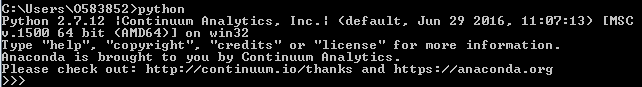
**Configuration Manual: Intraday Price Download and Bag of Words Model**

*Configuration: 64-bit Windows 7, Anaconda Python*

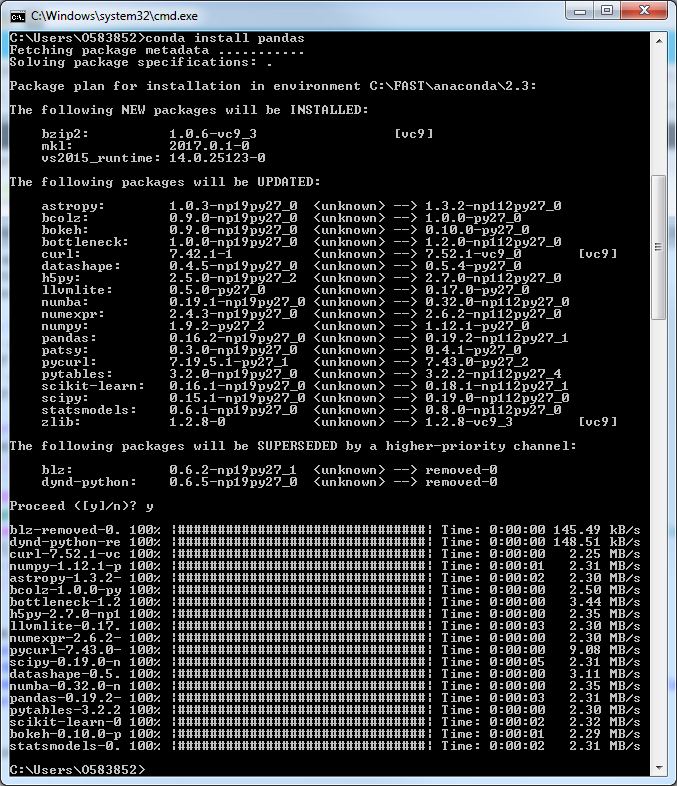
*Requires: Bloomberg Terminal (for data source)*

Anaconda Python for Windows can be downloaded here: <https://www.continuum.io/downloads#windows>

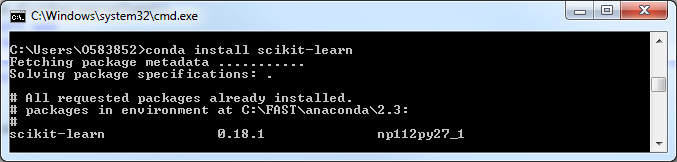
Our machine installed Anaconda Python using default settings to the directory: C:\FAST\. The version of Python installed was 2.7.



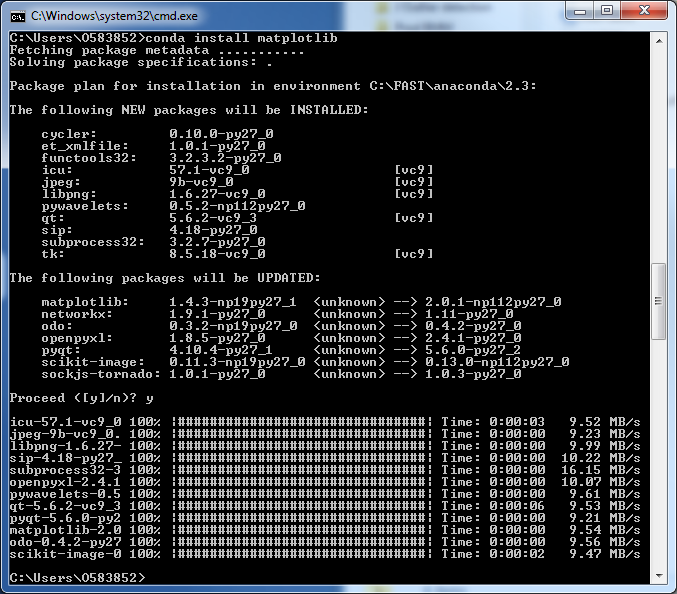
The pandas library is used and must be installed using command prompt: *conda install pandas*



