Analiza tweetów

Mikołaj Waśniewski

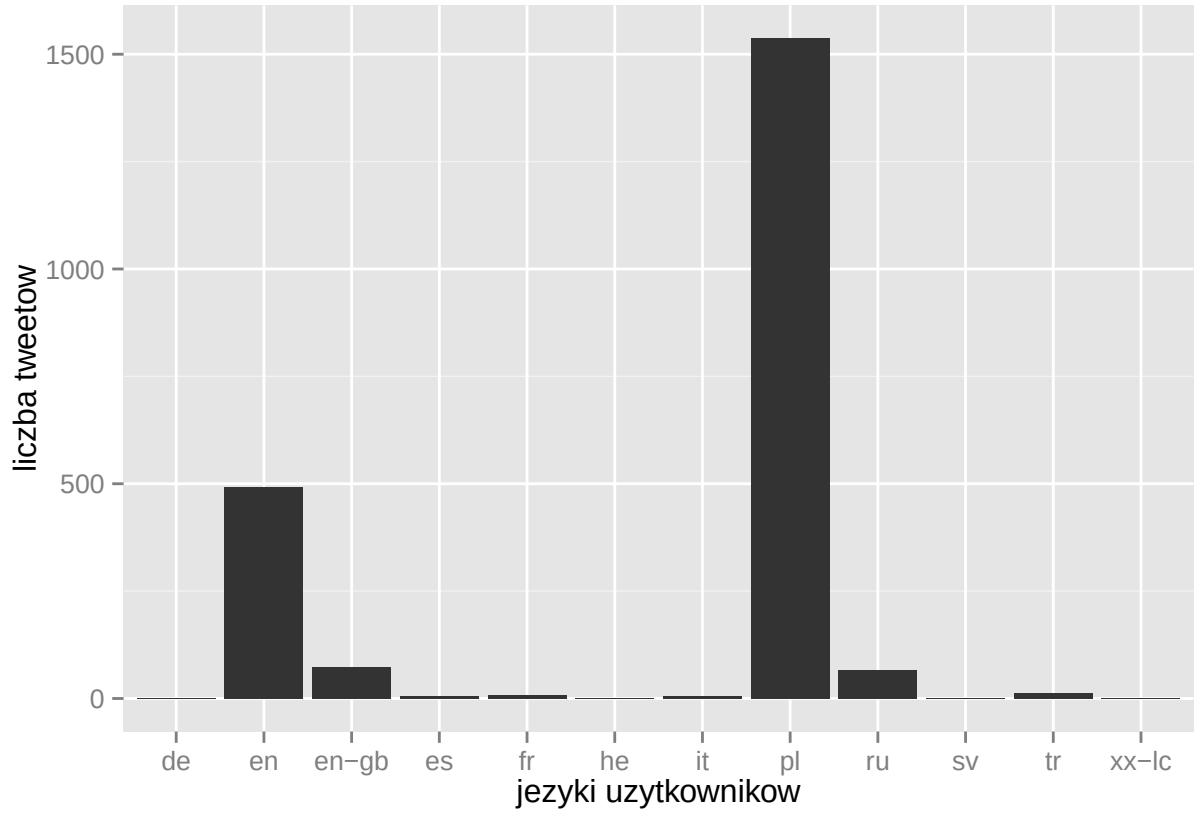
Tuesday, March 18, 2015

Wstep

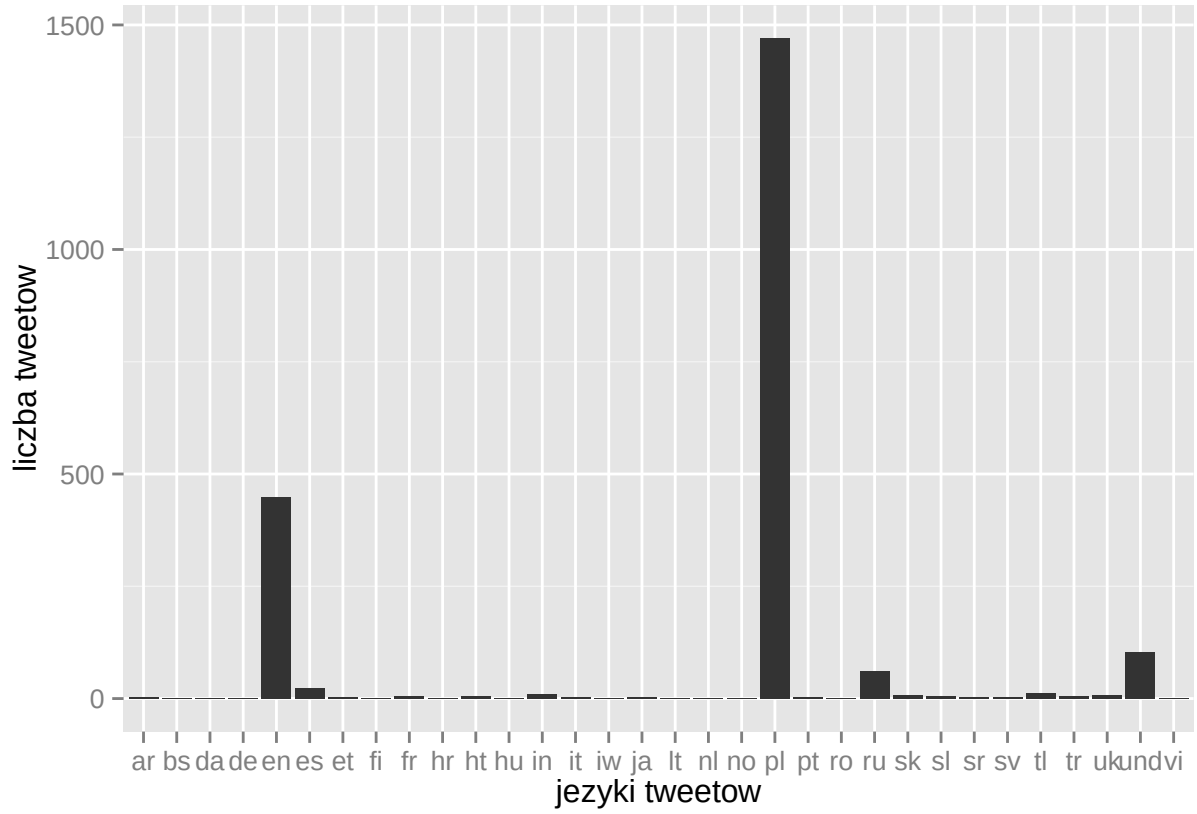
Do analizy postanowiłem wziąć tweety wysłane w Warszwie i okolicach. Podczaas selekcji tweetow ustawiłem parametr `locations= c(20.75,52,21.25,52.5)`. Tweety zbierałem cztery razy po jednej godzinie, udałe mi sie zebrać 2202 tweetów.

Jezyki

Na początku sporządziłem wykres jezykow uzytkowników pobranych tweetow (zmienna `user_lang`).



Następnie sporzadzieł wykres jezykow w jakich tweety zostały napisne (zmienna 'lang').



