Analiza danych profili obserwowanych przez oficjalne konto Uniwersytetu Warszawskiego na Instagramie

Michał Posiadała
nr. indeksu 438991

13 lutego 2022

1 Wstęp

Instagram powstał w 2010 roku i szybko stał się jednym z najbardziej popularnych social mediów. Ten prosty w obsłudze portal stał się niezwykle popularnym kanałem komunikacyjnym i marketingowym. Korzystają z niej celebryci, firmy, politycy ale i zwykli użytkownicy chcący publikować swoje zdjęcia.

Celem mojego projektu jest zbadanie profili obserwowanych przez oficjalne konto Uniwersytetu Warszawskiego (zwanych dalej sąsiadami tego profilu)

Badane dane zostały zebrane za pomocą napisanego przez mnie kodu w Pythonie, dotyczą one 50 profili, 49 sąsiadów obserwowanych przez @uniwersytetwarszawski oraz konta @uniwersytetwarszawski

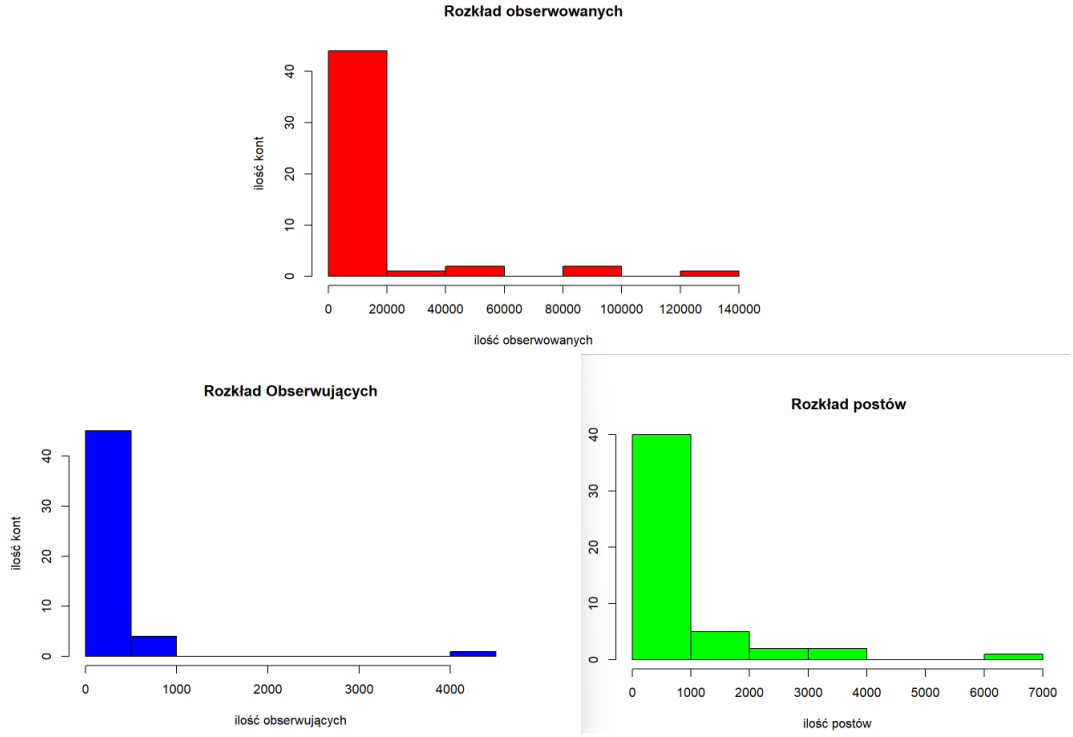
```
# załadowanie danych
dane <- read.csv("filtered_data.csv", header = TRUE, sep=","
,dec=".",encoding="UTF-8")
```

2 Czy ”relacje” internetowe zachowują się relacje ludzkie?

Rozkład relacji ludzkich w danej społeczności charakteryzuje się tym, że nie ulega on rozkładowi normalnemu (Gaussa), lecz przypomina on bardziej rozkład wykładniczy, tj. duża ilość osób ma małą liczbę znajomych, a mała liczba osób ma dużą liczbę

znajomych. Na podstawie poniższych wykresów widać, że rzeczywiście konta na Instagramie zachowują się bardzo podobnie, efekt jest niestety bardzo słabo widoczny ze względu na małą ilość kont które obserwuje profil @uniwersytetwarszawski. Co ciekawe, tą własność zachowuje również rozkład publikowanych postów.

```
hist(dane$edge_followed_by, main = "Rozkład obserwowanych", col="red
    ", xlab = "ilość obserwowanych", ylab = "ilość kont" )
hist(dane$edge_follow, main = "Rozkład Obserwujących", col="blue",
    xlab = "ilość obserwujących",ylab = "ilość kont" )
hist(dane$edge_follow, main = "Rozkład Obserwujących", col="blue",
    xlab = "ilość obserwujących")
hist(dane$edge_owner_to_timeline_media, main = "
     Rozkład postów", col="green",  xlab = "ilość postów",ylab = "
    ilość kont" )
```



