generalassembly-studio/DSI-course-materials

Data Science Immersive "Installfest"

DSI Computer Setup

Welcome to GA's Data Science Immersive! Before you start class, you'll need to download and install a few tools. Follow this guide to get your computer all set up, and let us know if you have any questions.

Part 1. Operating System

While you can be a data scientist on any operating system, most practicing data scientists choose a Unix-type operating system, typically either Apple's OS X or a popular linux distribution such as Ubuntu or Linux Mint.

- If you are already using Mac or Linux, great! Skip ahead to **Part 2** and get started with your installs.
- If you are using a Windows machine, follow the instructions below.

For Windows Users

If you are running Windows, you will need to install a virtual machine that runs Linux. We're recommending that you use VirtualBox with Ubuntu. Here's how to get started:

1. Download and install VirtualBox.
   - **Note**: Make sure to also install the "Extension Pack" for your version!
2.
   - **Note**: You'll need at least 2-3 gbs of free disk space.
3. Open VirtualBox and import the image
4. Double-click the Virtual Machine on the left; the password is "**ilovedatascience**"

Week 10

Some parts of week 10 will be run on a Virtualbox Virtual Machine. This is going to be deployed on all computers, both Windows and Mac or Linux, so make sure you follos the instructions

> **Instructor Note:** In order to complete the lessons and labs provided, students will need to install the custom virtual machine that allows them to spin up local versions of these tools. Instructions are as follows:

Instructions

1. Ensure you have a personal Dropbox account. If you don't, please to sign up for a free version. Due to the large size, they will need an account to download the VM files.
2. Download and install Virtualbox
3. Download and install Vagrant
4. Once you have all these installed, you can run the VM by doing:

```
cd dsi-bigdata-vm
vagrant up
vagrant ssh
```

Part 2. Anaconda and Python

In our class, we'll be working closely with tools that utilize the Python programming language. Anaconda is a popular cross-platform tool that helps install and manage python-related data science libraries.

1. Download Anaconda and follow the installation instructions package for your operating system.

2. Agree to the terms and let Anaconda go through its default installation. On OS X, there is a graphical installer.

3. Just type in the package name and Anaconda will install the package and any dependencies. Anaconda should install several packages by default, including:

    - **python**: a programming language very popular with data scientists
    - **jupyter**: an interface for creating interactive python notebooks, great for sharing analyses
    - **matplotlib**: a plotting library for python
    - **nltk**: a toolkit for natural language processing
    - **numpy**: a linear algebra library
    - **pip** & **setuptools**: software to manage and install python packages
    - **scikit-learn**: a toolkit for machine learning algorithms
    - **scipy** and **statsmodels**: statistical packages for python
    - **sqlite**: a popular, easy to use database

4. If one or more of these is missing, just select the packages in the installer, or run the following at the command line (remember, you don't need to type the `$`):

    ```
    $ conda install jupyter python matplotlib nltk numpy pip setuptools scikit-learn scipy sqlite statsmodels
    ```

5. Once Anaconda is installed, you can add additional python packages from the command line as follows:

    ```
    $ conda install gensim seaborn spacy
    ```

Anaconda may also update your packages at this time (which is perfectly ok!).

Just For Linux Users

On Ubuntu, if the `conda install` command fails for some reason, restart your terminal or source your `.bashrc` like so

```
$ source ~/.bashrc
```

Part 3. Confirm Your Python Installation

1. When you've gotten this far, open up a terminal and enter the Python interpreter:

```
$ python
```

Depending on your operating system, your terminal should return something like this:

```
user@vbox:~/Downloads$ python
Python 2.7.11 |Anaconda 2.5.0 (32-bit)| (default, Dec  6 2015, 18:08:45)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-1)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
>>>
```

2. Next, make sure that the necessary packages are installed. For example, to check that `matplotlib` is installed, type in your terminal:

```
>>>> import matplotlib
>>>> print matplotlib.__version__
1.5.1
```

You may see another version (which is OK). If you get an error like this:

```
$ import matplotlib
ImportError: No module named matplotlib
```

then you'll need to try to install the Python packages again.