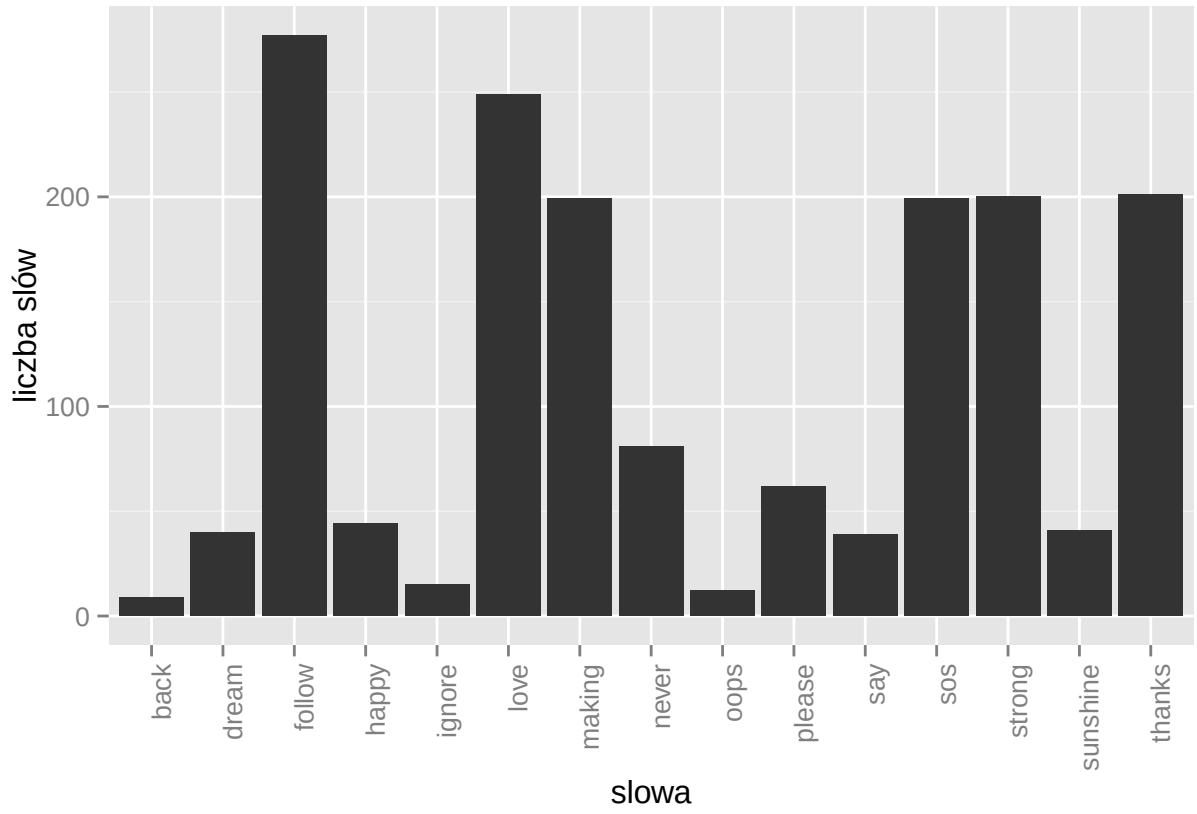
Słowa w tweetach

Tweety napisane w jezyju angielskim podzieliłem na słowa i na wykresie zamieściłem 15 najczęsciej występujacych słów.

```
tweets <- Tweets[Tweets$lang %in% c('en','en-gb','en-GB','uk'),1]
text1 <- sapply(tweets, stri_trans_tolower)
# usuniecie linkow
text1 <- sapply(text1, function(x){
   stri_replace_all_regex(x,"http[^ ]+|www[^ ]+","") }, USE.NAMES=FALSE)
# usuniecie hashtagow
text1 <- sapply(text1, function(x){ stri_replace_all_regex(x,"#[^ ]*","") },
                                                    USE.NAMES=FALSE)
# usuniecie odnosnkow do innego uyztkownika tweetera
text1 <- sapply(text1, function(x){ stri_replace_all_regex(x,"@[^ |^\\n|]+","") },
                                                    USE.NAMES=FALSE)
# usuniecie cyfr
text1 <- sapply(text1, function(x){ stri_replace_all_regex(x,"[0-9]*","") },
                                                      USE.NAMES=FALSE)
text1<-sapply(text1,removeWords,stopwords("english"),USE.NAMES=FALSE)
text1<-sapply(text1,removePunctuation,USE.NAMES=FALSE)
text1<-sapply(text1,stripWhitespace,USE.NAMES=FALSE)
without<-c("can","will","hemmings","hi","mutch","much","luke","ll","x","u")
words<-unlist(lapply(text1, stri_extract_all_words))
```

```
words<-words[!words%in%without]
words<-unname(words)
w<-head(sort(table(words), decreasing=T), n=15)

ggplot(data.frame("x"=names(w),"y"=w),aes(x=factor(names(w)), y=w),)+
   geom_bar(stat="identity")+xlab("slowa")+ylab("liczba słów")+
   theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Następnie utworzyłem chmurę 15 najczęściej występujących słów.

```
w2<-do.call(paste, c(as.list(words), sep=" "))
corp<-Corpus(VectorSource(w2))
term.matrix <- TermDocumentMatrix(corp)
term.matrix <- as.matrix(term.matrix)
commonality.cloud(term.matrix,max.words=15,random.order=FALSE)
```

Hashtagi

Najpopularniejsze hashtagi:

```
hash_tag<-unlist(stri_extract_all_regex(Tweets$text,"#[^ |^\\n|]+"))
head(sort(table(na.omit(hash_tag)), decreasing=TRUE),n=10)
```

```
##
##              #kca         #vote5sos #PolandLovesMechi             #TWUG
##              175               175                45                20
##           #Poland         #Warszawa              #Job             #Jobs
##               10                 8                 7                 7
##   #TEDxWarsaw2015      #TweetMyJobs
##                 7                 7
```

Mapa

Na mape naniosłem miejsca, z których zostału wysłane tweety.

```
lon<-Tweets$lon
lat<-Tweets$lat
lat<-lat[!is.na(lat)]
```

```
lon<-lon[!is.na(lon)]
qmap(location = "warsaw",color = 'bw',zoom = 12, messaging = TRUE, source = "google") +
   geom_point(data=data.frame("x"=lon,"y"=lat),aes(x=lon, y=lat))
```