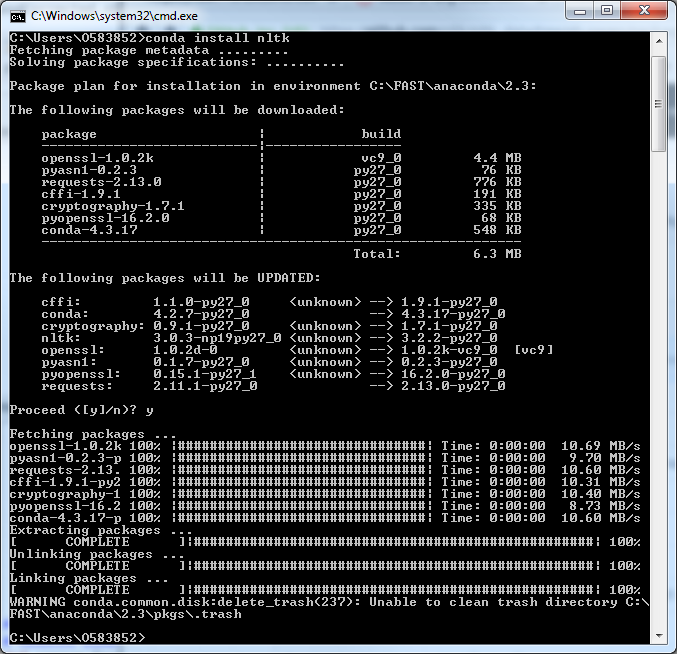
The scikit-learn package also needs to be installed: *conda install scikit-learn*



The graphing library matplotlib is also needed: *conda install matplotlib*



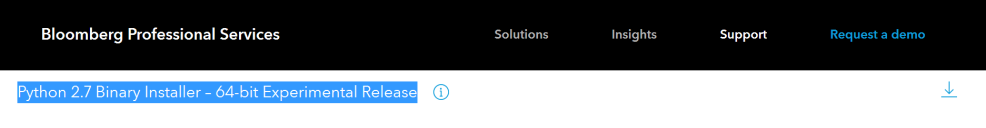
Finally, nltk is installed using the command: *conda install nltk*



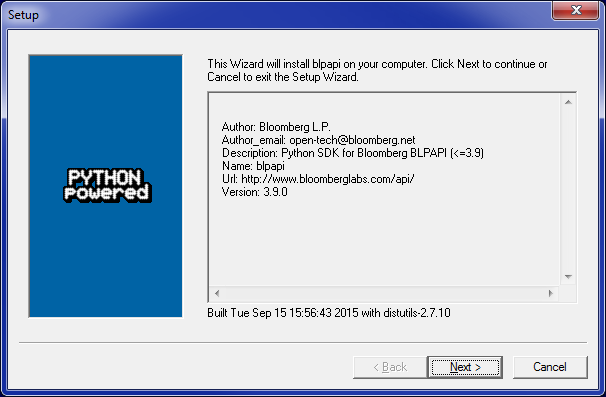
Anaconda automatically installs dependent libraries such as scipy.

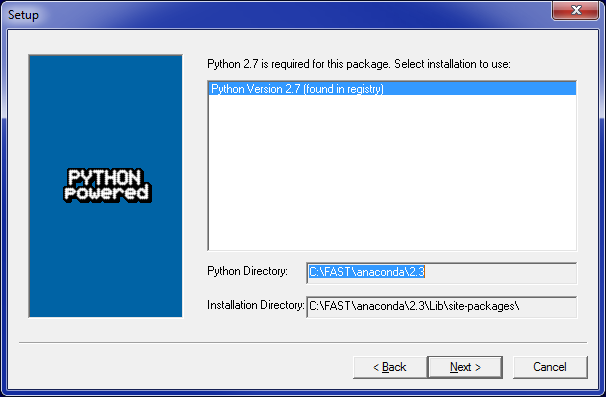
Next, download the self-extracting Bloomberg blpapi library executable from the Bloomberg website:

<https://www.bloomberg.com/professional/support/api-library/>

****

We are finished with the configuration once we run the downloaded file blpapi\_python\_3.9.0-win-amd64-py2.7.exe:





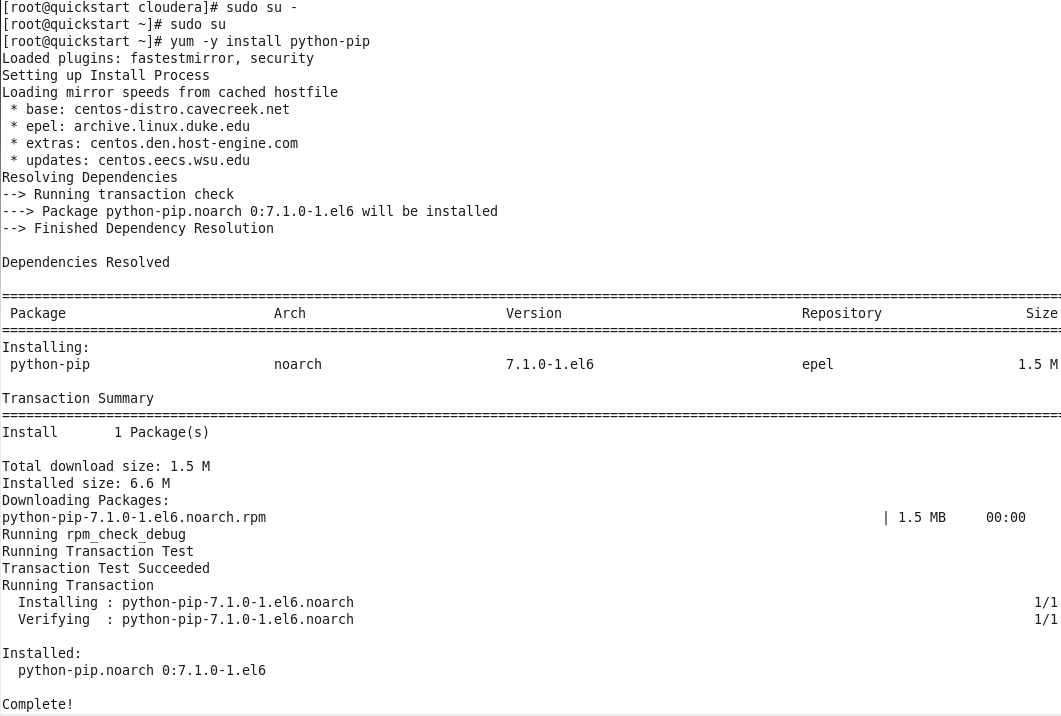
**Configuration Manual: Spark Streaming and Naïve-Bayes Model**

*Configuration: Cloudera Quickstart CDH 5.8 on Virtualbox (*[*https://www.cloudera.com/downloads.html*](https://www.cloudera.com/downloads.html)*)*

