PS4 Fretland

efretland

February 2018

1 Sources of Data for Scraping

As someone who is interested in sports analytics, I believe various football-related websites would be extremely useful to use. NFL has a basic statistics database, while Pro Football Reference has a more thorough database that is better organized and even more useful. There are also various fantasy football sites out there that would be useful, including Fleaflicker and ESPN.

On a different note, I would like to be able to scrape from the Bureau of Labor Statistics website. I would also like to scrape from financial markets websites in order to perform analyses on the NYSE.

2 R Script Answers

4) class(df) is type "SparkDataFrame" while class(df1) is data.frame.

5) a. Sepal Length Species 1 5.1 setosa 2 4.9 setosa 3 4.7 setosa 4 4.6 setosa 5 5.0 setosa 6 5.4 setosa

b. The two function exactly the same, although the syntax is different. Without sparkR, the operators need to be removed and the dataframe name isn't necessary after it's used the first time.

6) a. Sepal Length Sepal Width Petal Length Petal Width Species 1 5.8 4.0 1.2 0.2 setosa 2 5.7 4.4 1.5 0.4 setosa 3 5.7 3.8 1.7 0.3 setosa 4 7.0 3.2 4.7 1.4 versicolor 5 6.4 3.2 4.5 1.5 versicolor 6 6.9 3.1 4.9 1.5 versicolor

b. Again, the two function similarly but the syntax differs in the same way.

7) Sepal Length Species 1 5.8 setosa 2 5.7 setosa 3 5.7 setosa 4 7.0 versicolor 5 6.4 versicolor 6 6.9 versicolor

8) Species Mean Count virginica 6.588 50 versicolor 5.936 50 setosa 5.006 50

9) Sepal Length Sepal Width Petal Length Petal Width Species 1 5.1 3.5 1.4 0.2 setosa 2 4.9 3.0 1.4 0.2 setosa 3 4.7 3.2 1.3 0.2 setosa 4 4.6 3.1 1.5 0.2 setosa 5 5.0 3.6 1.4 0.2 setosa 6 5.4 3.9 1.7 0.4 setosa