Getting started with Python, IPython Notebook & GraphLab Create

It's important to emphasize that this specialization is **not** about providing training for a specific software package. The goal of the specialization is for your effort to be spent on learning the fundamental concepts and algorithms behind machine learning in a hands-on fashion. These concepts transcend any single package. What you learn here you can use whether you write code from scratch, use any existing ML packages out there, or any that may be developed in the future.

The learning approach in this specialization is to start from use cases and then dig into algorithms and methods, what we call a *case-studies approach.* We are very excited about this approach, since it has worked well in several other courses. The first course is focused on understanding how ML can be used in various cases studies, and the follow on courses will dig into the details of algorithms and methods for each of the main ML areas. In the first course, you will not be implementing algorithms from scratch, but rather building intelligent applications that use ML. In the subsequent course, we will be implementing and comparing a wide range of algorithms. To make it easy to implement the use cases we will be covering, we are recommending a particular set of software tools, but you can successfully complete the course with other tools out there.

Why Python

In this course, we are going to use the Python programming language to build several intelligent applications that use machine learning. Python is a simple scripting language that makes it easy to interact with data. Furthermore, Python has a wide range of packages that make it easy to get started and build applications, from the simplest ones to the most complex. Python is widely used in industry, and is becoming the *de facto* language for data science in industry. (R is another alternative language. However, R tends to be significantly less scalable and has very few deployment tools, thus it is seldomly used for production code in industry. It is possible, but highly discouraged to use R in this specialization.)

We will also use the IPython Notebook in our videos. The IPython Notebook is a simple interactive environment for programming with Python, which makes it really easy to share your results. Think about it as a combination of a Python terminal and a wiki page. Thus, you can combine code, plots and text to explain what you did. (You are not required to use IPython Notebook in the assignments, and should have no problem using straight up Python if you prefer.)

Why SFrame & GraphLab Create

There are many excellent machine learning libraries in Python. One of the most popular one today is scikit-learn. Similarly, there are many tools for data manipulations in Python; a popular example is Pandas. However, most of these tools do not scale to large datasets, including some we will tackle in this Specialization. In addition, in this specialization, we will cover a wide range of ML models, feature engineering transformation, and evaluation metrics. With most existing packages, you will have to install a combination of packages to get the tools that we need to tackle the use cases in this course. This is possible, but requires advanced knowledge of Python, which we feel will slow down most people's learning of the core concepts.

The main goal of this course is to learn core ML concepts, not how to use a specific software package. Thus, in this course, we recommend you use GraphLab Create, a package we have been working on for many years now, and has seen an exciting adoption curve, especially in industry with folks building real applications. GraphLab Create is a highly scalable machine learning library for Python, which also includes the SFrame, a highly-scalable library for data manipulation. A huge advantage of SFrame over Pandas is that with SFrame, you are not limited to datasets that fit in memory, which allows you to deal with large datasets, even on a laptop. (The SFrame API is very similar to Pandas' API. Here is a doc showing the relationship between the two of them.)

Licenses for SFrame & GraphLab Create

The SFrame package is available in open-source under a permissive BSD license. So, you will always be able to use SFrames for free.

GraphLab Create is free on a 1-year, renewable license for educational purposes, including Coursera. This software, however, has a paid license for commercial purposes.

For full disclosure!

GraphLab Create is very actively used in industry by a large number of companies. This package was created by a machine learning company called Dato. This company is spin off from a popular research project called GraphLab, which Carlos Guestrin, one of your two instructors, and his research group started at Carnegie Mellon University. In addition to being a professor at the University of Washington, Carlos is the CEO of Dato.

The reason we suggest you use GraphLab Create is not because Carlos is the CEO of Dato :), but because we very strongly believe using this software will make it much easier for us to follow the "case-study approach" we are taking in this specialization. In particular, it will let you focus on exploring each case study in this first course, without having to implement your own algorithms from scratch, and benefiting from the performance advantages that

GraphLab Create provides. ***In subsequent courses in the specialization, you will be implementing many of these algorithms from scratch, having had the foundation of seeing them perform in practice on real applications.***

We are happy, however, for you to use any tool(s) of your liking, by following the steps below. As you will notice, we are only grading the output of your programs, so the specific software tool is not the focus of the course.

