DS 501 Case Study 1: Twitter Analysis

James Greene

09-20-2025

Contents

0.1 Preparing the data

- Load provided TweetsDF.csv file into dataframe
- Next load packages to be used for data analysis and data blending
- Examine created dataframe to validate operations. In table 1, we have the column names listed for reference while examining the data.

x
text
favorited
favoriteCount
replyToSN
created
truncated
replyToSID
id
replyToUID
statusSource
screenName
retweetCount
isRetweet
retweeted
longitude
latitude

0.2 Cleaning the Data

- To analyze the word content of the text, we must isolate and clean the text itself.
- First we will isolate the text column of the dataframe by creating a new object.
- We will assign text_df name to this new object.
- User Defined Functions designed to remove URLs and other non-informative text will be created to clean the text data

```
tweetsDF$processed_text <- apply(tweetsDF['text'], 2, removeURLs)
tweetsDF$processed_text <- apply(tweetsDF['processed_text'],2, removeUsernamesWithRT)
tweetsDF$processed_text <- apply(tweetsDF['processed_text'],2, removeUsernames)
tweetsDF$processed_text <- apply(tweetsDF['processed_text'],2, removeHashtagSignOnly)
```

- Text vector passed through custom functions.
- We can now count the total number of words in the text to analyze

```
## [1] 6826
```

0.3 Analyzing the Data

- To analyze the the text data, we need isolate the useful information - words

- First we remove stop words with lexicon library

- Filter stem words via stemming technique. Now we have a final word count after cleaning the text data:

```
## [1] 1111
```

0.4 Visualizing the data

- Now that we have a data frame of useful information, we can now visualize it
- Table 2 displays the top 30 words used, with their counts

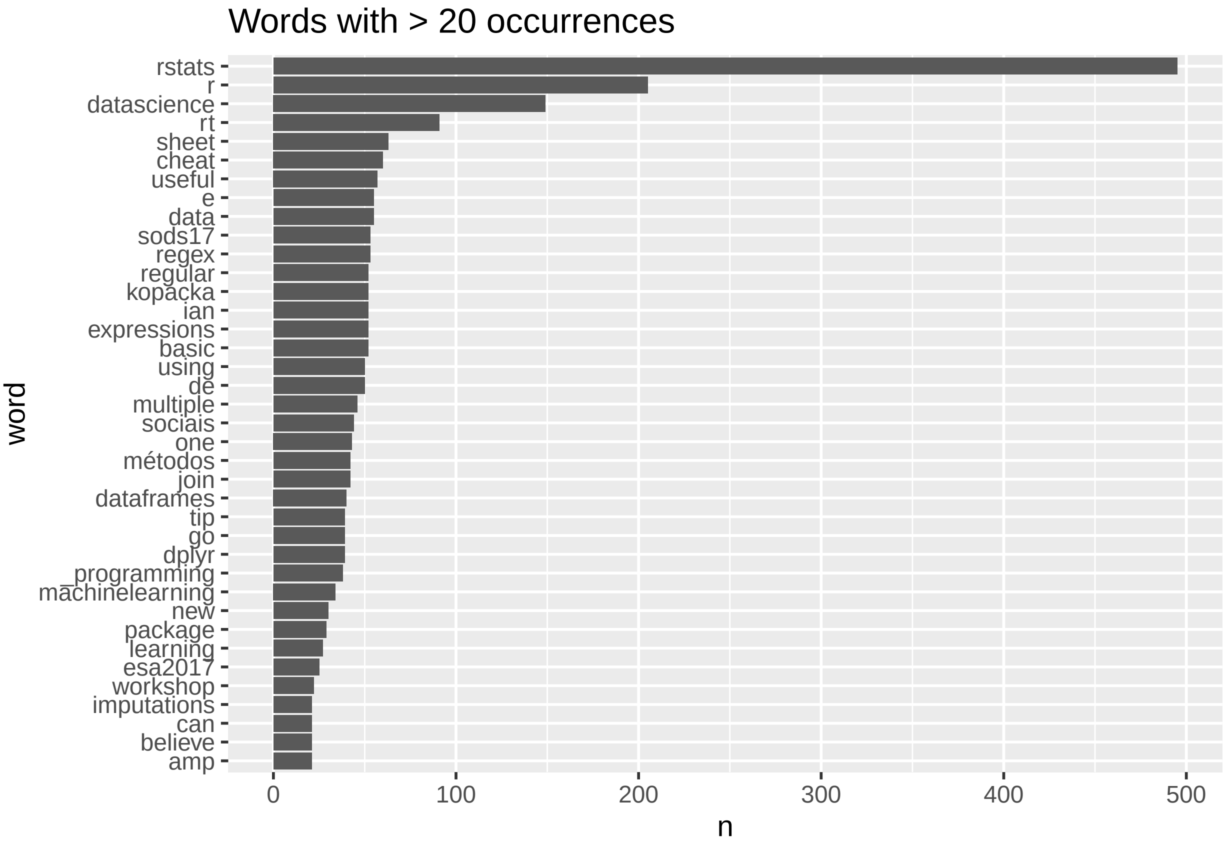
- With table 3, we also plot the most commonly used words:

