

Dumping big data

1/14/2019

Yuxing Yan¹

Abstract

In this paper, we present an easy way to dump big data. For example, it takes 2 minutes to download all the SEC quarterly index files from 1993 to 2018. 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 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 business schools and various Business Analytics and Data Science programs. Our method is extremely simple: download and install R and issue one-line R code. The total size of the data covered is around 80TB (terabyte, 1TB=10¹² byte). To process the downloaded data efficiently is beyond the scope of this paper. One solution seems obvious: learn a programming language.

JEC code: A2 I22 G00

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

¹ Department of Economics and Finance, Canisius College, email: yany@canisius.edu