Course Title

Data Science Specialization SwiftKey Capstone

Course Instructor(s)

- Jeff Leek
- Roger Peng
- Brian Caffo

You are a data scientist now

The goal of this data science specialization has been to give you the basic skills involved with being a data scientist. The goal of this capstone is to mimic the experience of being a data scientist. As a practicing data scientist it is entirely common to get a messy data set, a vague question, and very little instruction on exactly how to analyze the data. Our goal is to give you that same experience but with added support in the form of forums, discussion with instructors, feedback from SwiftKey and Coursera engineers, and a structured problem to solve. We hope that you will take advantage of the opportunity this project affords for you to demonstrate your skills and creativity.

Course Tasks

This course will be separated into 8 different tasks that cover the range of activities encountered by a practicing data scientist. They mirror many of the skills you have developed in the data science specialization. The tasks are:

- Understanding the problem
- Data acquisition and cleaning
- Exploratory analysis
- Statistical modeling
- Predictive modeling

- Creative exploration
- Creating a data product
- Creating a short slide deck pitching your product

You will hear about each of these tasks over the course of the capstone.

Assessements and Grading

To successfully complete the capstone project, you must receive a passing grade on all of the following assignments:

1. Quiz 1: Getting Started
2. Milestone Report: exploratory analysis of the data set + evaluation of at least three classmate submissions
3. Quiz 2: Natural Language Processing I
4. Quiz 3: Natural Language Processing II
5. Final Project: your data product and a presentation describing your final data product + evaluation of at least three classmate submissions

The quizzes will be standard multiple choice quizzes. The other components are graded by peer evaluation.

Your final grade will be calculated as follows

- Quiz 1 - 5%
- Milestone Report - 20%
- Quiz 2 - 10%
- Quiz 3 -10%
- Final Project -55%

Course dataset

This is the training data to get you started that will be the basis for most of the capstone. You must download the data from the link below and not from external websites to start.

- https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip

Later in the course you may use external data sets to augment your model as you see fit.

Differences of opinion

Keep in mind that currently data analysis is as much art as it is science - so we may have a difference of opinion - and that is ok! Please refrain from angry, sarcastic, or abusive comments on the message boards. Our goal is to create a supportive community that helps the learning of all students, from the most advanced to those who are just seeing this material for the first time.

Plagiarism

Johns Hopkins University defines plagiarism as "...taking for one's own use the words, ideas, concepts or data of another without proper attribution. Plagiarism includes both direct use or paraphrasing of the words, thoughts, or concepts of another without proper attribution." We take plagiarism very seriously, as does Johns Hopkins University.

We recognize that many students may not have a clear understanding of what plagiarism is or why it is wrong. Please see the following guide for more information on plagiarism:

http://www.jhsph.edu/academics/degree-programs/master-of-public-health/current-students/JHSPH-ReferencingHandbook.pdf

It is critically important that you give people/sources credit when you use their words or ideas. If you do not give proper credit -- particularly when quoting directly from a source -- you violate the trust of your fellow students.

The Coursera Honor code includes an explicit statement about plagiarism:

I will register for only one account. My answers to homework, quizzes and exams will be my own work (except for assignments that explicitly permit collaboration). I will not make solutions to homework, quizzes or exams available to anyone else. This includes both solutions written by me, as well as any official solutions provided by the course staff. I will not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.

Reporting plagiarism on course projects

Keep in mind that some components of the projects will be very similar across terms and so answers that appear similar may be honest coincidences. However, we would appreciate if you do a basic check for obvious plagiarism and report it during your peer assessment phase. Use the flagging function in the peer evaluation interface to alert staff to suspected plagiarism.

It is currently very difficult to prove or disprove a charge of plagiarism in the MOOC peer assessment setting. However, we will do our best working with the Coursera team to adjudicate plagiarism. We consider plagiarism of class projects for the capstone to be a waste of the effort you have put into this series and a serious violation of trust.

✓ Complete

👍 👎 ⚑