**Dictionary of functions for text analysis with R, in the order they appear**

**Day 2. Exercises 1-20**

|  |  |
| --- | --- |
| **group\_by(title)** | groups the data by title  **Package: dplyr** |
| **%>%** | Pipe operator. Allows you to pass data through multiple functions. A more efficient way to write code.  **Package: magrittr (used in dplyr)** |
| **unnest\_tokens(word, text)** | Split a column into tokens using the tokenizers package, splitting the table into one-token-per-row.  **Package: tidytext** |
| **anti\_join(stop\_words)** | Removes common words (a, an, the, etc) from the data set  **Package: tidytext** |
| **ungroup()** | Removes grouping of the data  **Package: dplyr** |
| **count(word, sort=TRUE)** | Does a word count  **Package: dplyr** |
| **filter()** | Filters the data set  **Package: dplyr** |
| **mutate(word= reorder(word, n)** | Creates a new variable (column), in this case, “word”. Also, reorder() sorts the data on the new column by count (n)  **Package: dplyr** |
| **ggplot(aes(word, n)) + geom\_col() + xlab(NULL) + coord\_flip()** | Creates a column chart with word on the X axis, and count on the Y axis, but flips X and Y so columns are horizontal  **Package: ggplot2** |
| **summarize(total = sum(n))** | There are two functions here. Summarize() produces a summary statistic. total = sum(n) adds up all the n values and assigns that to total. So the summary statistic is on the total variable  **Package: dplyr** |
| **left\_join(freq\_words, total\_words)** |  |
| **bind\_tf\_idf(word, title, n)** |  |
| **select(-total)** | **removes total from data set** |
| **arrange(desc(tf\_idf)** | sorts the data on tf\_idf column, in descending order |

creates a tf and idf ratio for each word, adds tf and idf as columns

**) —**