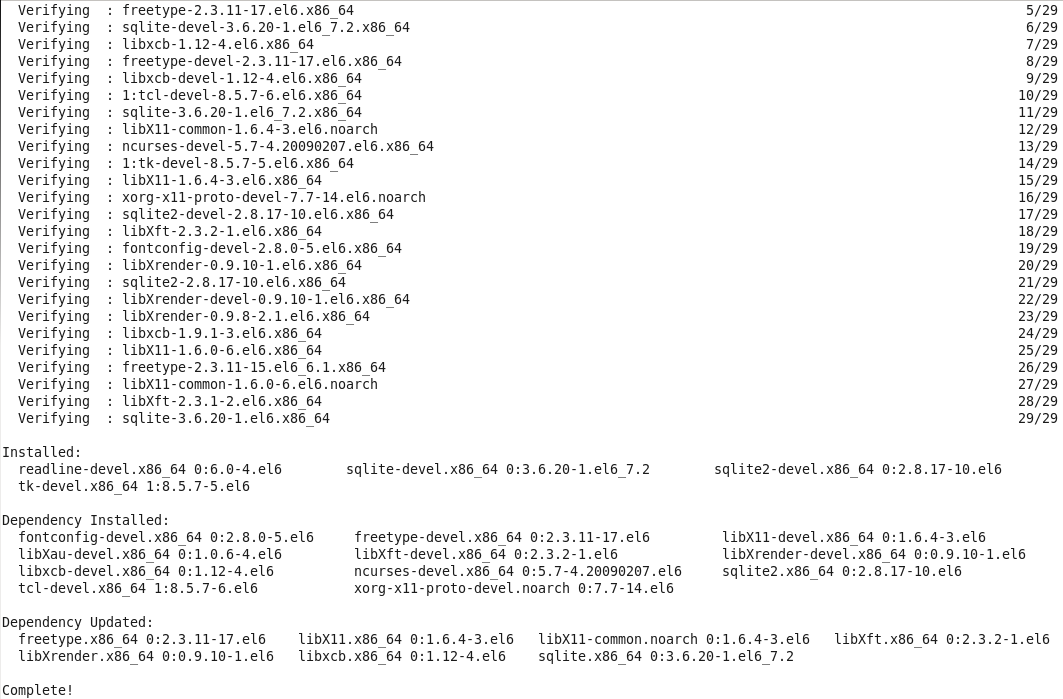
The NLTK 3 library must be installed in order to handle some of the natural language processing we do in cleaning tweet messages. NLTK 3 only supports Python 2.7, not 2.6 which is the default version in Cloudera and PySpark. With the following steps below, we create a virtual environment for Python 2.7 and install NLTK 3, sklearn, pandas, matplotlib, and scipy:

Using terminal in Cloudera, become root (password ‘cloudera’): *sudo su*

Useyum toinstall pip: *yum –y install python-pip*

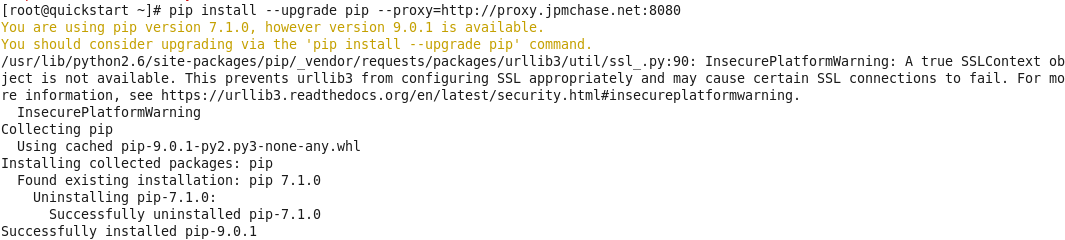


Use yum to install important libraries: *yum –y install readline-devel tk-devel sqlite-devel sqlite2-devel*

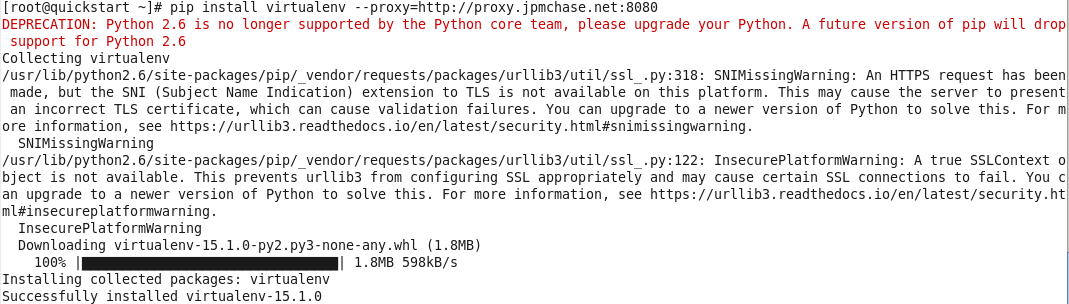


Upgrade pip: *pip install --upgrade pip*

Note: for pip commands, you can add a “--proxy=[PROXY\_SERVER]” parameter if behind proxy server (but otherwise don’t need to)

