PS2 - Data Science Tools

Lane Cazier

February 2021

Measurement Obviously, in the field of data science, data is very important. In order to get good data, care needs to be taken when gathering it. Good data needs to be measured well. If you are conducting research into what make a football team win, and you want to have one of your variables be the accuracy of a teams quarterback, you need a good way to measure this. It may be very important to a team's success, but if you cannot accurately measure it, then it won't be accurately reflected in the data or model.

Statistical Programming Languages In order to handle large data sets, it is often very useful to use a computer. In order most efficiently utilize a computer, you should know how to code and program. The most common languages used for data science are Python, R, and Julia. Using these will allow you to automate many basic processes and save time and effort.

Web Scraping One of the places data is frequently stored is the internet. Using programming languages, you can efficiently collect data from the internet for your own use.

Handling Large Data Sets Sometimes you will generate extremely large data sets when getting data to use for research and modeling. These can be handled using RDD(Resilient Distributed Datasets), useful when a data set is too large for a normal computer alone to handle. SQL is also a useful tool to polish datasets and turn them into a more useful and accessible form.

Visualisation Also important, after you obtain data, you need to be able to see, understand, and present it. Visualization is very important in this, and statistical programming languages have libraries that can be easily used to graph and display collected data, allowing you to more readily comprehend it.

Modeling An important part of the scientific method is testing hypothesis, and making predictions. Modeling helps test ideas and theories, as well as generate explanations, which can be used to help make predictions.