Table 1: Top 30 words by count

word	count
rstats	495
r	205
datascience	149
rt	91
sheet	63
cheat	60
useful	57
data	55
e	55
regex	53
sods17	53
basic	52
expressions	52
ian	52
kopacka	52
regular	52
de	50
using	50
multiple	46
sociais	44
one	43
join	42
métodos	42
dataframes	40
dplyr	39
go	39
tip	39
_programming	38
machinelearning	34
new	30

Table 2: Most Popular Tweets by Favorite Count

text
R Tip: How to join multiple dataframes in one go using dplyr. #rstats #DataScience https://t.co/jHVyMbTDGb
This -writing guide makes me think I could write mine own... ”Writing an R Package” by @jalapic... https://t.co/CMu
If you couldn’t make it to @noamross &amp; my #mgcv #rstats workshop at #ESA2017 yesterday, all the materials are
I, and all other SPSS users, love watching the Base-R wing of the #Rstats community arguing with the ggplot ”Corporate
We wrote a short history of spatial capabilities in R - https://t.co/011PdJxwH1. Comments/PRs welcome! #rstats... http
More on “The Part-Time R-User” https://t.co/EoluYigxy3 #rstats #DataScience
ICYMI ! @RStudioJoe outlines s (&amp; by category, too): ”June 2017 New Package Picks” https://t.co/ZAyMNJ
All materials for @naupakaz &amp; my intro #vegan #rstats workshop at #ESA2017 today are freely available https://t.
so I created a code that generates a code that generates an analysis report #rstats #rmarkdown https://t.co/bEx48nfL
Get started with the #Jupyter and #Rstats Notebooks with this tutorial - https://t.co/meYB3fQAic #DataScience https



0.5 Further Popularity Analysis

- Further popularity analysis can be performed with favorite and retweet counts.
- Table 3 shows the top 10 tweets, by number of favorits
- Lastly, Table 4 shows top 10 tweets, by number of retweets

Table 3: Most Popular Tweets by RT Count

text
RT @ClausWilke: Over the years, movies have converged to a length of ~100 min. 4 lines of code with ggjoy. #rstats http
RT @danielphadley: Add logos and gifs to plots in #rstats : https://t.co/i40eL4EbP3, or, Vincent Vega explains Cars htt
RT @Rbloggers: ggplot2 – Easy way to mix multiple graphs on the same page https://t.co/WbeK0Eg8C1 #rstats #Data
RT @dataandme: useful cheat sheet: ”Basic Regular Expressions in R” by Ian Kopacka https://t.co/q2AlmRjOnp #Reg
RT @tjpalanca: ”The point being that media isn’t biased in that your timeline is.” #rstats #databeersmnl Full article: ht
RT @Rbloggers: Machine Learning Explained: supervised learning, unsupervised learning, and reinforcement https://t.co/
RT @rOpenSci: [blog] Announcing the rOpenSci Fellowships Program https://t.co/4lgCMUR0yQ Application deadline Se
RT @dsquintana: New post: An #Rstats script to calculate statistical power for a random-effects meta-analysis https://t.
RT @R__Programming: New Grand Test added to ’Learn R By Intensive Practice’ video course #rstats https://t.co/EnY
RT @R__Programming: R Tip: How to join multiple dataframes in one go using dplyr. #rstats #DataScience https://t.co