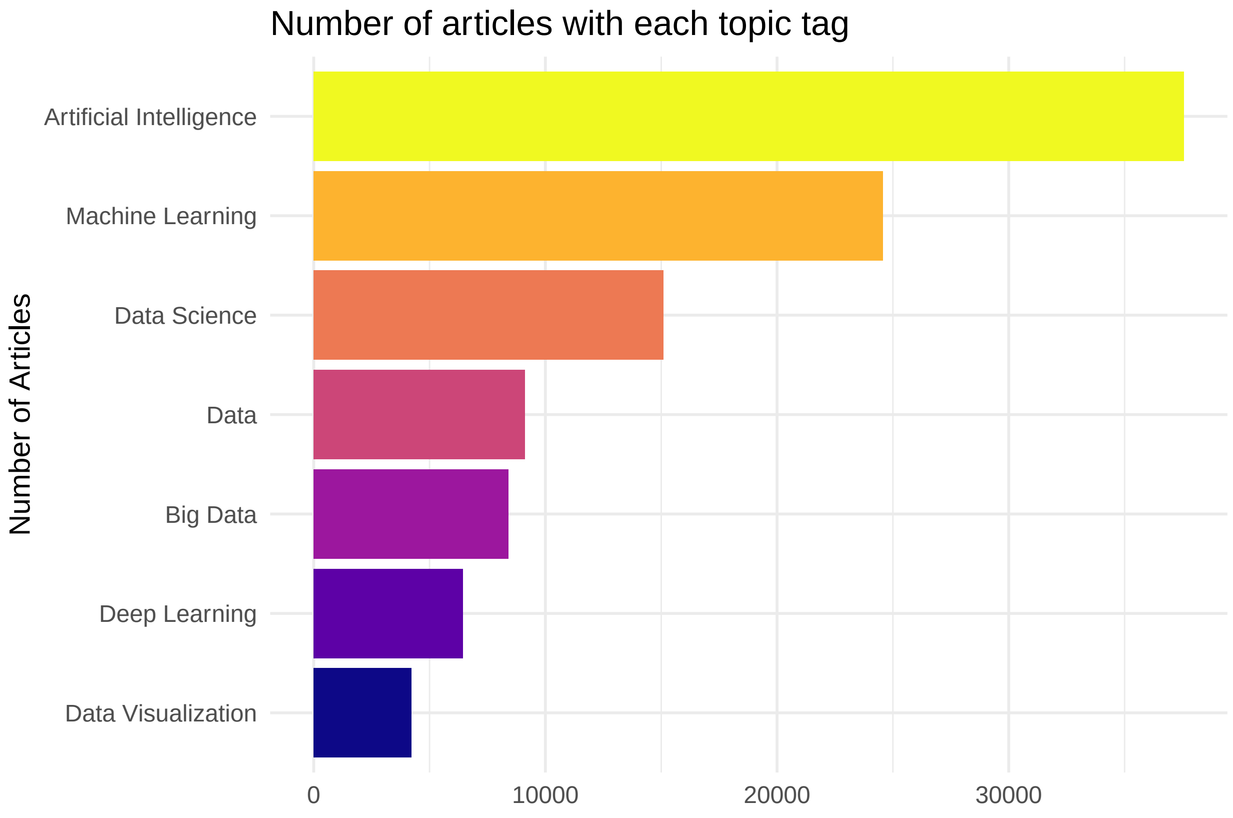
Medium Tidytext

Zachary Hamilton

03/12/2018

Using the slice function, only 1 row for each unique combination of title, subtitle, date, tag was kept. It appears that multiple copies of this same combination were either inadvertent duplicates of a story, or multiple parts of the same story. The most common tags for stories in this dataset are as follows:

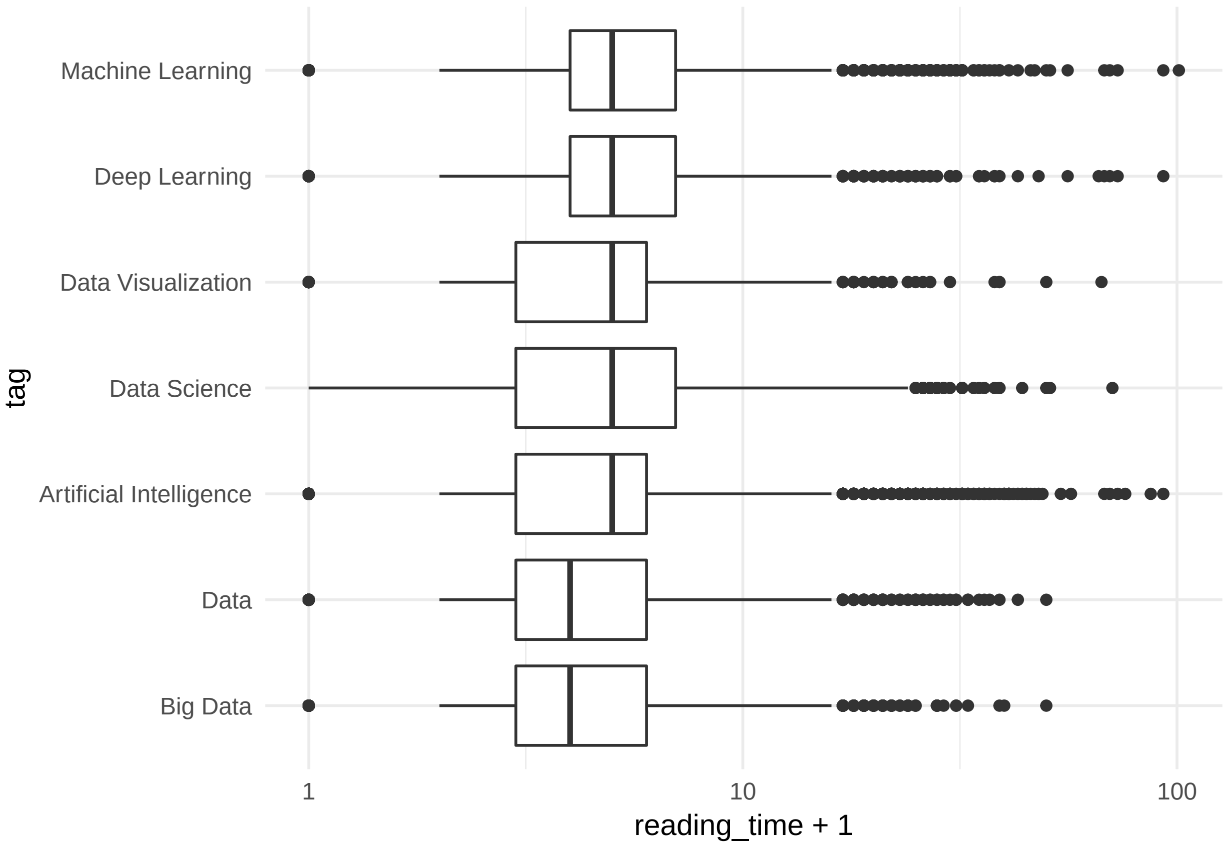
tag	n
Artificial Intelligence	37559
Machine Learning	24568
Data Science	15088
Data	9120
Big Data	8402
Deep Learning	6448
Data Visualization	4229

