Summary - Data clean-up

For this text mining and sentiment analysis project, the body of the tweet (“text”) and user account (“user\_id”) variables comprised the data set of AI related tweets. The Rtweet package allowed selecting only tweets which contained “artificial intelligence” or “AI” in the body of the text field. The project is to measure sentiments in the public discourse about AI. Thus, the assumption is that the general Tweeting public probably would not use more technical terms, like: machine learning and supervised or unsupervised learning.

The data set required minimal clean-up beyond selecting Tweets only English and omitting hash marks. The Data Story project milestone (Rmd file) describes in detail and provides code for this. It also explains words to be filtered and the word clouds used to prepare the “text” character strings for sentiment analysis.

The data set is in the AItwtdata.Rds file accompanying the final report.