Analiza tweetów

Piotr Smuda

14 marca 2015

Wprowadzenie

Do analizy postanowiłem wziąć tweety dotyczące trwającego w dniach 12-15 marca turnieju gier komputerowych *Intel Extreme Masters Katowice*. Jest to jeden z największych turniejów gier komputerowych, którego pula nagród wynosi ok. 2 mln zł.

Tweety

Przy doborze tweetów kierowałem się jedynie kluczami:

- IEM,
- Katowice,
- ESL,
- INTEL EXTREME MASTERS,

a same tweety zbierałem przez 3 godziny. Udało się zebrać blisko 11 tysięcy w języku angielskim (łącznie 15 tysięcy), z czego po wybraniu unikatowych zostało niecałe 4,5 tysiąca.

Obróbka tweetów

```
setwd("D:/Dokumenty/studia/8 semestr/R i Big Data/lab2")
parsedTweets <- parseTweets("iem.json", simplify = FALSE, verbose = TRUE)
```

```
## 15014 tweets have been parsed.
```

```
ktore<-which(parsedTweets[,"lang"]=="en")
parsedTweets<-parsedTweets[ktore,]
tweets<-parsedTweets[,"text"]
tweets<-stri_trans_tolower(tweets)
tweets<-sapply(tweets,removeWords,stopwords("english"),USE.NAMES=FALSE)
tweets<-sapply(tweets,stri_replace_all_regex,"(http[^ ]+|www.[^ ]+)","",USE.NAMES=FALSE)
tweets<-sapply(tweets,removePunctuation,USE.NAMES=FALSE)
tweets<-sapply(tweets,stri_replace_all_regex,"rt ","",USE.NAMES=FALSE)
tweets<-sapply(tweets,stripWhitespace,USE.NAMES=FALSE)
tweets<-unique(tweets)
```

Który klucz miał największy wpływ

```
iem<-which(stri_detect_regex(tweets,"iem")==TRUE)
esl<-which(stri_detect_regex(tweets,"esl")==TRUE)
katowice<-which(stri_detect_regex(tweets,"katowice")==TRUE)
intelextrememasters<-which(stri_detect_regex(tweets,"intel extreme masters")==TRUE)

klucze<-list(iem,esl,katowice,intelextrememasters)
przeciecie<-matrix(numeric(16),ncol=4)
for(i in 2:4)
{
   for(j in 1:(i-1))
   {
      przeciecie[i,j]<-length(intersect(klucze[[i]],klucze[[j]]))
   }
}
colnames(przeciecie)<-c("IEM","Katowice","ESL","INTEL EXTREME MASTERS")
rownames(przeciecie)<-c("IEM","Katowice","ESL","INTEL EXTREME MASTERS")
(przeciecie<-przeciecie[2:4,1:3])
```

