TECHNICKÁ UNIVERZITA V KOŠICIACH

FAKULTA ELEKTROTECHNIKY A INFORMATIKY

KATEDRA KYBERNETIKY A UMELEJ INTELIGENCIE

IDENTIFIKÁCIA AUTORSTVA

(Sémantický a sociálny web)

2017

BC. DÁVID KRON, BC. ERIK SILLING, BC. JOZEF SLOVÍK

OBSAH

ZADANIE 4

1. ANALÝZA SÚČASNÉHO STAVU 4

1.1. IDENTIFIKÁCIA AUTORÍT 4

1.2. DOLOVANIE AUTORÍT 4

1.3. TWITTER 5

2. NÁVRH RIEŠENIA 5

3. RIEŠENIE 6

3.1. IDENTIFIKÁCIA AUTORSTVA 6

3.1.1. POPIS RIEŠENIA 6

3.1.2. VÝSLEDKY 7

3.2. ANALÝZA PRISPIEVATEĽOV 9

3.2.1. POPIS RIEŠENIA 9

3.2.2. VÝSLEDKY 12

4. VYHODNOTENIE 13

ZDROJE 14

ZOZNAM OBRÁZKOV

OBR. 1 PRVOTNÉ NASTAVENIA 6
OBR. 2 PREPOJENIE PROSTREDIA RSTUDIO SO SLUŽBOU TWITTER 6
OBR. 3 VYHĽADANIE TWEETOV 7
OBR. 4 IDENTIFIKÁCIA AUTORSTVA 7
OBR. 5 ROZHRANIE SLUŽBY TWITTER 7
OBR. 6 TABUĽKA S TWEETMI 8
OBR. 7 GGPLOT IDENTIFIKÁCIE AUTORSTVA 8
OBR. 8 VÝBER PRISPIEVATEĽA 9
OBR. 9 VÝBER TWEETU A ODSTRÁNENIE NEVHODNÝCH SLOV 9
OBR. 10 ANALÝZA TWEETU 9
OBR. 11 PRIRADENIE AUTORA 10
OBR. 12 ANALÝZA POČTU KOMENTÁROV 10
OBR. 13 ANALÝZA POČTU SLOV 10
OBR. 14 ANALÝZA OBĽÚBENOSTI 11
OBR. 15 ANALÝZA ČASOVÉHO ÚSEKU 11
OBR. 16 ANALÝZA PÔVODU 11
OBR. 17 WORDCLOUD Z TWEETU 12
OBR. 18 TABUĽKA ANALÝZY PRISPIEVATEĽA 12

ZADANIE

Cieľom daného zadania je identifikácia autorstva konkrétneho príspevku. Tento cieľ možno definovať ako zodpovedanie nasledujúcich otázok:

- Kto je autorom príspevku ?
- Aký typ človek je prispievateľ ?

V nasledujúcich kapitolách bude popísaný postup vypracovania tohto zadania a to od analýzy súčasného stavu danej problematiky cez návrh, vytvorenie a prezentáciu navrhnutého riešenia až po vyhodnotenie.

1. ANALÝZA SÚČASNÉHO STAVU

1.1. IDENTIFIKÁCIA AUTORÍT

V prípade ak hovoríme o autorite, tak máme na mysli autoritu spravidla overenú. Táto autorita sa rozdeľuje na dva typy:

1. **Neformálna (prirodzená)** – autorita, ktorá má osobný profil, primerané sebavedomie, je posilňovaná rešpektom vedených ľudí. Vyznačuje sa čestnosťou, rozhodnosťou, prípadne predvídateľnosťou.
2. **Formálna** – autorita, ktorá má pozíciu, titul resp. funkciu v organizácii, jej status podlieha zmene a vyžaduje poslušnosť a podriadenosť. Platí, že formálna a prirodzená autorita môžu byť totožné, pričom formálna autorita sa môže zmeniť na prirodzenú.

Pri téme autorít na webe môžeme definovať dva typy autorít webu:

1. **Priateľ** – má veľké množstvo priateľov v rámci sociálneho webu a je podporovaný vzťahmi.
2. **Šíriteľ vplyvu (influencer)** – je často citovaný (odvolávajú sa na neho iní), vie zaujať iných, je autoritou, ktorá je podporovaná názormi a vedomosťami o objekte diskusie.

