Finalproject.R

Admin

2020-03-12

library(readxl)  
library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

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

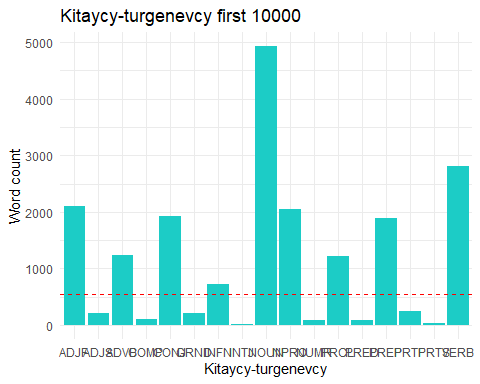
library(dplyr)  
x <- read\_excel("части речи по 10000.xlsx")  
View(x)  
x <- x%>%  
 fill(Corpus\_name)%>%  
 fill(ID)%>%  
 mutate(POS\_count = as.integer(POS\_number))%>%  
 select (-c(POS\_number))%>%  
 mutate(POS\_count\_expected = 1/18\*10000)

## Warning: в результате преобразования созданы NA

kitaycy <- x%>%  
 filter(grepl("^kitaycy", Corpus\_name)&ID==30)%>%  
 arrange(POS\_name)  
kitaycy %>%  
 ggplot(., aes(x=POS\_name, y=POS\_count)) +  
 geom\_histogram(stat='identity', fill="#1CCCC6") +   
 ylab("Word count") + xlab("Kitaycy-turgenevcy") +  
 ggtitle("Kitaycy-turgenevcy first 10000") +   
 geom\_hline(yintercept = kitaycy$POS\_count\_expected, color = "red",lty = 2) +  
 theme\_minimal()

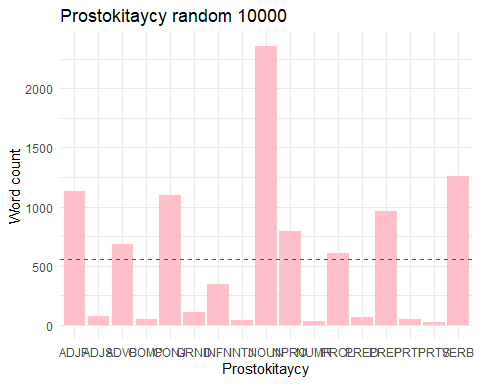
## Warning: Ignoring unknown parameters: binwidth, bins, pad

## Warning: Removed 1 rows containing missing values (position\_stack).



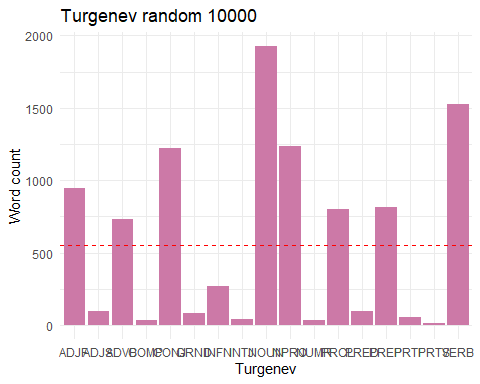
prostokitaycy <- x%>%  
 filter(grepl("^prostokitaycy", Corpus\_name)&ID==30)%>%  
 arrange(POS\_name)  
prostokitaycy %>%  
 ggplot(., aes(x=POS\_name, y=POS\_count)) +  
 geom\_histogram(stat='identity', fill="pink") +   
 ylab("Word count") + xlab("Prostokitaycy") +  
 ggtitle("Prostokitaycy random 10000") +   
 geom\_hline(yintercept = kitaycy$POS\_count\_expected, color = "red",lty = 2) +  
 theme\_minimal()

## Warning: Ignoring unknown parameters: binwidth, bins, pad



turgenev <- x%>%  
 filter(grepl("^turgenev", Corpus\_name)&ID==30)%>%  
 arrange(POS\_name)  
turgenev %>%  
 ggplot(., aes(x=POS\_name, y=POS\_count)) +  
 geom\_histogram(stat='identity', fill="#CC79A7") +   
 ylab("Word count") + xlab("Turgenev") +  
 ggtitle("Turgenev random 10000") +   
 geom\_hline(yintercept = kitaycy$POS\_count\_expected, color = "red",lty = 2) +  
 theme\_minimal()

## Warning: Ignoring unknown parameters: binwidth, bins, pad



ADJF <- x%>%  
 filter((grepl("ADJF", POS\_name)))  
  
x1kitturg <- ADJF%>%  
 filter((grepl("^kitaycy", Corpus\_name)))  
  
x1kitNEturg <- ADJF%>%  
 filter((grepl("^prostokitaycy", Corpus\_name)))  
  
table(x1kitturg$POS\_count,x1kitNEturg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1221, df = 1190, p-value = 0.2599  
## Chi-squared approximation may be incorrect

CONJ <- x%>%  
 filter((grepl("CONJ", POS\_name)))  
  
x2kitturg <- CONJ%>%  
 filter((grepl("^kitaycy", Corpus\_name)))  
  
x2kitNEturg <- CONJ%>%  
 filter((grepl("^prostokitaycy", Corpus\_name)))  
  
table(x2kitturg$POS\_count,x2kitNEturg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1073, df = 1054, p-value = 0.3351  
## Chi-squared approximation may be incorrect

VERB <- x%>%  
 filter((grepl("VERB", POS\_name)))  
  
x3kitturg <- VERB%>%  
 filter((grepl("^kitaycy", Corpus\_name)))  
  
x3kitNEturg <- VERB%>%  
 filter((grepl("^prostokitaycy", Corpus\_name)))  
  
table(x3kitturg$POS\_count,x3kitNEturg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1332, df = 1296, p-value = 0.2377  
## Chi-squared approximation may be incorrect

NOUN <- x%>%  
 filter((grepl("NOUN", POS\_name)))  
  
x4kitturg <- NOUN%>%  
 filter((grepl("^kitaycy", Corpus\_name)))  
  
x4kitNEturg <- NOUN%>%  
 filter((grepl("^prostokitaycy", Corpus\_name)))  
  
table(x4kitturg$POS\_count,x4kitNEturg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1258, df = 1225, p-value = 0.2501  
## Chi-squared approximation may be incorrect

x1turg <- ADJF%>%  
 filter((grepl("^turgenev", Corpus\_name)))  
  
table(x1kitturg$POS\_count,x1turg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1165.5, df = 1122, p-value = 0.1786  
## Chi-squared approximation may be incorrect

x2turg <- CONJ%>%  
 filter((grepl("^turgenev", Corpus\_name)))  
  
table(x2kitturg$POS\_count,x2turg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1184, df = 1156, p-value = 0.2771  
## Chi-squared approximation may be incorrect

x3turg <- VERB%>%  
 filter((grepl("^turgenev", Corpus\_name)))  
  
table(x3kitturg$POS\_count,x3turg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1332, df = 1296, p-value = 0.2377  
## Chi-squared approximation may be incorrect

x4turg <- NOUN%>%  
 filter((grepl("^turgenev", Corpus\_name)))  
  
table(x4kitturg$POS\_count,x4turg$POS\_count)%>%  
 summary()

## Number of cases in table: 37   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 1267.2, df = 1225, p-value = 0.1956  
## Chi-squared approximation may be incorrect