# RACHUNEK PRAWDOPODOBIEŃSTWA I STATYSTYKA PROJEKT ZALICZENIOWY - RAPORT

20 maja 2016

## 1 Zespół

## 2 Pozyskanie danych

## 3 Opis danych

## 4 Czyszczenie danych

## 5 Agregacja i transformacja danych

Jednostka statystyczna: Pojedyncze zalogowanie.
Zbiorowość generalna: Zbiór wszystkich zalogowań do systemu.
Zbiorowość próbna: Pojedynczy użytkownik (ID).

## 6 Eksploracja danych

a. Liczebność względna jednostek statystycznych z niekompletnymi danymi:

$$\delta = \frac{n}{m} \quad n - bldnych\ danych,\ m - poprawnych\ danych$$

b. Przedziały klasowe, rozpiętość przedziałów klasowych:

$$avg(x) = \frac{l_x}{d_o - d_p}$$

Gdzie: x - dany użytkownik, $l_x$ - łączna liczba logowań użytkownika, $d_o$ - data ostatniego logowania, $d_p$ - data pierwszego logowania

k - ilość użytkowników

$$Kl(x) = \begin{cases} 3 \quad dla \quad avg(x) > \frac{\sum_{i=0}^{k} avg(i)}{k} + \frac{max_{i=0..k}(avg(i))\cdot 2}{3} \\ 2 \quad dla \quad \frac{\sum_{i=0}^{k} avg(i)}{k} - \frac{min_{i=0..k}(avg(i))\cdot 2}{3} \leqslant avg(x) \leqslant \frac{\sum_{i=0}^{k} avg(i)}{k} + \frac{max_{i=0..k}(avg(i))\cdot 2}{3} \\ 1 \quad dla \quad avg(x) \leqslant \frac{\sum_{i=0}^{k} avg(i)}{k} - \frac{max_{i=0..k}(avg(i))\cdot 2}{3} \end{cases} \tag{1}$$

d. Wskaźnik natężenia zmian IP dla każdego użytkownika

$$W(n) = \frac{x_n}{\sum_{i=0}^{k} x_i} \cdot 100\%$$

$x_n$ - ilość zmian Ip dla rozpatrywanego użytkownika
$\sum_{i=0}^{k} x_i$ - suma ilości zmian ip po użytkownikach = suma wszystkich zmian ip

e. Współczynnik korelacji pomiędzy Ilością zabezpieczeń a czasem

$$cor(X,Y) = \frac{E(XY) - EXEY}{\sqrt{D^2X \cdot D^2Y}} = \frac{\frac{\sum_{i=0}^{k} z_i \cdot c_i}{k} - \frac{\sum_{i=0}^{k} z_i}{k} \cdot \frac{\sum_{i=0}^{k} c_i}{k}}{\sqrt{\frac{\sum_{i=0}^{k} \left( z_i - \frac{\sum_{j=0}^{k} z_j}{k} \right)^2}{k} \cdot \frac{\sum_{i=0}^{k} \left( c_i - \frac{\sum_{j=0}^{k} c_j}{k} \right)^2}{k}}}$$

$z_i$ - ilość zabezpieczeń w i-tym logowaniu
$c_i$ - czas i-tego logowania
$k$ - ilość logowań f. Wykres zmiany ilości logowań na dzień w czasie
Wykres na którym osią X jest czas, osią Y jest ilość logowań. Jednostka na osi X to dzień a każdy punkt to suma logowań w danym dniu do systemu.
h. Wykres z naniesionymi 3 wykresami: zmiany ilości kont danej klasy w czasie
Oś X - czas - jednska to dzień.
Oś Y - ilość kont danej klasy
Każy naniesiony wykres odpowiada każdej klasie
i. Wartość minimalna funkcji prawdopodobieństwa zmiennej losowej przedstawiającej ilość logowań w danym dniu.

$$M = min\{n_1, n_2, ...n_k\} \quad n_i - \ ilosc\ zalogowan\ i-tego\ uzytkownika$$

j. Prognoza obciążenia systemu dla danego dnia

$$p(mm/dd) = \frac{\sum_{i=0}^{ll} n(mm/dd/i)}{ll}$$

- prognoza to ilość logowań w danym dniu podzielona przez ilość takich dni (jeżeli logowania miały miejsce np w różnych latach)
Czyli suma logowań w danym dniu po wszystkich latach podzielona przez ilość tych lat. Dzięki temu uzyskujemy wartość oczekiwaną logowań na dany dzień.

l. Godzina i dzień tygodnia najbardziej dogodne do przeprowadzenia prac

$$x = min_{dt/hh}\left(\frac{n(dt/hh)}{\#dt/hh}\right)$$

n(dt/hh) - suma logowań w danym dniu tygodnia i godzinie po wszystkich takich dniach i godzinach podzielona przez ich ilość
m. Rozkład Poissona ilości połączeń na godzinę – prawdopodobieństwo, że w danym czasie będzie dokładnie n równoczenych
połączeń:

$$P(X = k) = \frac{\frac{n(dt/hh)}{\#dt/hh}^{k} \cdot e^{-\frac{n(dt/hh)}{\#dt/hh}}}{k!}$$

m. Sprawdzenie czy konto x wygasłe: 1 - wygasłe, 0 - nie:
$d_a$ - aktualna data, $d_o$ - data ostatniego

$$W(x) = \begin{cases} 1 & dla \quad d_a - d_o < 182 \\ 0 & wpp \end{cases} \tag{2}$$

# 7 Hipotezy badawcze

# 8 Weryfikacja hipotez

# 9 Wnioski

# 10 Uwagi