Z hľadiska prístupu môžeme autority rozdeliť do týchto dvoch skupín:

1. **Autority vo vede** – vedecké články na osobných stránkach, digitálne knižnice.
2. **Autority vo webových diskusiách** – diskusie k produktom, recenzie filmov alebo kníh, sociálne siete.

1.2. DOLOVANIE AUTORÍT

Pri téme dolovania autorít či už vo vedeckých článkoch alebo webových diskusiách poznáme dva typy dolovania autorít a to:

- dolovanie autorít zo štruktúry,
- dolovanie autorít z obsahu.

Pri dolovaní autorít vstupné dáta obsahujú:

- meno prispievateľa,
- polaritu príspevku,

- dĺžku príspevku,
- príspevky (reakcie),
- pozíciu príspevku v strome (štruktúre diskusie).

Tieto vstupné dáta vstupujú do procesu odhadu autority. Autorita nie je vzťahovaná k príspevkom, ale k prispievateľom, pričom integrácia všetkých informácií o prispievateľovi nepatrí k triviálnym úlohám. V priebehu procesu odhadu autority je vytváraný usporiadaný rebríček indikujúci prispievateľov, ktorí:

- prezentujú hlbokú znalosť problematiky,
- vyvolávajú mnoho reakcií,
- inicializujú najčastejšie prechod na novú tému.

1.3. TWITTER

Twitter je mikroblogovacia sociálna sieť umožňujúca svojim prispievateľom posielať a čítať správy ostatných prispievateľov. Tieto správy sa nazývajú tweety. Každý tweet môže obsahovať maximálne 280 znakov zobrazených na profile prispievateľa, preto hovoríme o mikroblogu. Twitter vlastní spoločnosť Twitter Inc. Jeho zakladateľmi sú Jack Dorsey, Noah Glass, Evan Williams, Biz Stone. Twitter je vo svete najviac rozšírený v USA.

2. NÁVRH RIEŠENIA

V tejto kapitole bude vysvetlené, ako chceme riešiť dané zadanie. V našom zadaní sme sa rozhodli pomocou text mining-ových metód analyzovať príspevky zo sociálnej siete Twitter a to pomocou štatistického jazyka R v prostredí RStudio. Naše riešenie zadania bude pozostávať z týchto postupných krokov:

1) **Dolovanie textových údajov zo služby Twitter** – toto dolovanie bude vykonané pomocou R balíkov twitteR, tm, wordcloud a ďalších. Balík twitteR poskytuje prístup k údajom služby Twitter, tm poskytuje funkcie na dolovanie textu a wordcloud vizualizuje výsledky dolovania vo forme cloud-u. V riešení budú použité aj ďalšie balíky, tie budú uvedené neskôr.
2) **Transformácia textov** – tweety (príspevky používateľov v službe Twitter) sa najprv prevedú na dátový rámec a potom na korpus (korpus sa vytvorí pomocou balíka tm). Potom vytvorený korpus prejde niekoľkými potrebnými transformáciami – napr. zmenou veľkých písmen na malé, odstránením interpunkcie alebo čísel atď.
3) **Identifikácia autorstva** – pri príspevkoch (tweetoch) bude identifikované, kto daný príspevok napísal, rovnako budú identifikované aj rôzne iné parametre.
4) **Analýza prispievateľov** – po identifikovaní autora príspevku a ďalších parametrov nasleduje analýza prispievateľov. Pri tejto analýze si vyberieme už známeho autora príspevku a začneme analyzovať rôzne parametre jeho písania – napr. to, či píše zväčša krátke alebo dlhé príspevky, či je obľúbeným prispievateľom, v ktorej časti dňa najčastejšie prispieva, z ktorej krajiny pochádza atď.
5) **Vyhodnotenie** – na konci riešenia zadania v skratke zhrnieme čo všetko bolo vykonané a aké výsledky boli dosiahnuté.