3 Jakie konta mają największą liczbę obserwujących?

W poniższej tabeli widać jakie profile w "sąsiedztwie" profilu Uniwersytetu Warszawskiego cieszą się największą popularnością.

```
sorted_dane<-dane[order(dane$edge_followed_by, decreasing = TRUE),]
df<-data.frame(sorted_dane)
df[1:5,c('full_name', 'edge_followed_by')]
```

```
                                full_name edge_followed_by
34                   Nowa Warszawa | Portal           120139
1                 Miasto Stołeczne Warszawa            97760
2                                 Go2Warsaw            88080
35 Warszawa Poland <U+0001F1F5><U+0001F1F1>            45956
13                       Sorbonne Université            41603
```

4 Rozkład ilości obserwujących w społeczności @uniwersytetwarszawski

Badanie ilości obserwujących,(nie ma niskiej ilości obserwujących ze względu na duże odchylenie standardowe).

```
śr<-mean(dane$edge_followed_by)
odch<-sd(dane$edge_followed_by)
maks<-max(dane$edge_followed_by)

obser.f<-factor(cut(dane$edge_followed_by, breaks=c(0,śr-odch, śr+
    odch, maks), labels=c("niska", "przeciętna", "duża")))
dane$ilość_obserwujących<-obser.f
pie(table(dane$ilość_obserwujących), labels = c("przeciętna", "duża"
    ), col = rainbow(length(dane)), cex=0.7, main = "Rozkład ilości
    obserwujących")
```

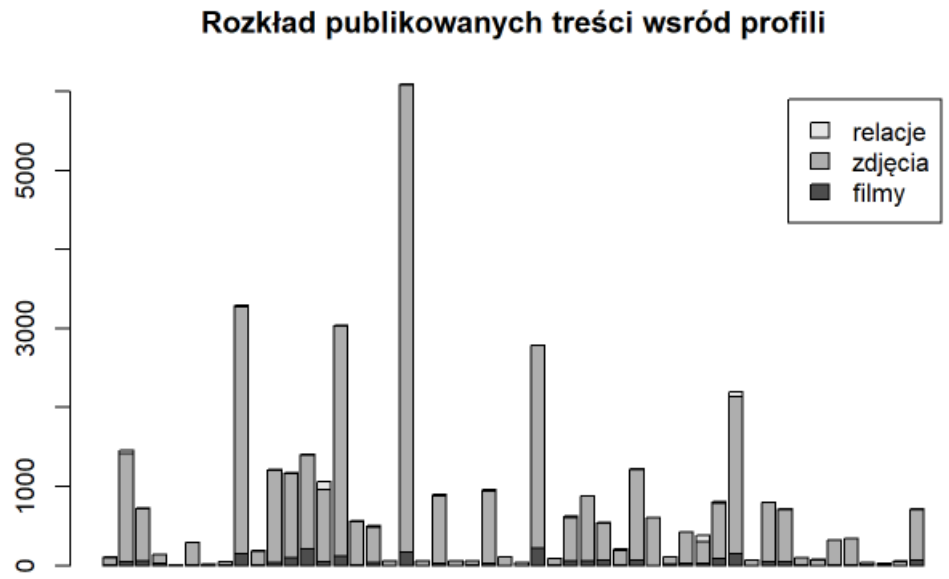
5 Co publikują badane konta?

Na pierwszej stronie każdego profilu możemy sprawdzić ile dany profil opublikował postów, filmów i wyroznionych relacji. Poniżej możemy zobaczyć jak ta liczba rozkłada się na tle wszystkich badanych profili. Jak widać poniżej, znaczną przewagę w publikacjach mają zdjęcia, nad drugim miejscu są filmy, a wyrónione relacje występują sporadycznie (średnia relacji to zaledwie 12 na konto, zdjęć to z kolei ponad 300).

```
y1<-aggregate(dane$edge_felix_video_timeline, by=list(dane$username)
    , sum, na.rm=TRUE)
y2<-aggregate(dane$edge_owner_to_timeline_media - dane$edge_felix_
    video_timeline, by=list(dane$username), sum, na.rm=TRUE)
y3<-aggregate(dane$highlight_reel_count, by=list(dane$username), sum
    , na.rm=TRUE)
tabela<-cbind(y1, y2$x, y3$x)
colnames(tabela)<-c("username", "filmy", "zdjęcia", "relacje")

tabela.m<-as.matrix(tabela)

barplot(t(tabela.m[,2:4]), main="Rozkład publikowanych treści wsród
    profili", legend.text = TRUE)
```



