Dumping Big Data

8/2/2020

Yuxing Yan[1]

Abstract

In this paper, we present an easy way to dump big data. For example, it takes about 2 minutes to download all the SEC (Securities Exchange Commission) quarterly index files from 1993 to 2020 (Q2). In total, there are 2 dozen big data sets, including the SEC quarterly index files, SEC Financial Statements, Financial Statements and Notes, SEC 10-K, 10-Q, 13f, S1, Forms 3, 4 and 5, Census SF1 (SF2), Census Demographic Profile, NYSE sample high-frequency trading data, all zip files from French's Data Library, TORQ and the like. Our method makes it trivial to access relatively large data for instructors and students at various Business Analytics/Data Analytics programs. Our method is extremely simple: download and install R, then issue one-line of R code. The total size of the data covered is around 80TB (terabyte, $1TB=10^{12}$ byte). To process the downloaded data efficiently is beyond the scope of this paper. One solution seems obvious though: learn a programming language.

JEC code: A2, I22, G00

Keywords: Big data, SEC filings, Census, NYSE high frequency data, R, dumping

[1] School of Business, SUNY at Geneseo, 1 College Circle, Geneseo, NY 14454, pyan@geneseo.edu.