Programowanie w R wina

Adam Matuszczyk_MSAD 2018/19

11 07 2019

Praca zaliczeniowa z "Programowania w R" zajęć prowadzonych w ramach MSAD 2018/19

__na podstawie analiz prowadzonych na zajęciach przez Artura Machno oraz mapa świata wg. win wykonanana na podstawie wpisu, na portalu kaggle.com przez uczestnika o loginie **Pozdniakov**

```
library("tidyverse")
library("tidytext")
library("data.table")
library("plotly")
library("magrittr")
library("DT")
library("ggjoy")
```

Przeprowadzić analizę danych tekstowych w opisach win względem kraju pochodzenia. Proponuje podzielić kraje na 4 kategorie: US, France, Italy, Other.

```
wine<-fread("d:/AGH/Programowanie w R/wine.csv", header  = TRUE, sep = ",")


wine <- wine[!is.na(price),][!country == "",] #chce usunąć wszystkie puste -bez informacji o kraju, cenach i punktach
```

MAPA WIN

```
m <- as.data.table(map_data("world"))
#unique(wine$country)[!(unique(wine$country) %in% unique(m$region))]
#m[region == "US", region := "USA"]
wine[country == "US", country := "USA"]
m2 <- merge(m,
            wine[,.(.N,
                    points = median(points, na.rm = T),
                    price = median(price, na.rm = T)),
                 by = country],
            by.x = "region",
            by.y = 'country',
            all.x = T,
            all.y = F,
            sort = F)

m2 <- m2[order(m2$order),]
m2[is.na(N), N:=0]
m2[,text:=sprintf("%s: %.0f wines <br>Median points: %.0f <br>Median price: %.0f$", region, N, points, price
)]

g <- ggplot(m2, aes(text = text))+
  geom_polygon(aes(long, lat, group = group, fill = N))+
  coord_equal()+
  scale_fill_gradient(low = '#c994c7', high = "#dd1c77", trans = "log", na.value = "#c994c7", breaks = c(0,
1, 10, 100, 1000, 10000))+
  theme_void()


gg <- ggplotly(g, tooltip = "text")
gg
```

Przeprowadzić analizę różnicy występowania słów w opisach w zależności od kraju pochodzenia wina Proponuje podzielić kraje na 4 kategorie: US, France, Italy, Other

Przeprowadzić analizę emocji w opisach win względem kraju pochodzenia Proponuje podzielić kraje na 4 kategorie: US, France, Italy, Other

porządkowanie danych

```
wine$country <- as.factor(wine$country)
fct_count(wine$country, sort = TRUE, prop = FALSE)
```

```
## # A tibble: 42 x 2
##    f             n
##    <fct>     <int>
##  1 USA       54265
##  2 France    17776
##  3 Italy     16914
##  4 Spain      6573
##  5 Portugal   4875
##  6 Chile      4416
##  7 Argentina  3756
##  8 Austria    2799
##  9 Australia  2294
## 10 Germany    2120
## # ... with 32 more rows
```

```
wine$country <- fct_lump(wine$country, 3)
fct_count(wine$country, sort = TRUE, prop = FALSE)
```

```
## # A tibble: 4 x 2
##   f          n
##   <fct>  <int>
## 1 USA    54265
## 2 Other  31961
## 3 France 17776
## 4 Italy  16914
```

uporządkowanie danych do analizy dalszej

```
wine_spy <- wine %>% select( V1, country, description ) %>% unnest_tokens(word, description)#tokenizacja
```

usuwanie zbędnych elementów w ramce danych

```
#wszystkie słowa
count(wine_spy, word, sort = TRUE)
```

```
## # A tibble: 34,201 x 2
##    word         n
##    <chr>    <int>
##  1 and     327806
##  2 the     206309
##  3 a       167413
##  4 of      162533
##  5 with    112581
##  6 this    106589
##  7 is       88580
##  8 wine     70197
##  9 flavors  60046
## 10 in       59546
## # ... with 34,191 more rows
```