6 Czy kategorie profili wybrane przez ich właścicieli są spójne z tym co wybrał dla nich Instagram?

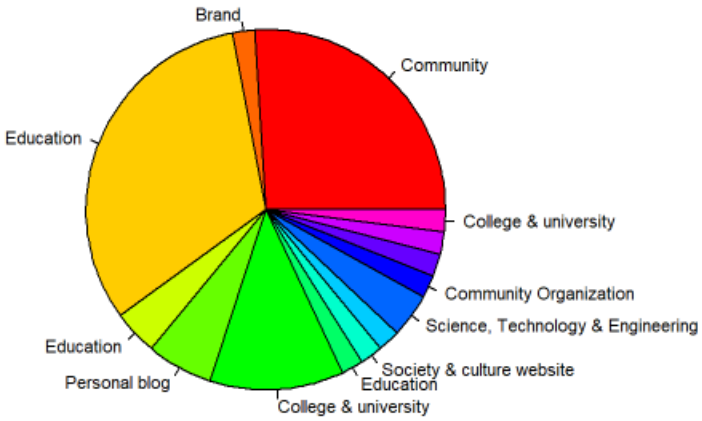
Każdy po założeniu swojego profilu może wybrać kategorię jakiej dotyczy profil. Lista jest bardzo długa i zawiera pozycje oczywiste (Diagram I), takie jak Bloger czy aktor, jak i bardziej ekstrawaganckie, na przykład Sklep z Domowymi Wypiekami. Jednak Instagram, lub bardziej jego właściciel (Meta, daw. Facebook) używają swoich własnych algorytmów i każdemu profilowi przyznają ich własne kategorie, które niezależnie od wyboru autora, trochę bardziej oddają specjalizację danego profilu (Diagram II). Kategorie te nie są widoczne z poziomu aplikacji czy też strony internetowej, są one ukryte w kodzie strony, który mój program w Pythonie zdołał wydobyć.
Podczas interpretacji obydwu diagramów, możemy dokonać pewnych obserwacji:

- Instagram wielu profilom nie przyznał żadnej kategorii, co widać przez miejsce zajmowane przez konta nie posiadające kategorii (pole czerwone na Diagramie II)

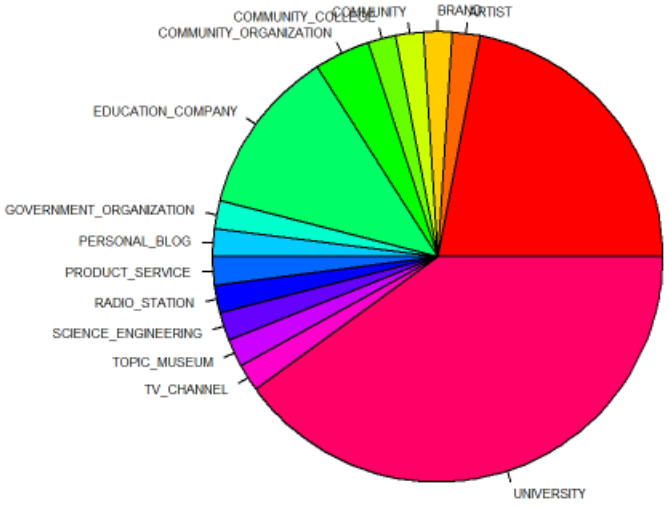
- Kategorie nadane przez serwis są w tym przypadku trafne, jako przykłady można wskazać kategorię RADIO_STATION nadane profilowi @radiokampus czy też kategoria UNIWERSITY przyznane oficjalnemu profilowi Politechniki Warszawskiej, gdzie twórcy tych profili nie wybrali żadnej z dostępnych kategorii.

```
pie(table(dane$category_name), labels = dane$category_name, col =
    rainbow(length(dane)), cex=0.7, main = "Kategorie ustawiane przez
     użytkowników (Diagram I)")
pie(table(dane$category_enum), col = rainbow(length(dane)),  cex
    =0.5, radius = 1, main = "Kategorie ustawiane przez Instagrama (
    Diagram II)")
```

Kategorie ustawiane przez użytkowników (Diagram I)



Kategorie ustawiane przez Instagrama (Diagram II)



7 Czy biografie są podobne?

Każdy profil, może zamieścić krótki opis czego dotyczy (tzw. biografię), i jest zwykle głównym źródłem informacji o danym koncie. Jednym z sposobów na szukanie wzorców w danej społeczności jest sprawdzenie jakie słowa pojawiają się najczęściej. Jak widać poniżej, wśród słów mających więcej niż 4 znaki, najczęściej pojawiają się słowa związane z Warszawą oraz Uniwersytetem.

```
library("stringr")
str<-paste(dane$biography, collapse = ', ')
str<-paste(str_extract_all(str, '\\w{4,}')[[1]], collapse=' ')

word_list<-as.list(strsplit(str, '[[:space:]]'))
sort(table(word_list),decreasing=TRUE)[0:5]
```

```
word_list
      Warsaw    University  Uniwersytetu Warszawskiego      official
          12            11            11            10             6
```