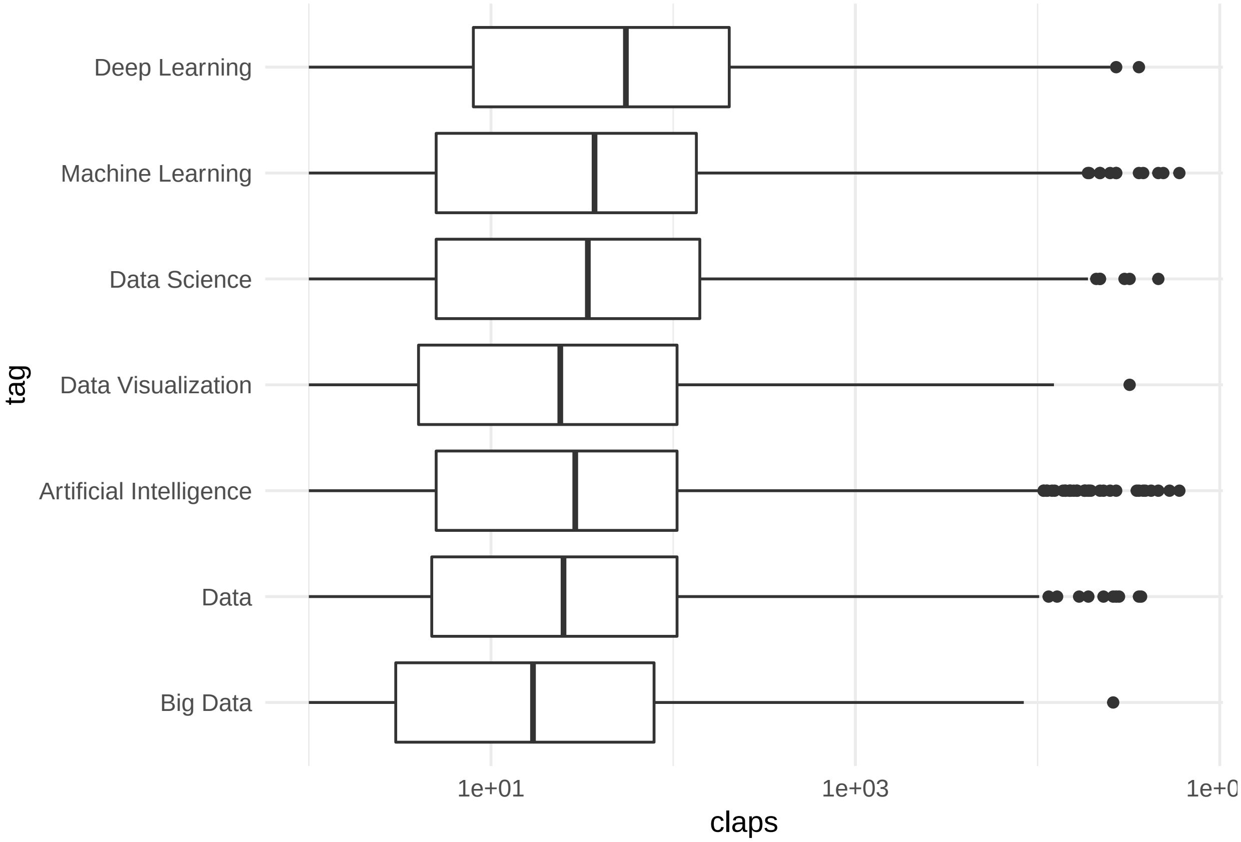


Number of Tags	Articles
1	53273
2	16649
3	5134
4	819
5	33

As you can see, many stories have multiple "tags" relating to the article topic. We might do some sort of correlation to see which tags are most often found with each other, or some sort of network analysis. For the purposes of tidy data and to allow for comparisons of key features across the different topics, each separate tag for a story has been given a separate row using the `gather` function.

First, lets have a look at the "reading time" and "clap" variables to see how these compared across the different topics.





No perceivable differences in reading time or claps across the different tags.

Upon inspecting the associated urls, it appears that articles given the title ".", ":" or "-" in this dataset are either written in a language other than english such as mandarin or arabic, or contain symbols that cannot be coded as simple text i.e. trademarkTM. As such they will not be discarded from the dataset.