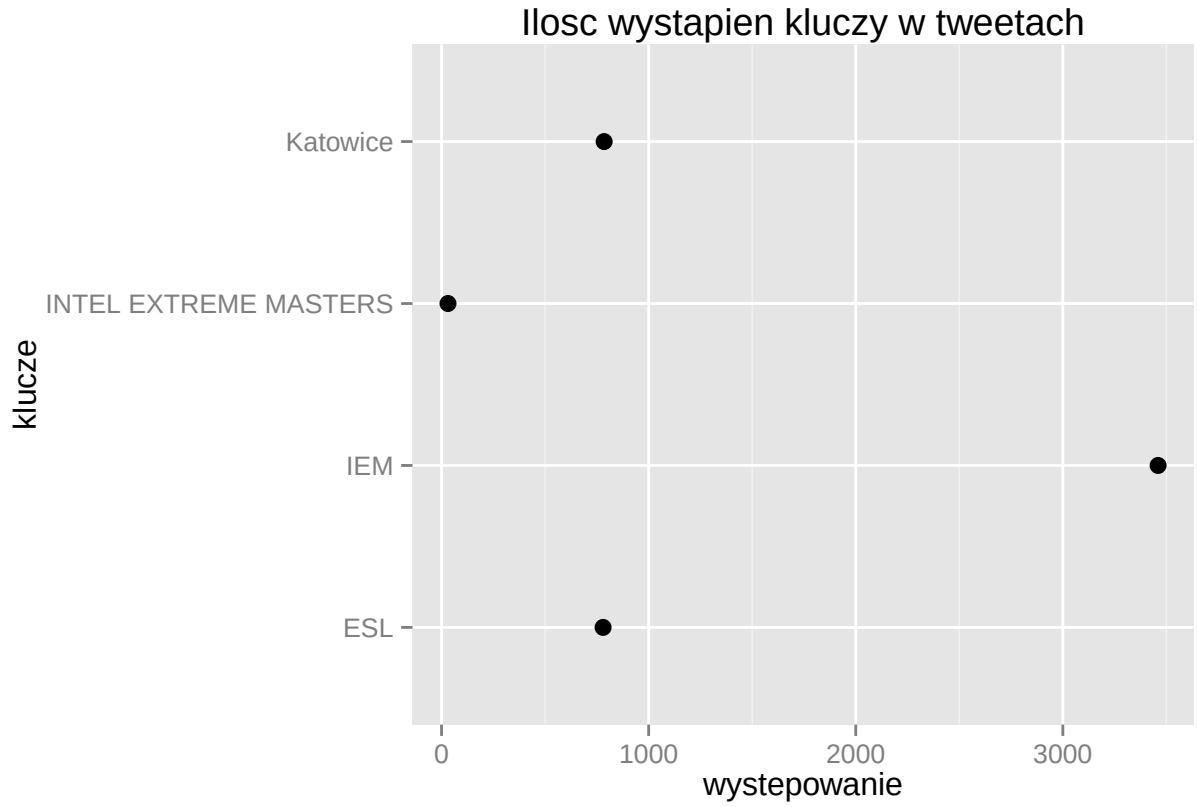
```
                     IEM Katowice ESL
Katowice             114        0   0
ESL                  428      248   0
INTEL EXTREME MASTERS  11        2   5
```

Powyższa macierz przedstawia w ilu tweetach pojawiały się na raz pary kolejnych kluczy.

```
liem<-length(iem)
lesl<-length(esl)
lkatowice<-length(katowice)
lintelextrememasters<-length(intelextrememasters)

licznosci<-data.frame(klucze=c("IEM","Katowice","ESL","INTEL EXTREME MASTERS"),
   wystepowanie=c(liem,lesl,lkatowice,lintelextrememasters))
licznosci<-licznosci%>%
   group_by(klucze)

ggplot(licznosci,aes(x=klucze, y=wystepowanie)) +
   geom_point(size=3) +
   coord_flip() +
   ggtitle("Ilosc wystapien kluczy w tweetach")
```