It's important to emphasize that this specialization is **not** about providing training for a specific software package. The goal of the specialization is for your effort to be spent on learning the fundamental concepts and algorithms behind machine learning in a hands-on fashion. These concepts transcend any single package. What you learn here you can use whether you write code from scratch, use any existing ML packages out there, or any that may be developed in the future. We are happy to hear that so many of you are enjoying this approach so far!

Using other ML packages

We strongly encourage you to use SFrame for this course.

You are welcome to use other ML packages, like scikit-learn, instead of GraphLab Create. However, we believe this will significantly slow down the your implementation tasks, especially for this first course.

The first course is focused on exploring the use cases we'll tackle throughout the specialization. A huge goal here is to familiarize ourselves with the core ML concepts that we will use the 5 follow-on courses. In those course, there will be much more implementation of ML algorithms, so the specific ML package becomes less important. But, in this first course, we want to move quickly through all the use cases, and GraphLab Create will help us do just that.

If you choose to use a different package, we will provide the data sets and the assignment questions will not depend specifically on GraphLab Create.

Learning outcomes

This reading will walk you through the steps you will need to follow to install and get started with Python, IPython Notebook, and GraphLab Create.

- Installing Python, IPython Notebook, and GraphLab Create
- Starting IPython Notebook
- Writing variables, functions and loops in Python

- Doing basic data manipulations in Python with SFrames

Getting started using these resources

You have two options: downloading and installing the required software or using a prepackaged version on a free instance on Amazon EC2.

Option 1: Downloading and installing all software on your own machine

- Download and install Python, iPython notebook and GraphLab Create. You can find the instructions here.

https://dato.com/learn/coursera/

- There are many Python resources available online. Here is a good place for documentation.
- For GraphLab Create, there is also a lot of information available online. Here are some starting points.

Learning Concepts about the Tools	https://dato.com/learn/
The User Guide	https://dato.com/learn/userguide/
More Detailed API Docs	https://dato.com/products/create/docs/

Option 2: Using a free Amazon EC2 with all the software pre-installed

If you do not have a 64-bit computer, you will not be able to run GraphLab Create. Additionally, some of you may want a simple experience where you don't have to download the course content and install everything locally. Here, we'll address these situations!

Amazon EC2 offers free cloud computing hours with what they call micro instances. These instances are all we need to do the work for this course. We have created an image for one such instance that is easy to launch and contains all the course content. **This will allow you to run everything you need for this course in the cloud for free, without having to install anything locally.** (You do need to create an Amazon EC2 account and have internet access.)

You can find step-by-step instructions here:

https://dato.com/download/install-graphlab-create-aws-coursera.html

We note that installing all the software on your own local machine may be the right option for most people; especially since you can run locally everything without needing to be online to do the homeworks. But, the option using Amazon EC2 should be a great alternative.

Watch the videos on getting started with IPython Notebook and SFrames

If you haven't done so yet, before you start, we recommend you watch the videos where we go over Python, IPython notebook and SFrames using GraphLab Create.

Download the data and sample code and familiarize yourself with the notebooks

Before doing the assignments in this course, familiarize yourself with the two notebooks we covered in the videos:

- Download the notebook that covers getting started with Python: Getting started with iPython Notebook.ipynb
- Download the notebook that covers getting started with SFrames: Getting Started with SFrames.ipynb
- Download the simple people dataset: people-example.csv
- Save all these files in the same directory (where you are calling iPython notebook from). **Not sure where to save the files? See this guide.**

Familiarize yourself with the notebooks

Make sure:

- You've downloaded and installed Python, IPython Notebook, and GraphLab Create.
- Started up IPython Notebook from the directory you downloaded the files above, by launching graphical launcher (for Windows and Mac) or by typing

```
ipython notebook
```

in your command line (for Linux).

Now you are ready to get started! Familiarize yourself with Python and SFrames, as well as writing code with the IPython Notebook. From here, you will be ready to do all the assignments in the course, and build awesome intelligent applications that use machine learning!