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.

2 Hashinig

Zadanie 11 — **("Rolling hash"')** Rozważamy metodę haszowania opartą na wzorze

$$h_{r,p}([x_0, \ldots, x_k]) = \sum_{i=0}^{k} x_i \cdot r^i \mod p$$

1. Zastosuj metodę Hornera do implementacji tej metody haszowania i oszacuj złożoność obliczeniową tej metody.
2. Załóżmy, że p jest liczbą pierwszą. Rozważamy ciąg $[x_0, \ldots, x_m]$. Niech $0 \leqslant a < b < m$. Pokaż, że można wyznaczyć $h_{r,p}[x_{a+1}, \ldots, x_{b+1}])$ można wyznaczyć z $h_{r,p}[x_a, \ldots, x_{a+b}])$ w stałym czasie.
3. Załóżmy, że p jest liczbą pierwszą. Niech $\vec{x}$ i $\vec{y}$ będą ciągami długości r. Losujemy z jednakowym prawdopodobieństwem liczbę r ze zbioru $\{0, \ldots, p-1\}$. Pokaż, że
$$\Pr[h_{r,p}(\vec{x}) = h_{r,p}(\vec{y})] \leqslant \frac{r-1}{p} \ .$$
4. Zapoznaj się z algorytmem Rabina-Karpa wyszukiwania wzorca w w ciągu t. Pokaż, że jeśli to tego algorytmu zastosujemy funkcję haszującą $h_{r,p}$ z p będącym liczbą pierwszą taką, że $p > |t|^2$ zaś r jest losową liczbą ze zbioru $\{0, \ldots, p-1\}$, to algorytm ten działa w średnim czasie $O(|s| + |t|)$ ($|x|$ oznacza tu długość ciągu x).

Zadanie 12 — Do n urn wkładamy niezależnie k kul (rozważamy rozkład jednostajny). Niech $L_{n,k}$ oznacza wartość oczekiwaną liczby pustych urn. Oblicz
1. $\lim_{n\to\infty} \mathrm{E}\left[\frac{L_{n,n}}{n}\right]$
2. $\lim_{n\to\infty} \mathrm{E}\left[L_{n,n \ln n}\right]$
3. $\lim_{n\to\infty} \mathrm{E}\left[\frac{1}{\sqrt{n}}(n - L_{n,\sqrt{n}})\right]$

Zadanie 13 — Rozważamy dwie zmienne losowe X, Y o wartościach w zbiorze $\{1, \ldots, n\}$. Niech $\Pr[X = i] = \Pr[Y = i] = p_i$ dla $i \in \{1, \ldots, n\}$.
1. Pokaż, że $\Pr[X = Y] = \sum_{i=1}^{n} p_i^2$
2. Pokaż, że $\Pr[X = Y]$ przyjmuje wartość minimalną dla rozkładu jednostajnego na $\{1, \ldots, n\}$.

Zadanie 14 — Niech $f(x) = \ln(x)\ln(1-x)$ dla $x \in (0, 1)$.
1. Pokaż, że $f(x) = f(\frac{1}{2} - x)$.
2. Pokaż, że $\lim_{x\to 0} f(x) = 0$.
3. Pokaż, że f osiąga maksimum w punkcie $x = \frac{1}{2}$.
4. Naszkicuj wykres funkcji f.

Zadanie 15 — Oprogramuj w języku Python Filtr Blooma. Skorzystaj z funkcji MurMurHash z biblioteki mmh3 (instalacja: pip install mmh3). Do implementacji tablicy wykorzystaj tablicę bitów (skorzystaj z bibliteki bitarray). Filtr zrealizuj jako obiekt. Przetestuj działanie filtru Blooma na słowach z pliku Hamlet.txt (użyj 8 funkcji haszujących, ustaw rozmiar tablicy na liczę słów w Hamlet.txt).

3 Sampling

Zadanie 16 — Pokaż, że jeśli $\mathcal{H}$ jest $(k+1)$-niezależną rodziną haszującą, to jest również k-niezależną rodziną haszującą.

Zadanie 17 — Pokaż, że 2-niezależna rodzina funkcji haszujących jest rodziną uniwersalną.

