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 $\mathrm{E}\left[(1 + \alpha)^{C_n}\right]$

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

Zadanie 6 — Niech C_n będzie wartością klasycznego licznika Morris'a po n krotnym wywołaniu funkcji onInc().

1. Pokaż, że $\mathrm{E}\left[4^{C_n}\right] = 1 + \frac{3}{2}n(n+1)$.
2. Pokaż, że $\mathrm{var}\left[2^{C_n}\right] = \frac{1}{2}n(n-1)$.
3. Skorzystaj z nierówności Jensena dla wartości oczekiwanej zmiennej losowej do pokazania, że $\mathrm{E}\left[C_n\right] \leqslant \log_2(n+1)$.

Zadanie 7 — Załóżmy, że $X_1, \ldots, X_m$ są niezależnymi zmiennymi losowymi o wartości oczekiwanej μ oraz wariancji σ^2. Niech

$$L = \frac{X_1 + \ldots + X_m}{m} .$$

1. Pokaż/sprawdź, że $\mathrm{E}\left[L\right] = \mu$ oraz $\mathrm{var}\left[L\right] = \frac{1}{m}\sigma^2$.
2. Pokaż, że $\Pr[|L - \mu| \geqslant \epsilon\mu] \leqslant \frac{\sigma^2}{\epsilon^2}$.

Zadanie 8 — Rozważamy ciąg $B_1, \ldots, B_n$ niezależnych zdarzeń, takich, że $\Pr[B_1] = \ldots = \Pr[B_n] = \frac{3}{4}$. Niech X oznacza liczbę sukcesów, czyli $X = \sum_{i=1}^n X_i$, gdzie $X_i = 1$ jeśli zaszło zdarzenie B_i oraz $X_i = 0$ w przeciwnym przypadku.

1. Korzystając z nierówności Czernoffa dla rozkładu dwumianowego pokaż, że
$$\Pr[X \leqslant \frac{1}{2}n] \leqslant \exp\left(-\frac{n}{24}\right)$$
2. Niech $\delta > 0$. Pokaż, że jeśli $n \geqslant 24 \ln \frac{1}{\delta}$, to $\Pr[X \leqslant \frac{n}{2}] \leqslant \delta$.
3. Skorzystaj z następującej wersji nierówności Czernoffa
$$\Pr[X \leqslant \mu - \lambda], \Pr[X \geqslant \mu + \lambda] \leqslant \exp\left(-\frac{2\lambda^2}{n}\right)$$
dla zmiennej losowej X o rozkładzie dwumianowym $\mathrm{Binom}(n, \mu)$ do wzmocnienia wyników z poprzednich dwóch punktów.

Zadanie 9 — Niech $x_1, \ldots, x_n$ będzie ciągiem liczb rzeczywistych. Rozważamy dwie funkcje $f(x) = \sum_{i-1}^n |x_i - x|$ oraz $g(x) = \sum_{i=1}^n (x_i - x)^2$

1. Pokaż, że funkcja g osiąga minimum w średniej arytmetycznej liczb $x_1, \ldots, x_n$
2. Pokaż, że funkcja h osiąga minimum w medianie ciągu $x_1, \ldots, x_n$.
Wskazówka: Możesz założyć, że $x_1 \leqslant x_x \leqslant \cdots \leqslant x_n$. Przyjrzy się najpierw pomocniczej funkcji $\phi(x) = |x_1 - x| + |x_n - x|$.

Zadanie 10 — Niech $x_1, \ldots, x_n$ będzie ciągiem liczb rzeczywistych oraz niech $a < b$ będą dowolnymi liczbami rzeczywistymi. Załóżmy, że

$$|\{i \in \{1, \ldots, n\} : x_i \in (a, b)\}| > \frac{n}{2} .$$

Pokaż, że wtedy mediana ciągu $x_1, \ldots, x_n$ należy do odcinka (a, b).
Wskazówka: Rozważ oddzielnie przypadek parzystego i nienarzystego n.

c.d.n.
Powodzenia,
Jacek Cichoń