Medium Articles Analysis

### Read the dataset

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.1

## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.3.0 v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.5.1

## Warning: package 'tibble' was built under R version 3.5.1

## Warning: package 'tidyr' was built under R version 3.5.1

## Warning: package 'readr' was built under R version 3.5.1

## Warning: package 'purrr' was built under R version 3.5.1

## Warning: package 'dplyr' was built under R version 3.5.1

## Warning: package 'stringr' was built under R version 3.5.1

## Warning: package 'forcats' was built under R version 3.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

medium\_dataset\_csv <- read\_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2018/2018-12-04/medium\_datasci.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## title = col\_character(),  
## subtitle = col\_character(),  
## author = col\_character(),  
## publication = col\_character(),  
## url = col\_character(),  
## author\_url = col\_character()  
## )

## See spec(...) for full column specifications.

medium\_dataset\_processed <- (medium\_dataset\_csv) %>% select(-x1)

### display the count of top 10 publication

medium\_dataset\_processed %>% count(publication, sort = TRUE) %>% head(10)

## Warning: package 'bindrcpp' was built under R version 3.5.1

## # A tibble: 10 x 2  
## publication n  
## <chr> <int>  
## 1 <NA> 44072  
## 2 Towards Data Science 3135  
## 3 Hacker Noon 1178  
## 4 Becoming Human: Artificial Intelligence Magazine 671  
## 5 Chatbots Life 450  
## 6 Data Driven Investor 351  
## 7 Chatbots Magazine 350  
## 8 SyncedReview 320  
## 9 Planeta Chatbot todo sobre los Chatbots y la Inteligencia Artific~ 294  
## 10 DEEPAERODRONES 273

### Top 10 most popular author

medium\_dataset\_processed %>% count(author, sort = TRUE) %>% head(10)

## # A tibble: 10 x 2  
## author n  
## <chr> <int>  
## 1 Yves Mulkers 487  
## 2 Synced 328  
## 3 <NA> 286  
## 4 DEEP AERO DRONES 279  
## 5 AI Hawk 245  
## 6 Peter Marshall 209  
## 7 Ilexa Yardley 162  
## 8 Alibaba Cloud 160  
## 9 Corsair's Publishing 156  
## 10 Jae Duk Seo 150

### aggregate the tags & find the most common tags

medium\_dataset\_processed %>% summarize\_at(vars(starts\_with("tag\_")),sum)

## # A tibble: 1 x 8  
## tag\_ai tag\_artificial\_~ tag\_big\_data tag\_data tag\_data\_science  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 13763 29580 8686 9420 15424  
## # ... with 3 more variables: tag\_data\_visualization <dbl>,  
## # tag\_deep\_learning <dbl>, tag\_machine\_learning <dbl>

### reshaping the above data

medium\_gathered <-medium\_dataset\_processed %>%   
 gather(tag,value,starts\_with("tag\_")) %>%  
 mutate(tag = str\_remove(tag,"tag\_"))%>%   
 filter(value == 1)  
  
medium\_gathered %>% View  
medium\_gathered %>% count(tag,sort = TRUE)

## # A tibble: 8 x 2  
## tag n  
## <chr> <int>  
## 1 artificial\_intelligence 29580  
## 2 machine\_learning 25175  
## 3 data\_science 15424  
## 4 ai 13763  
## 5 data 9420  
## 6 big\_data 8686  
## 7 deep\_learning 6612  
## 8 data\_visualization 4334

### find the most popular tag based on the number of claps

medium\_gathered %>%   
 group\_by(tag) %>%  
 summarize(median\_claps = median(claps)) %>%  
 arrange(desc(median\_claps))

## # A tibble: 8 x 2  
## tag median\_claps  
## <chr> <dbl>  
## 1 deep\_learning 25  
## 2 machine\_learning 12  
## 3 data\_science 11  
## 4 artificial\_intelligence 5  
## 5 data\_visualization 5  
## 6 ai 4  
## 7 data 3  
## 8 big\_data 1

### find the average number of claps for an article in towards data science

medium\_gathered %>%   
 group\_by(publication) %>%  
 filter(publication== "Towards Data Science") %>%  
 summarize(median\_claps = median(claps)) %>%  
 arrange(desc(median\_claps))

## # A tibble: 1 x 2  
## publication median\_claps  
## <chr> <dbl>  
## 1 Towards Data Science 215

### find the most popular tag based on the number of claps

medium\_gathered %>%  
 group\_by(tag) %>%  
 summarize(median\_claps = median(claps)) %>%  
 arrange(desc(median\_claps))

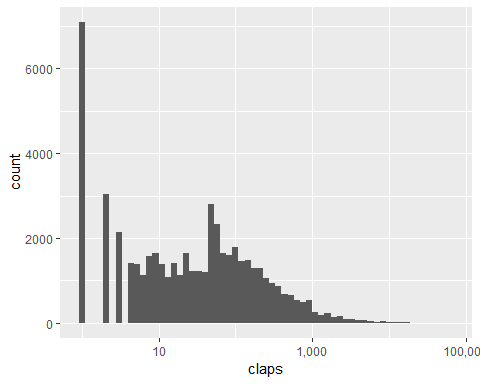
## # A tibble: 8 x 2  
## tag median\_claps  
## <chr> <dbl>  
## 1 deep\_learning 25  
## 2 machine\_learning 12  
## 3 data\_science 11  
## 4 artificial\_intelligence 5  
## 5 data\_visualization 5  
## 6 ai 4  
## 7 data 3  
## 8 big\_data 1

### distribution of claps

medium\_dataset\_processed %>%  
 ggplot(aes(claps)) +  
 geom\_histogram(binwidth = 0.08) +  
 scale\_x\_log10(labels = scales::comma\_format())

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 25729 rows containing non-finite values (stat\_bin).