3. Finally, you can check the installation and versions of *all* the python libraries. You can do this a couple of different ways.

A) By running this jupyter-notebook:

```
$ cd Downloads
$ jupyter-notebook
```

- Open the notebook by selecting the notebook file
- From the `Kernel` menu select `Restart & run all`

If you see any errors then you'll need to reinstall the library that posts the error. Otherwise you should see a bunch of version numbers.

B) By typing `pip freeze` in the terminal:

- Open a terminal window

```
$ pip freeze
```

You will see a list of all the python packages currently installed with their version numbers in the terminal window.

Part 4. Git

1. We'll also be using git -- a popular version control system used to share code with others -- extensively along with Github. The installation instructions at Github are very good, and you should also make a Github account while you are there (if you don't have one already).

2. To check if your git installation is successful, open a new terminal window and try to run `git` from the command line:

   ```
   $ git --version
   ```

   The output should be something like this:

   ```
   $ git --version
   git version 2.5.0
   ```

3. Next, you'll need to tell git your name and email. Make sure to use the same email address that you use for github:

   ```
   $ git config --global user.name "Your Name"
   $ git config --global user.email your.name@example.com
   ```

   These identifiers will be added to your commits and show up when you push your changes to github from the command line!

Part 5. PostgreSQL

PostgreSQL is a database, similar to MySQL, that we'll be using later in class. Install Postgres with the following steps:

1. Follow the instructions for your operating system below

   Mac Users

   - Download Postgres.app from www.postgresapp.com
   - Move the Postgres.app to your 'Applications' folder.
   - Open the Postgres.app (using "right-click + open" since it is an application that isn't from the Mac App Store)
   - Look for the elephant in the the menu bar.

   Linux Users

```
$ sudo apt-get install postgresql postgresql-contrib postgresql-client
```

2. You need to add yourself as a user in postgres so you can access the `psql` console seamlessly. Following the commands below, replace `dsi-student` with *your own user-name* and type *your own password* when prompted.

If you are running Ubuntu, use "**ilovedatascience**" as your password.

```
$ sudo -i su - postgres
$ createuser dsi-student --superuser --password
$ createdb dsi-student
$ exit
```

Test that this works by typing `psql`. You should be presented with the postgres shell. To exit type `\q`.

Part 6. Required Tools

1. We'll be using Slack, a popular messaging platform, for our class communications.
    - Click on the installation instructions for your platform to install the Slack desktop app. You can also sign into Slack using a web interface or via their mobile app!

> Note: Add additional market & cohort-specific channel instructions here, as needed.

2. Chrome is Google's popular web browser, and it comes with a complete set of developer tools built-in. We'll use Chrome to examine code, debug scripts, and view back-end processes. If you don't already have Chrome, make sure to download and install it now.

3. Tableau is a popular dashboard creation system for visualizing data. As a data scientist you'll need to create visualizations that make your analyses accessible to colleagues, stakeholders, and decision makers.
    - Install the sofware for your operating system.
    - For now just sign up for a trial account! We'll provide you with a license key for the full version during your first week of class.

4. Import.io is a popular webscraping tool.
    - Grab the download and follow the installation instructions for your OS.

Part 7. Text Editors

A data scientist frequently writes scripts to process data, perform analysis, and create visualizations, webpages, and other end products, so you'll need a good text editor. If you don't already have a preference, try Atom or Sublime. Both editors are available for most platforms.

> Instructors should modify these options based on their preferences.

1. Download the editor of your choice from their website.

2. Install the package by double clicking the file icon or from the command line
3. Run your editor from the applications menu, or from the command line, like so:

```
$ subl
$ atom
```

This example would open up Sublime or Atom, respectively. Whichever editor you choose, be sure to practice using it!

Configure Git with your Text Editor

Finally, you'll want to tell `git` which editor it should use for your commits.

- If you choose to use Sublime, you would type:

```
$ git config --global core.editor "subl --wait --new-window"
```

- If you choose to use Atom, you would type:

```
$ git config --global core.editor "atom --wait"
```

That's it! Now you're ready to begin GA's Data Science Immersive. See you on the first day of class!