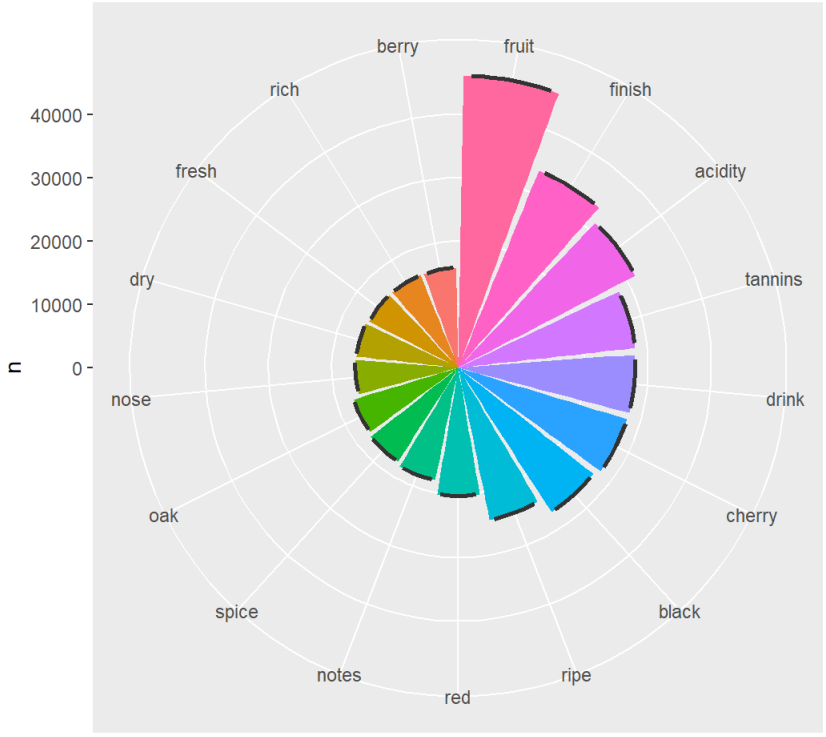
```
#wszystkie słowa bez "stop_words"

count(wine_spy %>%
        anti_join(stop_words), word, sort = TRUE)
```

```
## # A tibble: 33,534 x 2
##    word        n
##    <chr>   <int>
##  1 wine    70197
##  2 flavors 60046
##  3 fruit   46126
##  4 aromas  37445
##  5 palate  36683
##  6 finish  33634
##  7 acidity 31501
##  8 tannins 28095
##  9 drink   27935
## 10 cherry  27815
## # ... with 33,524 more rows
```

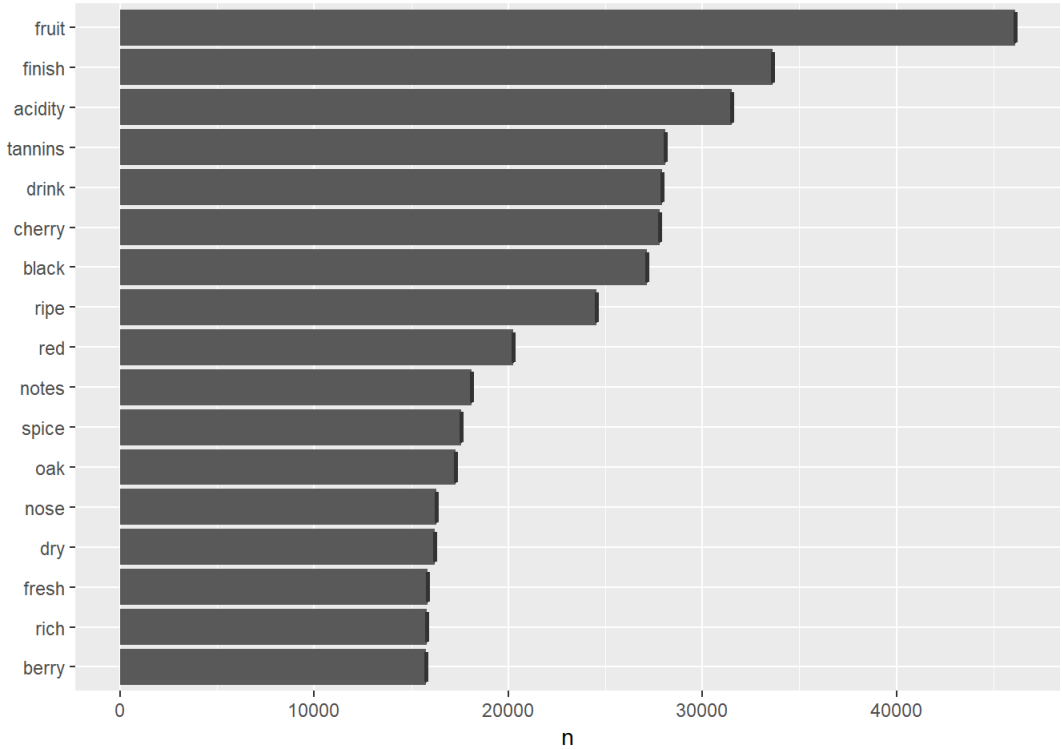
```
#dodawanie elementów zbędnych do stop words
# słowa specyficzne dla win

my_stop_words <- rbind(stop_words, tibble(word = c("wine", "flavors", "aromas", "palate"),lexicon = "my"))

# modyfikujemy ramkę "win_spy"

wine_spy <- wine_spy %>% anti_join(my_stop_words)
```

