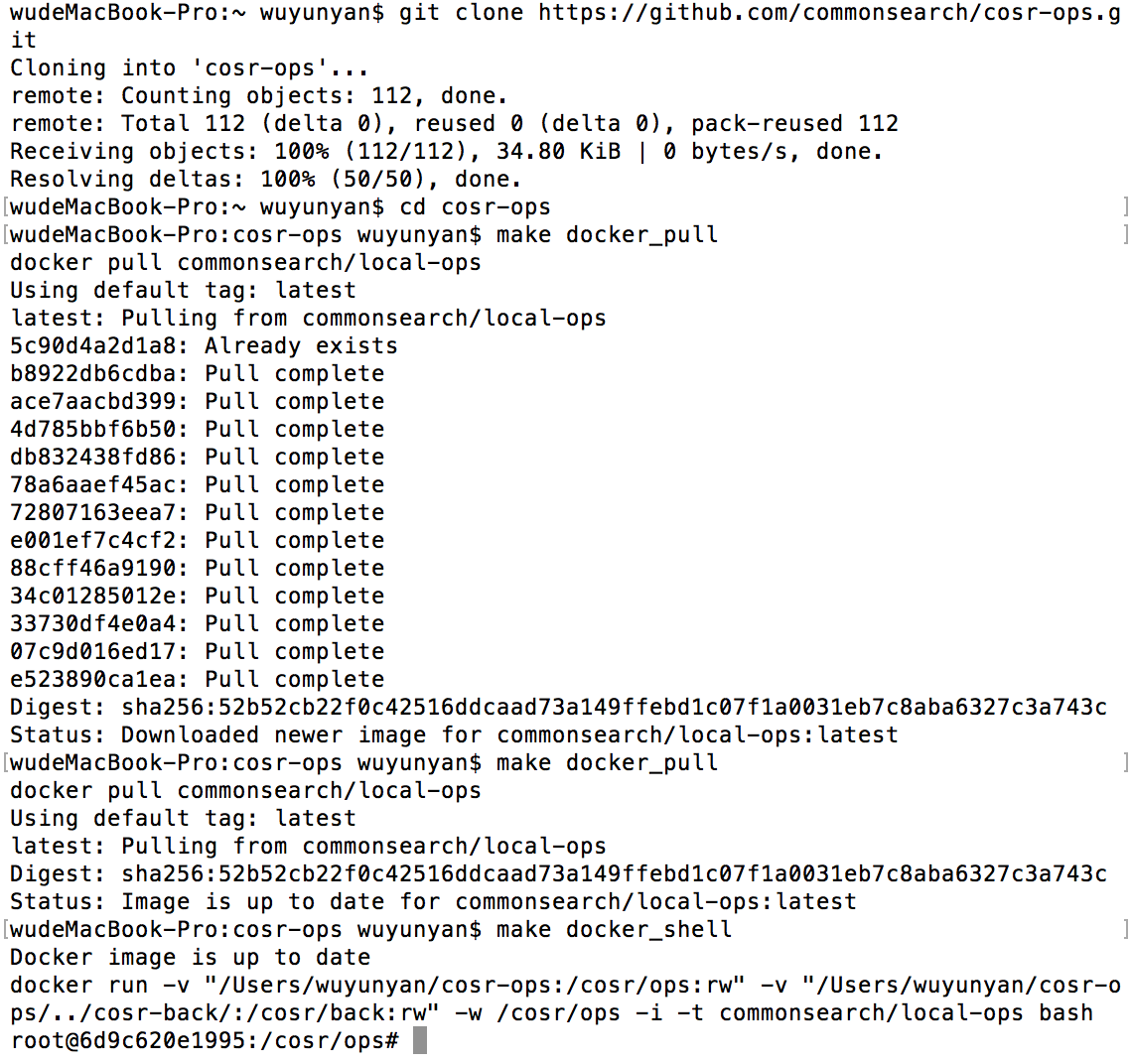
git clone https://github.com/commonsearch/cosr-ops.git

cd cosr-ops

2) Next, you should download our pre-built image from the Docker Hub:

make docker\_pull

If succeed,you will get:



1. You’re done!

To open a shell inside the Docker image with all the tools installed, just type:

make docker\_shell

**Part C: Do a test run on your local machine**

In this test, we will use one document source ([Common Crawl](https://www.commoncrawl.org/)) and two plugins (one to filter documents, and one to dump our list of backlinks) to collect all links from domain un.org contained key words: sicence and technology.

