Exercise 10.1 Strings and Text Mining

Strings Processing

Dataset: jnlactive.csv

The dataset lists all active journal titles published by Elsevier in 2019 and can also be found in their website.

Hint: str_remove() and str_split() from stringr package may be helpful to perform the analysis below.

1. What is the distribution of words[1] in a journal title?
2. What is the longest number of words in a journal title?
3. What is the name of the journal with the longest number of words in title?
4. What is the number of words in the entire collection of Journal Titles?
5. What is the number of unique words in the entire collection of Journal Title?
6. What are the top 20 most common words in the Journal Title collection?

Instructor Solution to be uploaded after class: journaltitle.R

Text Mining

7. Execute all the journal title analysis listed above using Text Mining approach (quanteda Rpackage). What's the difference compared to the string processing approach? Which approach is easier to you?

   Instructor solution to be uploaded after class: journaltitle2.R

[1] Number of words in title. e.g. How many 1 word title, 2 word title, 3 word title, …etc?