Collaboratory at Columbia University

Bootcamp Overview

The data science bootcamp is an intensive one week course. In each day of the course we will cover a particular set of topics which require prior preparation. In order for you to better prepare yourself with the material covered on a given day we ask that you go over a set of videos and reading assignments. It is mandatory that you cover all the prior assignments before the given day. Below is the list of prior assignments for each day of the bootcamp.

Video Assignments | *Free edX Modules*

1. **DS102x** Machine Learning for Data Science and Analytics http://bit.ly/1SH4dk3
2. **DS101x** Statistical Thinking for Data Science and Analytics http://bit.ly/1Vzjdnn
3. **DS103x** Enabling Technologies for Data Science and Analytics: IOT http://bit.ly/2iYCimE

Additional links to required materials are listed below for each days' lesson plans. Refer to module number as needed.

Mandatory Assignments

DAY 1 | Introduction to Data Science		
Video/Module(s)	Data Science	Time Involved
DS101x 2c:	Simple Visualization and Summaries	18:07
DS101x 2e:	Data Collection & Sampling	15:39
	Probability	Time Involved
DS101x 2f:	Introduction to Probability	33:33
DS101x 3a:	Conditional Probability	13:38
Read Article	Regression Analysis	*Access via Google Drive
DS101x 3g:	Regression Analysis 3 & 4	10:17
DS101x 3h:	Regression Analysis 4 & Remarks	11:14
DAY 2 | Algorithms		
Video/Module(s)	Intro to Algorithms	Time Involved
DS102x 1b:	Tools to Analyze Algorithms	23:33
YouTube	Andrew Ng's, "Gradient Descent"	11:30 http://bit.ly/1MOczcV
YouTube	Andrew Ng's, "G.D.: Intuition"	11:50 http://bit.ly/2iQxwuh
	Intro to Machine Learning	Time Involved
DS102x 4a:	What is Machine Learning	17:06
DS102x 4c:	Classification	16:31
DS102x 4d:	Linear Classifiers	22:48
YouTube	Georgia Tech: Naive Bayes	7:52 http://bit.ly/2jc8Fxf
YouTube	Udacity: Bayes Rule for Classification	2:09 http://bit.ly/2jjEvfL

*Google Drive Files: https://goo.gl/sxBeS8 | Requires UNI or GMAIL login

Folder contains materials which are property of Columbia University. Please do not redistribute!

DAY 3 | MACHINE LEARNING		
VIDEO/MODULE(S)	INTRO TO MACHINE LEARNING	TIME INVOLVED
Coursera	Geoffrey Hinton's "The Idea of Full Bayesian Learning"	7:27 http://bit.ly/2i10KoO
	PROBABILISTIC MODELING	TIME INVOLVED
DS102x 5b:	Probabilistic Modeling I	6:51
DS102x 5c:	Probabilistic Modeling II	10:19
DS101x 3j:	Clustering Text	27:40
	EVALUATION	TIME INVOLVED
DS102x 4g:	Cross Validation	17:18
DS102x 2g:	Statistical Inference – Confidence Intervals	20:50
DS102x 2h:	Statistical Inference – Significant Tests	19:28
DAY 4 | NEURAL NETWORKS & NATURAL LANGUAGE PROCESSING		
VIDEO/MODULE(S)	NEURAL NETWORKS	TIME INVOLVED
YouTube	Geoffrey Hinton's "What are Neural Networks"	8:30 http://bit.ly/2hZ3uja
YouTube	Geoffrey Hinton's "Some Simple Models of Neurons"	8:23 http://bit.ly/2iQDQ4T
YouTube	Geoffrey Hinton's "A Simple Example of Learning"	5:38 http://bit.ly/2i11HO4
YouTube	Geoffrey Hinton's "Types of Neural Network Architecture"	7:28 http://bit.ly/2iZ2DRm
YouTube	Geoffrey Hinton's "Modeling Sequences: A Brief Overview"	17:23 http://bit.ly/2iKdp0n
YouTube	Geoffrey Hinton's "Long-term Short-term (LSTS) Memory"	9:15 http://bit.ly/2iMqnIJ
YouTube	Geoffrey Hinton's "Dropout"	8:35 http://bit.ly/2iZaqi8
YouTube	Geoffrey Hinton's "Convoluted Nets for Digital Recognition"	16:01 http://bit.ly/2iQIZtO
	NATURAL LANGUAGE PROCESSING	TIME INVOLVED
DS103x 4g:	Application of NLP	8:00
DS103x 4b:	NLP News	16:06
DS103x 4k:	Syntax and Parsing	13:50

SOFTWARE REQUIREMENTS

The following outlines all of the required software packages along with install instructions.

In this bootcamp we will be using Jupyter Notebook with iPython3 which is included in the Anaconda distribution. You can download Anaconda with python version 3.5 following the instructions from the official website: https://www.continuum.io/downloads

Software Requirements – Continued

The Anaconda distribution also contains most of the packages that we will be using during the lab sessions, except for the following three: gensim, seaborn, and tensorflow. Please use the following terminal commands for these three packages:

>> conda install gensim

>> conda install -c jjhelmus tensorflow=0.12.0rc0

NOTE TO WINDOW USERS:

We suggest using the Anaconda terminal to run these commands.

For additional installation instructions and troubleshooting during installation, we recommend following the official online resources for the specific package. In instances where these resources are not helpful in troubleshooting the problem we recommend that you search online community postings (e.g. stackoverlow.com). For additional help please see us before the start of the bootcamp on Monday morning at 9am when we will be offering support to participants that still have issues with setting up their computers. Due to time limitations we won't be able to help you install all of the components if you haven't previously tried to do so on your own.

Once you are done installing all the required components and packages you would need to verify that your iPython environment is working. To do so open Jupyter notebook and load the following iPython file located in the "cu_datascience_bootcamp" folder:

cu_datascience_bootcamp/sample_notebook.ipynb

Below is the link to the "cu_datascience_bootcamp" folder which is located on Google drive and it is accessible using your UNI or gmail account: https://goo.gl/sxBeS8

Folder contains materials which are property of Columbia University. Please do not redistribute!