Big Data
Lista zadań

Jacek Cichoń, WiT, PWr, 2022/23

1 Wstęp

Zadanie 1 — Pobierz plik z kilkoma dramatami Szekspira ze strony wykładu. Wybierz jeden z dramatów.

1. Oczyść wybrany plik. Podziel go na słowa.
2. Usuń z niego "Stop Words" i usuń z niego słowa o długości mniejszej lub równej 2.
3. Zbuduj chmurę wyrazów (word cloud) z otrzymanej listy. Możesz skorzystać np. z serwisu http://www.wordclouds.com/

Celem tego zadania jest wygenerowanie mniej więcej takiego obrazka (dla poematu "Pan Tadeusz"):

Zadanie 2 — To jest kontynuacja poprzedniego zadania.

1. Zastosuj część funkcji które napisałeś do realizacji poprzedniego zadania do wyznaczenia indeksów TF.IDF dla wszystkich wyrazów z dokumentów w dramatów Szekspira znajdujących się w pliku ze strony wykładu.
2. Zbuduj chmury wyrazów oparte o TF.IDF dla wszystkich rozważanych dramatów.

Zadanie 3 — Pokaż, że jeśli chcesz jednoznacznie wyreprezentować każdą z liczb ze zbioru $\{0, 1, \ldots, n\}$ za pomocą b bitów to $b \geqslant \lceil \log_2(n+1) \rceil$.

Zadanie 4 — Pokaż, że jeśli $x = \sum_{k=0}^{s} a_k 2^k$, gdzie $a_i \in \{0, 1\}$ oraz $a_s = 1$ to $s = \lceil \log_2(x+1) \rceil$

Zadanie 5 — Rozważmy następującą modyfikację licznika Morrisa: ustalamy liczbę $\alpha > 0$ oraz rozważamy tak oprogramowany licznik:

init :: C =0
onInc :: if $\left(random() < \left(\frac{1}{1+\alpha} \right)^C \right)$ then C = C+1
onGet :: return (?????)

Niech C_n oznacza wartość zmiennej C po n wywołaniach metody onInc.

1. Wyznacz $E[(1+\alpha)^{C_n}]$

2. Uzupełnij funkcję onGet tak aby otrzymać nieobciążony estymator liczby użyć metody onInc.

c.d.n.
Powodzenia,
Jacek Cichoń