najczęściej wystepujące słowa

```
wine_spy %>% count(word, sort = T) %>% filter(n > 15000) %>% mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = word, y = n, fill= factor(word))) +
  geom_col(show.legend= FALSE)  +
  xlab(NULL)+
  geom_boxplot(show.legend = FALSE) +
  coord_polar(direction = -1)
```

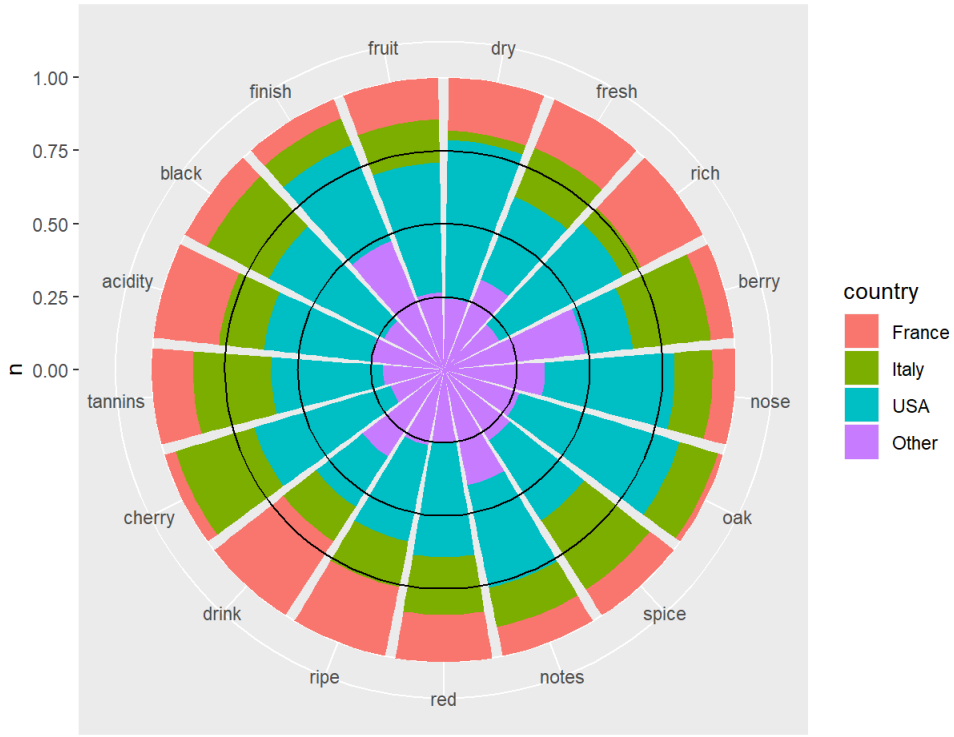



```
wine_spy %>% count(word, sort = T) %>% filter(n > 15000) %>% mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = word, y = n)) +
  geom_col()  +
  xlab(NULL) +
  geom_boxplot()+
  coord_flip()
```