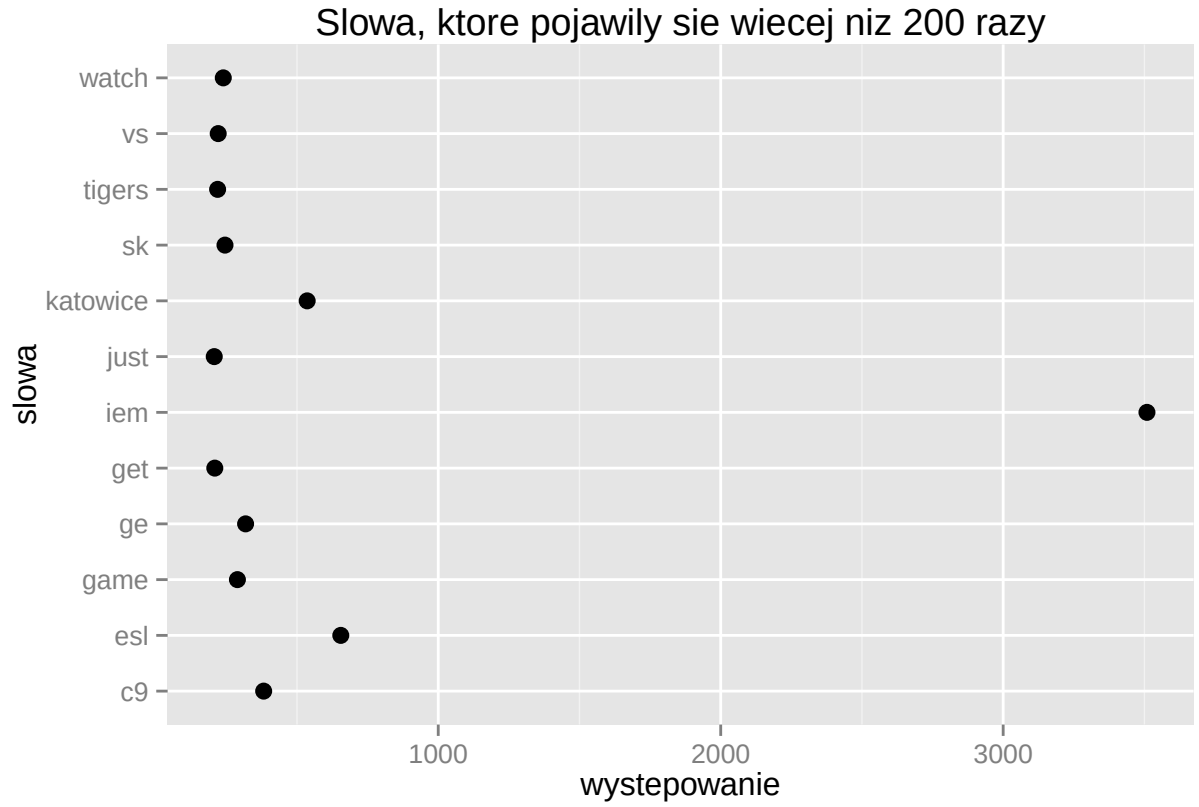


Najczęstsze słowa

```
words<-table(unlist(stri_extract_all_words(tweets)))

slowa<-data.frame(slowa=names(words[words>200]),
   wystepowanie=words[words>200])

ggplot(slowa,aes(x=slowa, y=wystepowanie)) +
   geom_point(size=3) +
   coord_flip() +
   ggtitle("Slowa, ktore pojawily sie wiecej niz 200 razy")
```



Słowa składające się z dwóch znaków w tym przypadku mają rację bytu, ponieważ są to skróty od nazw zespołów biorących udział w turnieju.

Mini analiza sentymentu

```
slownik_pozytywne<-read.table("D:/Dokumenty/studia/8 semestr/R i Big Data/lab2/positive-words.txt")
slownik_negatywne<-read.table("D:/Dokumenty/studia/8 semestr/R i Big Data/lab2/negative-words.txt")
old_tweets<-tweets
tweets<-stri_extract_all_words(tweets)
wydzwiek<-sapply(tweets,function(lista)
  {
     n<-length(lista)
     wartosci<-numeric(n)
     for(i in seq_len(n))
     {
        ktory_pozytywne<-which(lista[i]==slownik_pozytywne)
        ktory_negatywne<-which(lista[i]==slownik_negatywne)
        if(length(ktory_pozytywne)>0)
        {
           wartosci[i]<-1
        }
        if(length(ktory_negatywne)>0)
        {
           wartosci[i]<--1
```

```
        }
      }
      lista<-sum(wartosci)
   })

sentyment<-data.frame(wydzwieki=c("pozytywny","negatywny","neutralny"),
   ilosc=c(length(wydzwiek[wydzwiek>0]),length(wydzwiek[wydzwiek<0]),
      length(wydzwiek[wydzwiek==0])))

ggplot(sentyment,aes(x=wydzwieki, y=ilosc)) +
   geom_point(size=3) +
   coord_flip() +
   ggtitle("Podzial tweetow ze wzgledu na sentyment")
```