1. Open a console in the cosr-back by running:

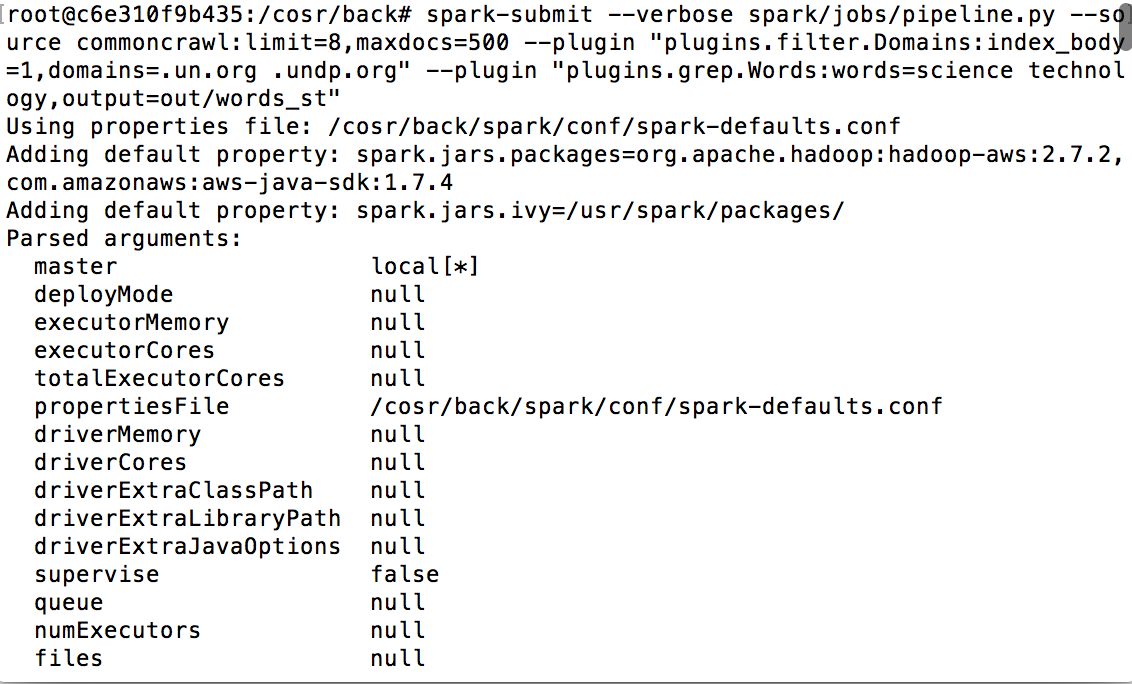
cd cosr-back

make docker\_shell

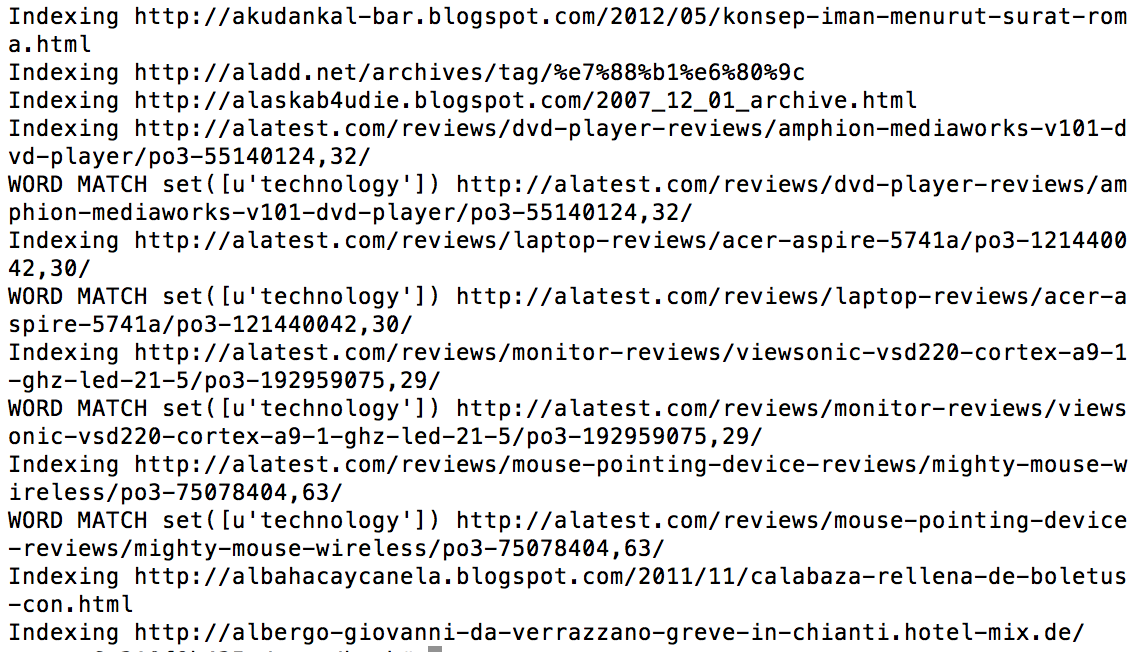
1. To build the command line for your job and run:

spark-submit --verbose spark/jobs/pipeline.py --source commoncrawl:limit=8,maxdocs=500 --plugin "plugins.filter.Domains:index\_body=1,domains=.[un.org](http://un.org/) .[undp.org](http://undp.org/)" --plugin "plugins.grep.Words:words=science technology,output=out/words\_st"

If succeed, you will get:



When finished, you can see:



This means Spark has filtered the links and got the ones we want. You can find them on your local machine like:



Also, you can change the command to suit your own goal to analyze the web.

If need, you can find more detailed instruction to build the command by your own from the following link on which we based to develop this handout:

<https://about.commonsearch.org/developer/tutorials/analyzing-the-web-with-spark-on-ec2>

However, if you want to handle large amounts of data, your local machine can not satisfy your need. That’s why we will connect to cloud service, such as AWS EC2, to deal with the big data problem in next steps.