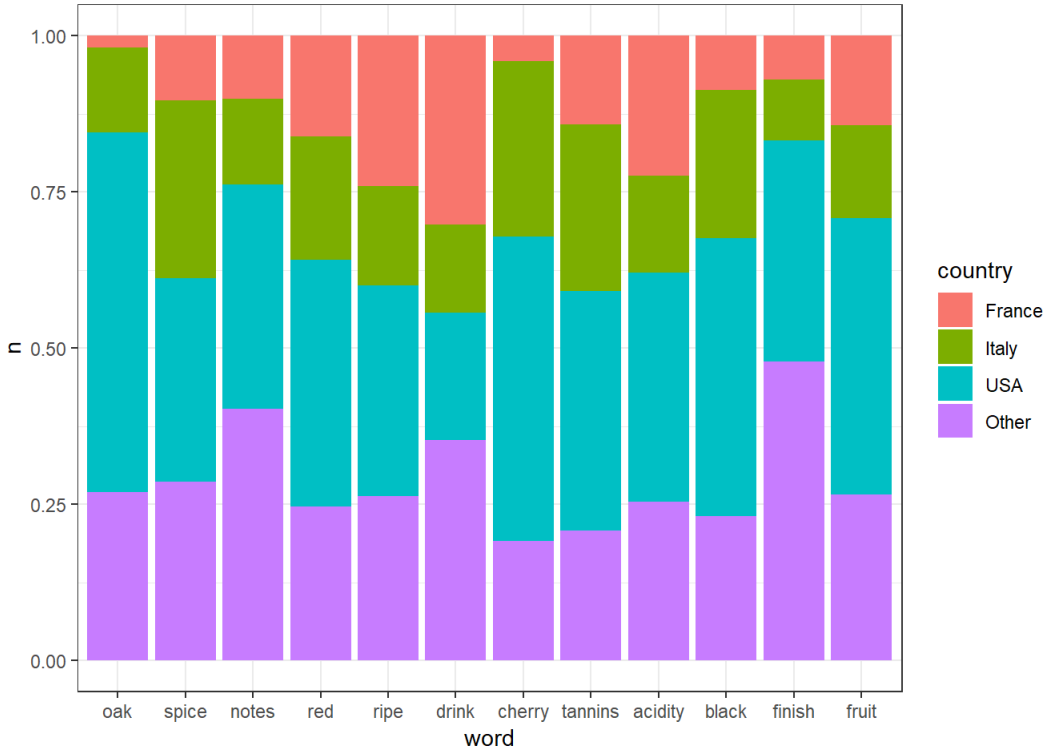


Najczęściej występujące słowa ze względu na kraj pochodzenia wina

```
wine_spy %>%
  left_join(wine_spy %>% select(c(V1, country))) %>%
  count(country, word, sort = T) %>%
  semi_join(wine_spy %>%
              count(word) %>%
              filter(n > 15000),
            by = "word") %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = word, y = n, fill = country)) +
  geom_col(position = "fill") +
  geom_hline(yintercept = 0.25) +
  geom_hline(yintercept = 0.5) +
  geom_hline(yintercept = 0.75) +
  xlab(NULL) +
  coord_polar()
```




```
wine_spy %>%
  left_join(wine_spy %>% select(c(V1, country))) %>%
  count(country, word, sort = T) %>%
  semi_join(wine_spy %>%
              count(word) %>%
              filter(n > 17000),
            by = "word") %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = word, y = n, fill = country)) +
  geom_col(position = "fill") +
  theme_bw()
```



LOUGHRAN

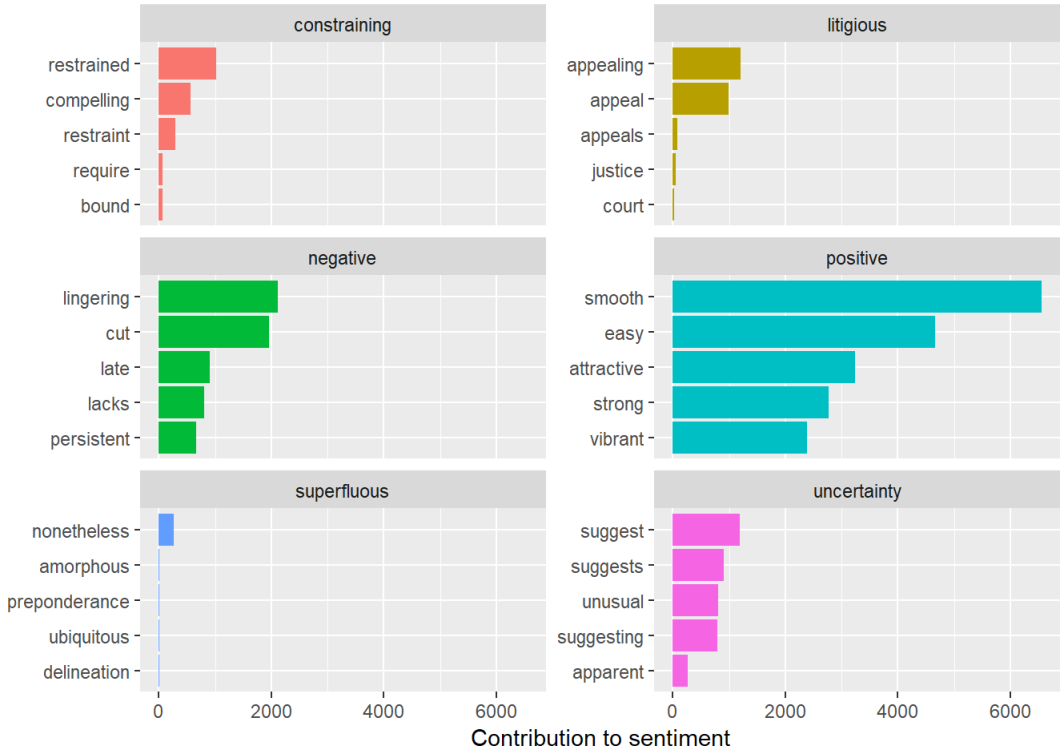
```
loughran<- get_sentiments("loughran")  # niestety słownik "nrc" wyleciał z pakietu "tidytext"
unique(loughran$sentiment)
```

```
## [1] "negative"     "positive"     "uncertainty"  "litigious"
## [5] "constraining" "superfluous"
```

```
#loughran

loughran_df <-
  wine_spy %>%
  left_join(loughran)  %>%
  group_by(V1, sentiment) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  spread(key = sentiment, value = n, fill = 0) %>%
  select(V1:uncertainty)
```