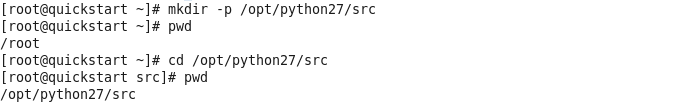


Install virtualenv: *pip install virtualenv*



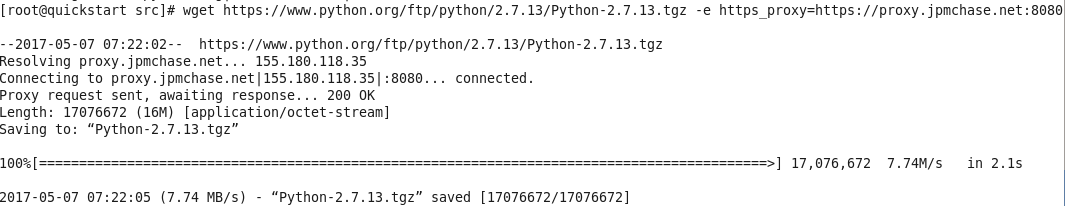
Make directory for Python 2.7: *mkdir –p /opt/python27/src*

Navigate to the new folder: *cd /opt/python27/src*

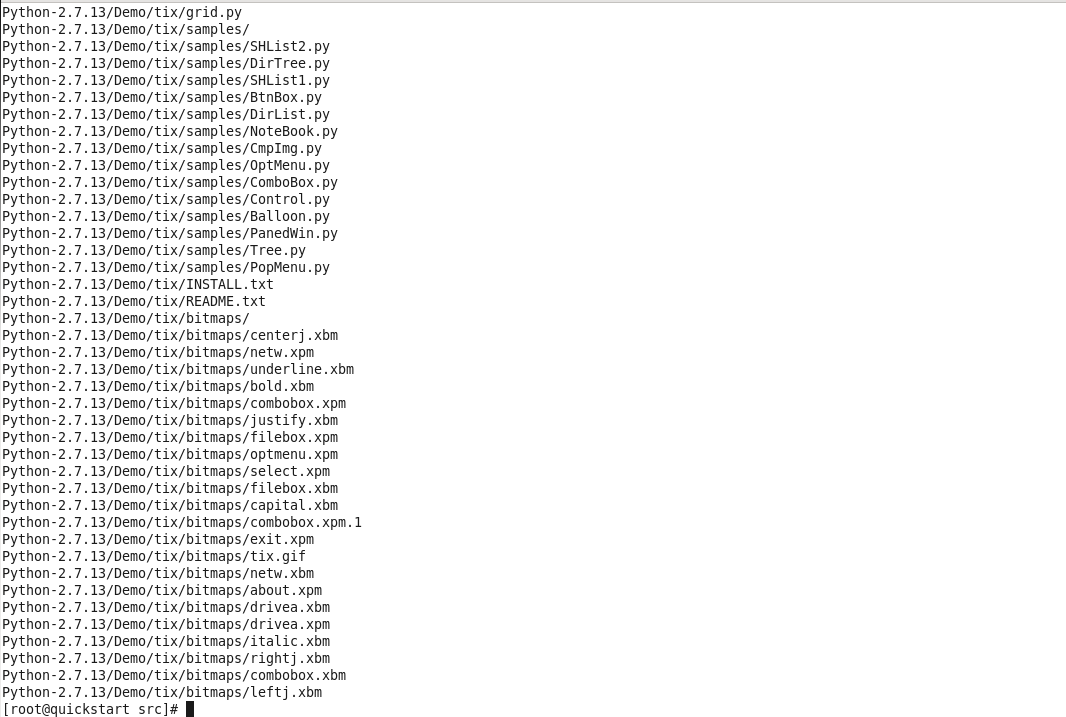


Download python 2.7: *wget https://www.python.org/ftp/python/2.7.13/Python-2.7.13.tgz*

Note: for wget commands, you can add the following parameter to get around proxy servers: “-e https\_proxy=[PROXY\_SERVER]” but again only if this is needed.

