Data Science for Economists - Spring 21

Problem Set 4

Syed Waleed Mehmood Wasti

February 25, 2021

Q7. I am interested in finding out how the engagement level of a certain tweet of an individual impacts his/her subsequent tweet. So basically if a person tweets and gets a good response on it in terms of the likes and comments then how likely is he/she to tweet sooner rather than later. Also is the subsequent tweet likely to be on a similar topic or something very diverse.

Practice with JSON files (R exercise part 1)

a N/A

b N/A

c N/A

d Ans: mydf is: [1] "tbl_df" "tbl" "data.frame"
mydf$date is [1] "character"

e First 10 rows: Command= head(mydf,10);
Result = A tibble: 10 x 6
date description lang category1 category2 granularity
<chr> <chr> <chr> <chr> <chr> <chr>
1 1 Tiberius, under order of Augustu... en By place Roman Em... year
2 1 Gaius Caesar and Lucius Aemilius... en By place Roman Em... year
3 1 Gaius Caesar marries Livilla, da... en By place Roman Em... year
4 1 Quirinius becomes a chief adviso... en By place Roman Em... year
5 1 Areius Paianeius becomes Archon ... en By place Roman Em... year
6 1 The "Yuanshi" era of the Chine... en By place Asia year
7 1 Confucius is given his first roy... en By place Asia year
8 1 Emperor Ping of Han China begins... en By place Asia year
9 1 Former regent Dong Xian commits ... en By place Asia year
10 1 Sapadbizes, Yuezhi prince and Ki... en By place Asia year

Practice with sparklyr (R Exercise part 2)

1. N/A
2. N/A
3. N/A
4. N/A
5. N/A

6. N/A

7. class(df1) is: [1] “tbl_df” “tbl” “data.frame”
   class(df) is: [1] “tbl_spark” “tbl_sql” “tbl_lazy” “tbl”

   Yes. They are both different.

8. Yes, the column names are different.

9. (a) Sepal_Length Species
   <dbl> <chr>
   1 5.1 setosa
   2 4.9 setosa
   3 4.7 setosa
   4 4.6 setosa
   5 5 setosa
   6 5.4 setosa

10. (a) Source: spark<?> [?? x 5] Sepal_Length Sepal_Width Petal_Length Petal_Width Species
    <dbl> <dbl> <dbl> <dbl> <chr>
    1 5.8 4 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 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

11. Source: spark<?> [?? x 2]
    Sepal_Length Species
    <dbl> <chr>
    1 5.8 setosa
    2 5.7 setosa
    3 5.7 setosa
    4 7 versicolor
    5 6.4 versicolor
    6 6.9 versicolor

12. Source: spark<?> [?? x 3]
    Species mean count
    <chr> <dbl> <dbl>

1 virginica 6.59 50
2 versicolor 5.94 50
3 setosa 5.01 50

13. (a) Source: spark<?> [?? x 3]
Species mean count
<chr> <dbl> <dbl>
1 virginica 6.59 50
2 versicolor 5.94 50
3 setosa 5.01 50

(b) Species mean count
1 setosa 5.006 50
2 versicolor 5.936 50
3 virginica 6.588 50