Problem Set 4

Matt Mullins

February 2018

1 Question 1

Describe the type of data and sources you would be interested in scraping from:

1. Since 2018 was the safest year for flying ever, I would be interested in researching the previous 15-20 years of accidents from each of the major airlines to uncover any insights such as who is the safest, what causes the most accidents, when are accidents most likely to occur, etc. Link- https://aviation-safety.net

2. I think it would be interesting to look at how Twitter can be used as a tool to increase stock prices during an acquisition. I would use a combination of scarping tweets related to specific acquisitions or investments and combine that with the stock prices over the same range of days and perhaps further to see how the performance continues after the hype decreases.

3. It would be interesting to scrape recruiting rankings from NCAA basketball data on ESPN and use that info to determine how the team performs the following two to three years in order to see how important those rankings are for team success.

2 Question 2

Answer to questions in PS4:

5.D what type of object is this?
data.frame SparkRDataFrame

6.5/6 if you try the same command on DF1 what happens? Why do you think this happens?

I receive an error message when trying to execute these different commands with the df1 data frame. I think it is because when I try to execute the commands in SparkR it does not like or is unable to manipulate this data frame because of the way that it was created.