3. RIEŠENIE

3.1. IDENTIFIKÁCIA AUTORSTVA

V tejto podkapitole bude popísané riešenie pre identifikáciu autorstva. Po tomto popise prejdeme už ku konkrétnym výsledkom, ktoré sme dosiahli po správnom vykonaní riešenia.

3.1.1. POPIS RIEŠENIA

1. **Prvotné nastavenia** – v tejto časti sme nastavili domovský adresár pre projekt v Rstudiu a takisto sme stiahli a nainštalovali balíky potrebné pre prepojenie prostredia Rstudia so službou Twitter a následný data mining zo získaných textov. Táto časť riešenia je platná aj pre podkapitolu s analýzou prispievateľov.

```
#nastavenie domovskeho priecinku
setwd("~/TUKE_5.rocnik_ZS/Semanticky_a_socialny_web/Analysign_Autorship")

#nastavenia twittru
install.packages(c("devtools", "rjson", "bit64", "httr","httpuv"))
#kniznice
library(devtools)
library(twitteR)
library(stringr)
library(ggplot2)
library(NLP)
library(tm)
library(RSentiment)
library(RColorBrewer)
library(wordcloud)
```

Obr. 1 Prvotné nastavenia

2. **Prepojenie prostredia RStudio so službou Twitter** – tu sme vytvorili prepojenie prostredia RStudio so službou Twitter.

```
#apikey
api_key <- "Z5PIfvkQExaVshBj0D58IKW1e"
#apisecret
api_secret <- "Y53GB2Q0lNpXYW2qDETSSPHpSYBMpVFnGKjKL7BgSwyQyhmnG6"
#accesstoken
access_token <- "938133282813829126-tavd9hUqPgR98IuLQ4w3NyM6iBweUMD"
#accesstokensecret
access_token_secret <- " SAPDkfD5bOLd59H7mn24ndeXbXHdjyvaY6BHfCV78fOAO"
#nadviazanie spojenia
setup_twitter_oauth(api_key,api_secret)
```

Obr. 2 Prepojenie prostredia RStudio so službou Twitter

3. **Vyhľadanie príspevkov (tweetov)** – tu boli po vytvorení prepojenia na službu Tweeter hľadané príspevky – napr. hľadáme 200 tweetov, ktoré obsahujú hashtag #xiaomi:

```
#autor prispevku
#pocet tweetov
n=200
tweet='#xiaomi'

#zadavame tweet
rdmTweets <- searchTwitter(tweet, n=n)
```

Obr. 3 Vyhľadanie tweetov

4. **Identifikácia autorstva** – v poslednej časti tohto opisu riešenia sme identifikovali autora príspevku a vykonali viacero možných úprav a modifikácií, ktoré sme zakončili vykreslením grafického zobrazenia výstupu – tzv. ggplotom.

```
#vytvorenie datoveho ramca
datfram <- do.call("rbind", lapply(rdmTweets, as.data.frame))
#vsetky stlpce
names(datfram)
#prve tri riadky
head(datfram,3)
#obmedzenie pre zobrazenie viac ako jedneho vyskytu
counts=table(datfram$screenName)
cc=subset(counts,counts>1)
barplot(cc,las=2,cex.names =0.6)
#odstranene pomocou znakovej sady
datfram$text=sapply(datfram$text,function(row) iconv(row,to='UTF-8'))
#pomocna funkcia na odstranenie @ pri mene...
trim <- function (x) sub('@','',x)
##funkcie na pridanie novych stlpcov
#vyberam kto ma spravu
datfram$to=sapply(datfram$text,function(tweet) str_extract(tweet,"^(@[[:alnum:]_]*)"))
datfram$to=sapply(datfram$to,function(name) trim(name))
#sposob urcenia RT
datfram$rt=sapply(datfram$text,function(tweet) trim(str_match(tweet,"^RT (@[[:alnum:]_]*)")[2]))
#vykreslenie poctu mien v grafe
ggplot()+geom_bar(aes(x=na.omit(datfram$rt)))+xlab(NULL)+theme(axis.text.x = element_text(angle=-45, hjust=0, vjust=1))
```