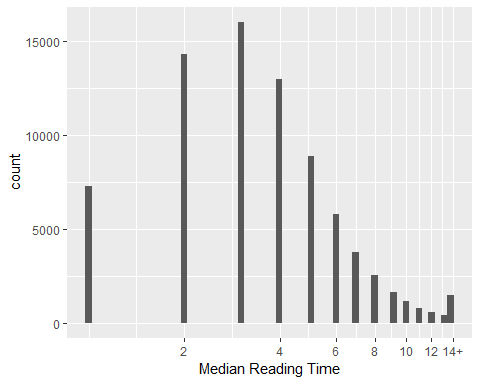


### distribution of reading time

medium\_dataset\_processed %>%  
 mutate(reading\_time = pmin(14,reading\_time)) %>%  
 ggplot(aes(reading\_time)) +  
 geom\_histogram(binwidth = 0.02) +  
 scale\_x\_log10(breaks =seq(2,14,2),  
 labels = c(seq(2,12,2),"14+")) +  
 labs(x = "Median Reading Time")

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 814 rows containing non-finite values (stat\_bin).



Q1 <- c(10,3,7,9)  
Q2 <- c(9,4,6,4)  
  
pmax(Q1,Q2) #parallel maxima

## [1] 10 4 7 9

### reading time associated with various tags

medium\_gathered %>%  
 group\_by(tag) %>%  
 summarize(reading\_time = mean(reading\_time)) %>%  
 arrange(desc(reading\_time))

## # A tibble: 8 x 2  
## tag reading\_time  
## <chr> <dbl>  
## 1 deep\_learning 5.08  
## 2 machine\_learning 4.76  
## 3 data\_science 4.69  
## 4 artificial\_intelligence 4.36  
## 5 data\_visualization 4.26  
## 6 ai 4.14  
## 7 data 4.05  
## 8 big\_data 3.97

## Text Mining

library(tidytext)

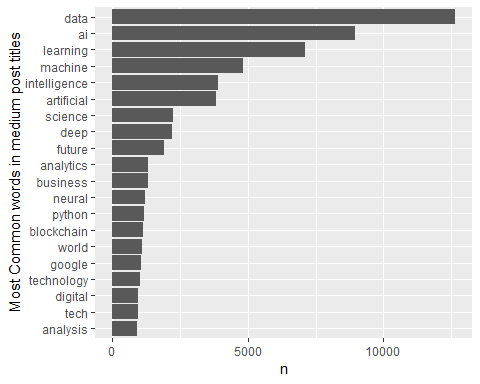
## Warning: package 'tidytext' was built under R version 3.5.2

words <- medium\_dataset\_processed %>%   
   
 filter(!is.na(title)) %>% #removing missing data  
 unnest\_tokens(word,title) %>% # tokenizing  
 anti\_join(stop\_words, by= "word") %>% # removing stop words  
 transmute(post\_id = row\_number(),subtitle,year,reading\_time,claps,word) %>%  
 filter(word!= 'de',str\_detect(word,'[a-z]'))   
   
 class(words)

## [1] "tbl\_df" "tbl" "data.frame"

### Most Common words in medium post titles

words %>%  
 count(word,sort = TRUE) %>%  
 head(20) %>%  
 mutate(word = fct\_reorder(word,n)) %>%  
 ggplot(aes(word,n)) +  
 geom\_col() +  
 coord\_flip()+  
 xlab("Most Common words in medium post titles")



words %>%  
 add\_count(word)%>% #adds a column with count of the word  
 filter(n>500) %>%  
 group\_by(word) %>%  
 summarize(median\_claps = median(claps),  
 geometric\_mean\_claps = exp(mean(log(claps +1))),  
 occurences = n()) %>%  
 arrange(desc(median\_claps))

## # A tibble: 49 x 4  
## word median\_claps geometric\_mean\_claps occurences  
## <chr> <dbl> <dbl> <int>  
## 1 tensorflow 53 36.6 704  
## 2 building 36 27.0 669  
## 3 deep 35 27.7 2194  
## 4 guide 35 27.6 569  
## 5 networks 24 25.1 764  
## 6 blockchain 23 19.2 1123  
## 7 model 23 19.2 623  
## 8 neural 22 23.3 1199  
## 9 network 21 20.8 759  
## 10 build 20 20.1 536  
## # ... with 39 more rows

The words that got the most claps were ‘tensorflow,building,guide,deep learning, neural networks’ and the words that got the least no of claps were ‘marketing, trends’ and so on

rm(medium\_dataset\_processed)  
rm(medium\_gathered)  
rm(medium\_dataset\_csv)  
rm(Q1)   
rm(Q2)  
gc()

## used (Mb) gc trigger (Mb) max used (Mb)  
## Ncells 2269588 121.3 4172378 222.9 4172378 222.9  
## Vcells 10337771 78.9 32229094 245.9 35841690 273.5