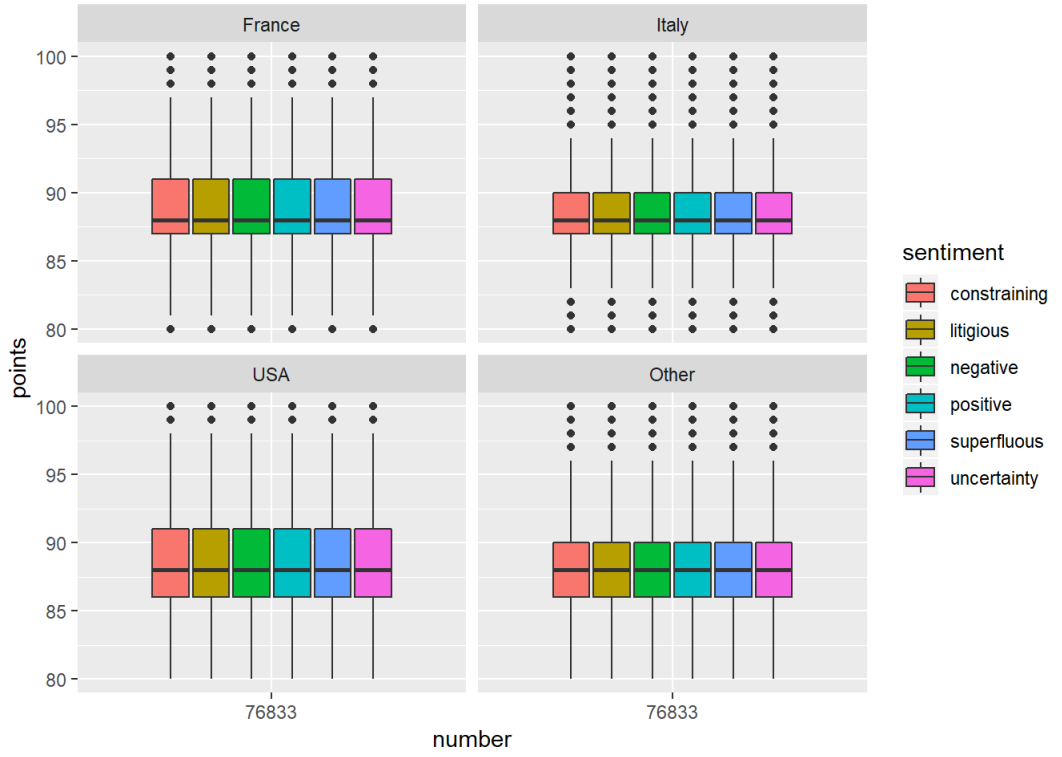
Najpopularniejsze słowa emocjonalne w opisach win

```
wine_spy %>%
  inner_join(loughran) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup() %>%
  group_by(sentiment) %>%
  top_n(5) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y",   ncol = 2) +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```



Związek emocji opisu z oceną wina w podziale na kraje

```
wine %>% left_join(loughran_df, by = "V1") %>%
  gather(key = "sentiment",
         value = "n_sentiment",
         constraining : uncertainty) %>%
  ggplot(aes(x = factor(sum(n_sentiment)),
             y = points,
             fill = sentiment)) +
  geom_boxplot() +
  labs(x = "number") +
  facet_wrap( ~ country, nrow = 2)
```



```
wine %>% left_join(loughran_df, by = "V1") %>%
  gather(key = "sentiment",
         value = "n_sentiment",
         constraining : uncertainty) %>%
  ggplot(aes(x = factor(sentiment),
             y = points,
             fill = sentiment)) +
  geom_boxplot(show.legend = FALSE) +
  labs(x = "number") +
  facet_wrap(. ~ country, nrow = 2)
```