Obr. 4 Identifikácia autorstva

3.1.2. VÝSLEDKY

1. **Prepojenie prostredia RStudio so službou Twitter** – po vykonaní prvotných nastavení (nainštalovanie potrebných balíkov, nastavenie balíkov) sme vykonali prepojenie Rstudia so službou Twitter. Po spustení tejto časti sa nám zobrazí rozhranie služby Twitter, ktoré požaduje naše prihlásenie sa:

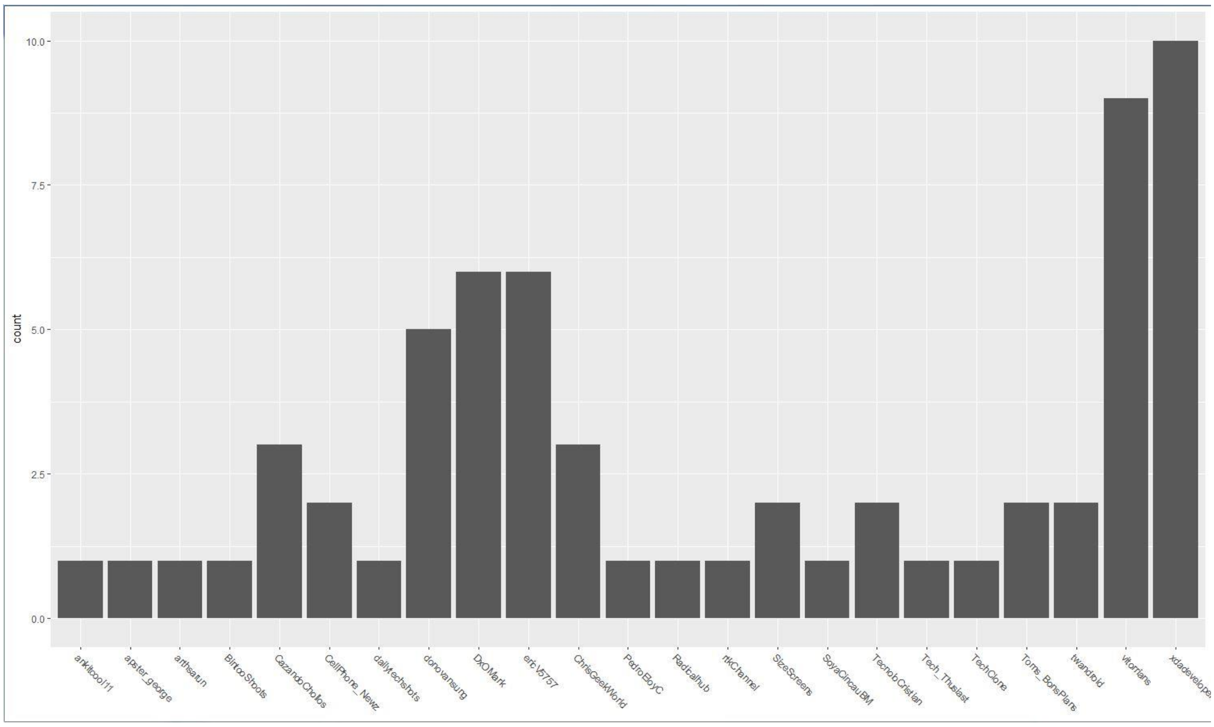
Obr. 5 Rozhranie služby Twitter

2. **Vyhľadanie príspevkov (tweetov)** – po prepojení so službou Twitter bola vytvorená tabuľka obsahujúca 200 tweetov, ktoré obsahujú hashtag #xiaomi:

