Running Spark Jobs Locally

It's generally much easier to test your code locally (on a smaller data set, one assumes) before uploading to the Cluster. Fortunately, Spark makes that easy.

Local Spark Jobs: your computer (Linux, OSX)

This assumes a Linux-like environment. I believe these instructions work more-or-less the same on OSX.

From the Spark download page *(https://spark.apache.org/downloads.html)* , get Spark version 3.0.0 (which is the version we'll be using on the cluster), "Pre-built for Hadoop 2.7 and later", and click the "download Spark" link. Unpack that somewhere you like. Set a couple of environment variables so things start correctly. (This must be done each time you log in/create a new terminal.)

```
export PYSPARK_PYTHON=python3
export PATH=${PATH}:/home/you/spark-3.0.0-bin-hadoop2.7/bin
```

If you encounter Java version problems (often visible as `Py4JJavaError`), you may need to install Java or OpenJDK version 8 and ask Spark to use it:

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
```

On OSX, at least one student has found this to be the correct setting:

```
export JAVA_HOME="/Library/Internet Plug-Ins/JavaAppletPlugin.plugin/Contents/Home/"
```

Then you can start the pyspark shell or a standalone job:

```
pyspark
spark-submit sparkcode.py
```

While the job is running, you can access the web frontend at http://localhost:4040/ *(http://localhost:4040/)* .

If you're using the pyspark shell and want the IPython REPL instead of the plain Python REPL, you can set this environment variable:

```
export PYSPARK_DRIVER_PYTHON=ipython3
```

Local Spark Jobs: your computer with pip

In theory, Spark can be pip-installed:

```
pip3 install --user pyspark
```

… and then use the `pyspark` and `spark-submit` commands as described above.

I haven't had good luck with pip + pyspark in the past, but they may have updated their installer on the Spark side. Feedback appreciated.

You may also have to configure your Java version as with other Spark installs:

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
```

Local Spark Jobs: CSIL Linux

Spark is installed on the CSIL Linux workstations (to run in local-only mode). You need to specify a Java runtime and that you're using Python 3, but then the standard commands should work:

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PYSPARK_PYTHON=python3
pyspark
spark-submit sparkcode.py
```

While the job is running, you can access the web frontend at http://localhost:4040/ *(http://localhost:4040/)* .

If you're using the pyspark shell and want the IPython REPL instead of the plain Python REPL, you can set this environment variable:

```
export PYSPARK_DRIVER_PYTHON=ipython3
```

Updated Fri Aug. 28 2020, 11:21 by ggbaker.