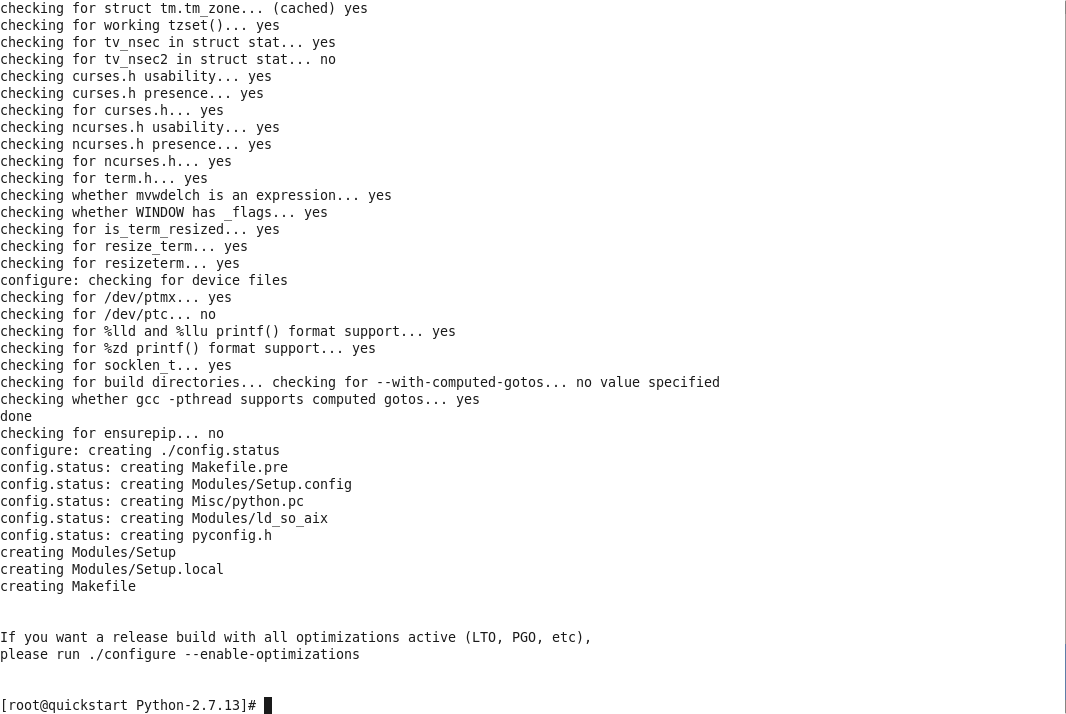


Unzip the file: *tar –xvf Python-2.7.13.tgz*

**

Change directory: *cd Python-2.7.13*

Run configure script: *./configure --prefix=/opt/python27/*

**

Run the make procedure: *make*



Run the make install procedure: *make install*



Exit root session: *exit*

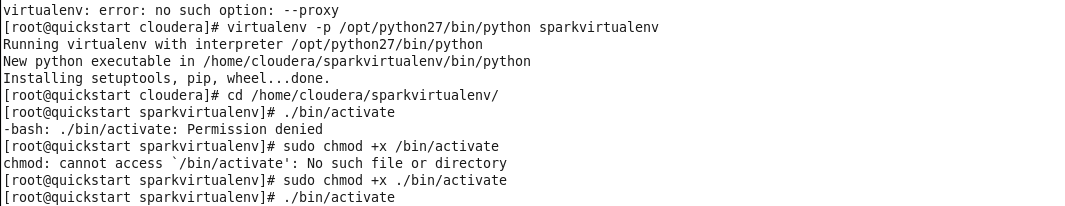
Go to this directory*: cd /home/cloudera/*

Create virtual environment: *virtualenv –p /opt/python27/bin/python sparkvirtualenv*