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id
1	iPhoneAddict : Notre TOP 10 du meilleur de la techno en...	FALSE	1	NA	2017-12-09 10:04:08	TRUE	NA	939435486388670464
2	#DxOMark #Xiaomi #Mi #Note #minote3 #smartphone ...	FALSE	0	NA	2017-12-09 09:58:53	TRUE	NA	939434164251742208
3	Wer hat schon Lust zum #staubsaugen? Er schon í ˝í¸ #X...	FALSE	0	NA	2017-12-09 09:58:41	TRUE	NA	939434115220287488
4	RT @CellPhone_Newz: Kantar: #Huawei, #Xiaomi, #Appl...	FALSE	0	NA	2017-12-09 09:53:13	FALSE	NA	939432738460807169
5	RT @xdadevelopers: Semi-functional LineageOS 14.1 bui...	FALSE	0	NA	2017-12-09 09:45:01	FALSE	NA	939430675865849856
6	#Xiaomi Mi Max 2 review https://t.co/BkLJYa1sVz @gadg...	FALSE	0	NA	2017-12-09 09:41:53	FALSE	NA	939429885830942720
7	Kantar: #Huawei, #Xiaomi, #Apple, Vivo, Oppo account ...	FALSE	0	NA	2017-12-09 09:35:50	TRUE	NA	939428363630006273
8	RT @xdadevelopers: Semi-functional LineageOS 14.1 bui...	FALSE	0	NA	2017-12-09 09:34:18	FALSE	NA	939427978592825344
9	#Xiaomi Redmi 5 Đ¸ Redmi 5 Plus ĐżŃ€ĐľŃ‚ĐµŃŃ‚Đ¸Ń€ĐľĐ...	FALSE	0	NA	2017-12-09 09:33:12	FALSE	NA	939427700464390145
10	RT @Tech_Thusiast: Check out my speed test comparison...	FALSE	0	NA	2017-12-09 09:31:39	FALSE	NA	939427311052578816
11	#Xiaomi Redmi 5 Đ¸ Redmi 5 Plus ĐżŃ€ĐľŃ‚ĐµŃŃ‚Đ¸Ń€ĐľĐ...	FALSE	3	NA	2017-12-09 09:30:24	FALSE	NA	939426997683523584
12	RT @xdadevelopers: Semi-functional LineageOS 14.1 bui...	FALSE	0	NA	2017-12-09 09:29:53	FALSE	NA	939426866947014656
13	#Xiaomi Mi Note 3: ÎĎ€ÎąÎ±ĎÎµ Ď¸ÎżĎ…Ď‚ 90 Ď€ĎŚÎ´Ď¸ÎżĎ...	FALSE	0	NA	2017-12-09 09:29:16	FALSE	NA	939426712542171136
14	Aparat w Xiaomi Mi Note 3 jest jednym z lepszych na ryn...	FALSE	0	NA	2017-12-09 09:29:09	TRUE	NA	939426682930216960
15	Äiá»‡n thoáş¡i #Xiaomi Redmi 5 vĂ Redmi 5 Plus https://...	FALSE	0	NA	2017-12-09 09:28:58	FALSE	NA	939426636247851008
16	#Xiaomi > #Apple Let the shitstorm happen	FALSE	0	NA	2017-12-09 09:26:00	FALSE	NA	939425891003793409
17	Make your home cool with this #xiaomi product https://...	FALSE	0	NA	2017-12-09 09:16:03	FALSE	NA	939423385389584386
18	RT @xdadevelopers: Semi-functional LineageOS 14.1 bui...	FALSE	0	NA	2017-12-09 09:11:32	FALSE	NA	939422250884722689
19	#Xiaomiâ€™s â€#MiPadâ€™ #TrademarkBlocked for Cop...	FALSE	0	NA	2017-12-09 09:10:37	FALSE	NA	939422019925561345

Obr. 6 Tabuľka s tweetmi

3. **Identifikácia autorstva –** po vykonaní viacerých úprav a modifikácií sme vykonali identifikáciu autora, ktorú sme zobrazili do ggplot-u:



Obr. 7 GGplot identifikácie autorstva

3.2. ANALÝZA PRISPIEVATEĽOV

Táto podkapitola bude mať podobné dve časti ako predchádzajúca podkapitola, avšak nebude sa zameriavať na identifikáciu autorstva, ale na analýzu prispievateľov.

3.2.1. POPIS RIEŠENIA

1. **Prvotné nastavenia + 2. Prepojenie prostredia RStudio so službou Twitter** – tieto dve kroky sú krokmi, ktoré už boli realizované pri identifikácii autorstva.
3. **Výber prispievateľa** – tu sme si vybrali prispievateľa, ktorého sme chceli ďalej podrobnejšie analyzovať. Napr. na obrázku je možné vidieť výber prispievateľa s menom „Ajith_kmr“:

```
##analyza prispievatela
#vyberiem si meno prispievatela
name="Ajith_kmr"
```

Obr. 8 Výber prispievateľa

4. **Výber tweetu a odstránenie nevhodných slov** – v tejto časti sme vybrali tweet, ktorý sme chceli v ďalšej časti analyzovať, no ešte predtým sme definovali funkciu pre výpočet počtu slov v danom tweete a takisto aj odstránili nevhodné slová.

```
#vyber textu twetu
data = as.data.frame.matrix(datfram[1], header=TRUE,quote=NULL)
#vyberame tweet podla mena
data.rand <- data[grep(name, datfram$screenName),]
#funkcia na vypocet poctu slov
countSpaces <- function(s) { sapply(gregexpr(" ", s), function(p) { sum(p>=0) } ) }
foo <- transform(data.rand, baz = countSpaces(data.rand))
#odstranenie nevhodnych slov
data.rand=gsub("[^[:graph:]]"," ",data.rand)
data.rand=gsub("https://t.co/","",data.rand)
corpus <- Corpus(VectorSource(data.rand))
corpus <- tm_map(corpus, removeNumbers)#cisla
corpus <- tm_map(corpus, removePunctuation)#znaky
corpus <- tm_map(corpus, stripWhitespace)#medzeri
corpus <- tm_map(corpus, removeWords, stopwords('english'))#slova bez vyznamu
matica <- DocumentTermMatrix(corpus)
```

Obr. 9 Výber tweetu a odstránenie nevhodných slov

5. **Analýza tweetu** – tu sme vybraný tweet analyzovali z pohľadu viacerých základných údajov – času, obľúbenosti, počtu komentárov či množstva pozitívneho resp. negatívneho textu. Na konci sme v závislosti od použitého hashtag-u vykreslili wordcloud.

```
#dlzka prispevku
pris=foo[1,2]-matica$ncol
word=foo[1,2]
#pocet komentov
pockom=sum(datfram$retweetCount[grep(name, datfram$screenName)])
#oblubenost
oblub=sum(datfram$favoriteCount[grep(name, datfram$screenName)])
#cas prispevku
caspri=datfram$created[grep(name, datfram$screenName)]
#krajina prispievatela
krajpris=datfram$language[grep(name, datfram$screenName)]
#priatel
pria=datfram$rt[grep(name, datfram$screenName)]
#pozitivne a negativny text
dataframe <- data.frame(text=sapply(corpus, identity), stringsAsFactors=F)
sentiment=calculate_sentiment(dataframe)
#workcloud
wordcloud(corpus,scale=c(3,0.2),max.words = n/2, min.freq = 0.01,colors=brewer.pal(8, "Dark2"))
```

Obr. 10 Analýza tweetu

6. **Analýza prispievateľa** – posledným bodom tohto opisu riešenia je samotná analýza prispievateľa. Najprv bolo potrebné priradiť autora/prispievateľa, ktorého príspevky sme chceli analyzovať:

```
#nacitanie tabulky
tabulka=read.table("Analyza.csv", header = FALSE, sep = ";",stringsAsFactors=FALSE)
#meno autora
tabulka[1,2]=name
#twet
tabulka[2,2]=tweet
```

Obr. 11 Priradenie autora

Neskôr sme začali s jednotlivými analýzami prispievateľa, ktorými boli:

a) Analýza počtu komentárov, ktoré prispievateľ dostáva na svoje tweety:

```
#Ma komentare
#malo #stredne #vela
if(pockom<5){
  tabulka[3,2]="malo komentarov"
}else if(pockom<15){
  tabulka[3,2]="stredne vela komentarov"
}else{
  tabulka[3,2]="vela komentarov"
}
```