Change into virtualenv*: cd /home/cloudera/sparkvirtualenv/*

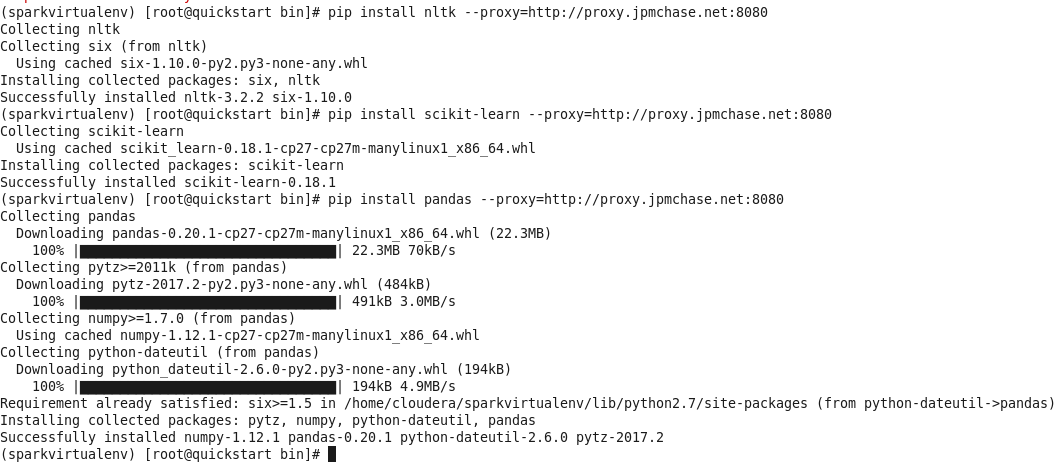
Change permissions to run activate: *sudo chmod +x ./bin/activate*

Activate the virtualenv: *source ./bin/activate*

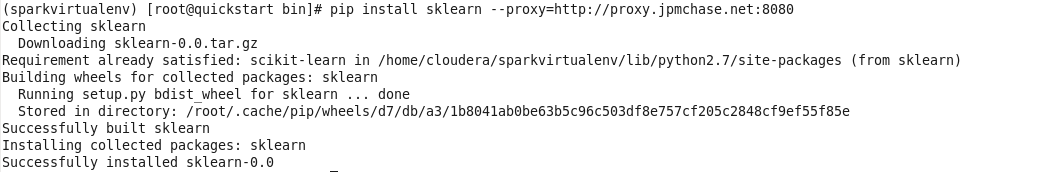
**

*pip install nltk*

*pip install pandas*

**

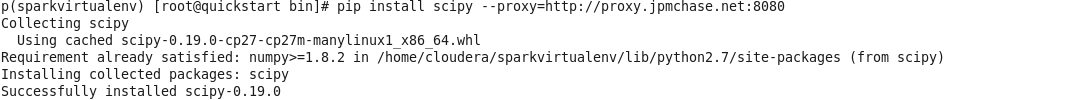
*pip install sklearn*

**

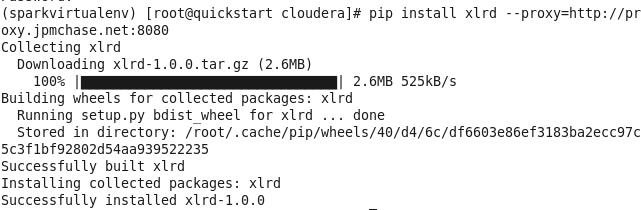
*pip install matplotlib*

**

*pip install scipy*

**

*pip install xlrd*

**

To test if this works, open pyspark and import nltk, sklearn, matplotlib, and pandas libraries. We also download NLTK, which is necessary to use the PerceptronTagger later. Note that the version of Python is the newly installed 2.7.13.