Obr. 12 Analýza počtu komentárov

b) Analýza počtu slov, ktoré prispievateľ používa:

```
#Pouziva slova
#malo #stredne #vela
if(word<10){
  tabulka[4,2]="malo slov"
}else if(word<20){
  tabulka[4,2]="stredne vela slov"
}else{
  tabulka[4,2]="vela slov"
}
```

Obr. 13 Analýza počtu slov

c) Analýza obľúbenosti prispievateľa:

```
#Je oblubeny
if(oblub<5){
  tabulka[5,2]="malo oblubeny"
}else if(oblub<15){
  tabulka[5,2]="stredne oblubeny"
}else{
  tabulka[5,2]="vela oblubeny"
}
```

Obr. 14 Analýza obľúbenosti

d) Analýza časového úseku, kedy používateľa najčastejšie pridáva tweety:

```
#Pridava prispevky
#morning #day #night
caspri=gsub(Sys.Date(), "" , caspri, perl=TRUE)
caspri=gsub(" ", "" , caspri, perl=TRUE)
caspri=gsub("[:*]", "" , caspri, perl=TRUE)
if(as.integer(caspri)<90000){
  tabulka[6,2]="rano"
}else if(as.integer(caspri)<190000){
  tabulka[6,2]="cez den"
}else{
  tabulka[6,2]="vecer"
}
```

Obr. 15 Analýza časového úseku

e) Analýza pôvodu prispievateľa:

```
#Je z krajiny
if(krajpris=="en"){
  tabulka[7,2]="U.S.A"
}else if(krajpris=="fr"){
  tabulka[7,2]="France"
}else{
  tabulka[7,2]="other"
}
```

Obr. 16 Analýza pôvodu

3.2.2. VÝSLEDKY

1. **Výber prispievateľa, výber tweetu, odstránenie nevhodných slov a analýza tweetu** – pri nami vybratom prispievateľovi s menom „Ajith_kmr“ sme si zvolili tweet, ktorý sme následne očistili o nevhodné slová a analyzovali vykreslením wordcloudu:

Obr. 17 Wordcloud z tweetu

2. **Analýza prispievateľa** – analýza prebehla vo viacerých bodoch, ktoré sú podrobnejšie popísané v podkapitole ,,Popis riešenia“. Analýzou prispievateľa sme sa dostali k nasledujúcej tabuľke:

Filter

	V1	V2
1	Meno	ajith_kmr
2	tweet	#xiaomi
3	Ma komentare	stredne vela komentarov
4	Pouziva slova	stredne vela slov
5	Je oblubeny	malo oblubeny
6	Pridava prispevky	rano
7	Je z krajiny	U.S.A
8	Ma priatela	DxOMark
9	Jeho text je	Positive

Obr. 18 Tabuľka analýzy prispievateľa

Z tabuľky vyplýva, že:

- tweet prispievateľa má stredne veľa komentárov,
- prispievateľ používa stredne veľa slov,
- prispievateľ je málo obľúbený a príspevky pridáva zväčša ráno,
- prispievateľ je z USA, má priateľa DxOMark a jeho príspevky sú pozitívne.

4. VYHODNOTENIE

Výsledkom našej práce bolo vytvoriť aplikáciu, ktorá by dokázala vyhodnocovať autora a určiť jeho charakteristiku z príspevku. Zamerali sme sa na spracovanie príspevkov služby Twitter a môžeme konštatovať, že sme dosiahli ciele, ktoré sme si na začiatku deklarovali. Naša aplikácia s interakciou od prispievateľa dokáže získavať tweety a na tento tweet aplikovať svoj program, ktorého výsledkom je identifikácia autora a určenie charakteristík jeho príspevkov. Samozrejme potenciál tejto aplikácie sa dá ďalej rozvíjať a aplikácia sa dá ďalej zlepšovať.

ZDROJE

[1] MACHOVÁ, Kristína – KONCZ, Peter: Metódy dolovania v konverzačnom obsahu so zameraním na analýzu sentimentu. FEI TU v Košiciach. 2013

[2] https://www.r-bloggers.com/getting-started-with-twitter-analysis-in-r/

[3] https://rdatamining.wordpress.com/2011/11/09/using-text-mining-to-find-out-what-rdatamining-tweets-are-about/