*pyspark*

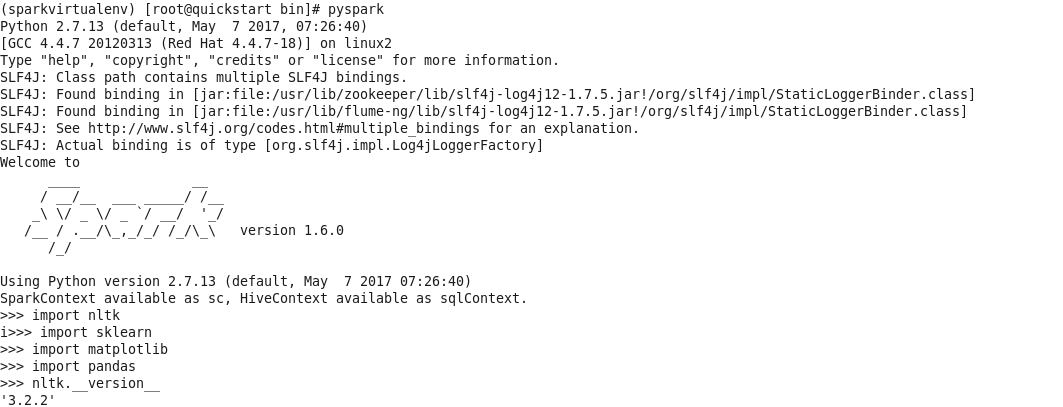
*import nltk*

*import sklearn*

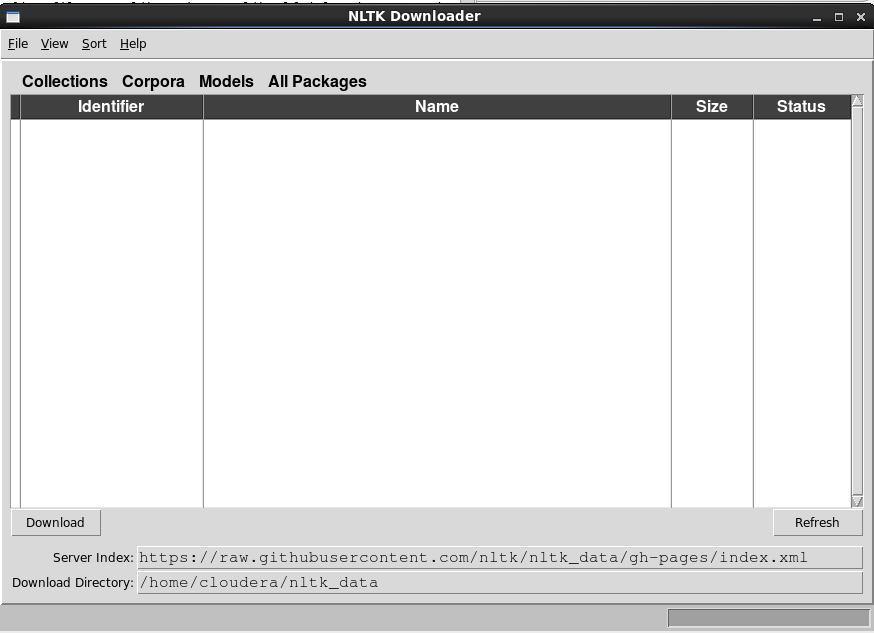
*import matplotlib*

*import pandas*

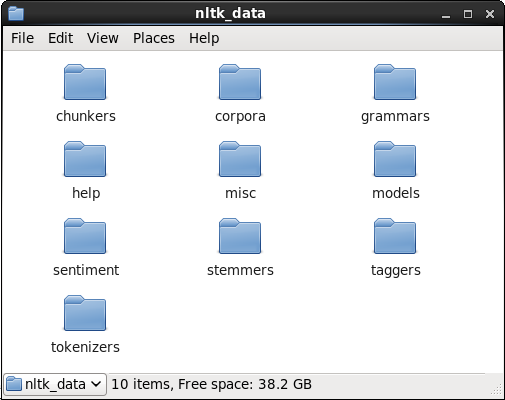
*nltk.\_\_version\_\_*



*nltk.download()*



Most configurations would show a list of items and installation would be complete after installing everything. In our test installation, the proxy server remained an issue, so we took the following steps. The optional command *nltk.set\_proxy(‘[PROXY\_SERVER]’)* allows for a proxy server but it did not work either. We downloaded the zip files from the NLTK github site (<https://github.com/nltk/nltk_data/tree/gh-pages/>) and saved and unzipped the contents of the packages to the folder /home/cloudera/nltk\_data/



This concludes our installation instructions for running spark streaming and running the Naïve Bayes model.