Zadanie 18 — Pokaż, że jeśli $\mathcal{H}$ jest 2- niezależną rodziną funkcji haszujących z U do V, to dla dowolnych $x \in U$ oraz $v \in V$ mamy
$$\Pr_{h \leftarrow \mathcal{H}}[h(x) = y] = \frac{1}{|V|} \ .$$

Zadanie 19 — Załóżmy, że $1 < n < p$. Niech X będzie zmienną losową o rozkładzie jednostajnym na zbiorze $\{0, \ldots, p-1\}$. Niech $\phi(x) = x \mod n$. Wyznacza rozkład zmiennej losowej $\phi \circ X$.

Zadanie 20 — Załóżmy, że $h : U \to R$ jest funkcją różnowartościową. Pokaż, że jednoelementowa rodzina $\mathcal{H} = \{h\}$ jest rodziną uniwersalną ale nie jest 2-niezależna.

Zadanie 21 — Jak z liczb $S = \sum_{i=1}^n x_i$, $SS = \sum_{i=1}^n x_i^2$ oraz n możesz wyznaczyć wariancję $\frac{1}{n}\sum_{i=1}^n (x-\mu)^2$, gdzie μ oznacza średnią $\mu = \frac{1}{n}\sum_{i=1}^n x_i$?

Zadanie 22 — Zaimplementuj prostą (z użyciem generatora liczb pseudolosowych po wczytaniu każdego elementu) wersję algorytmu **R** Vittera. Pobierz z sieci notowania dzienne bitcoina z ostatnich 5 lat (możesz posłużyć się biblioteką Pandas języka Python), wydobądź z danych notowania otwarcia, wygeneruj losową próbkę 40 elementów i wygeneruj wykresy notowań i losowej próbki.

Zadanie 23 — Ustalmy liczby naturalne $1 \leqslant k \leqslant n$. Rozważamy przestrzeń probabilistyczną $[n]^k = \{X \subseteq \{1,\ldots n\} : |X| = k\}$ z prawdopodobieństwem jednostajnym ($\Pr[X] = \binom{n}{k}^{-1}$). Ustalmy zbiór $A \subseteq \{1,\ldots,n\}$ taki, że $|A| = k-1$. Rozważmy następujący proces: (1) losujemy $B' \in [n]^k$; (2) z wylosowanego B' usuwamy losowo wybrany element (każdy z prawdopodobieństwem $\frac{1}{k}$) i otrzymujemy zbiór B.

1. Sprecyzuj powyższe rozumowanie korzystając z przestrzeni probabilistycznej $[n]^k \times \{1,\ldots,k\}$
2. Wyznacz prawdopodobieństwo otrzymania zbioru A.

Zadanie 24 — **(Własności dystrybuanty)** Celem tego zadania jest przypomnienie sobie podstawowych własności dystrubuant zmiennych losowych o wartościach w liczbach rzeczywistych.

1. Niech F_X będzie dystrybuantą zmiennej losowej X (czyli $F_X(x) = \Pr[X \leqslant x]$).
   Pokaż, że $\lim_{x\to-\infty} F(x) = 0$ oraz $\lim_{x\to\infty} F(x) = 1$
2. Niech F_X będzie dystrybuantą zmiennej losowej X. Pokaż, że F jest prawostronnie ciągła w każdym punkcie x, czyli, że dla każdego $a \in \mathbb{R}$ mamy $\lim_{x\to a+} F(x) = F(a)$.
3. Załóżmy, że F_X jest ostro rosnąca oraz, że $\mathrm{rgn}(F) = (0,1)$. Niech U będzie zmienną losową o rozkładzie jednostajnym w odcinku $(0,1)$. Pokaż, że zmienna losowa $F^{-1} \circ U$ ma taki sam rozkład, co zmienna X.
4. Załóżmy, że F jest dystrybuantą zmiennej losowej X. Uogólnioną odwrotnością dystrybuanty F nazywamy funkcję zdefiniowaną wzorem

   $$F^{\leftarrow}(p) = \inf\{x : F(x) \geqslant p\}\ .$$

   Zbadaj podstawowe własności tej funkcji (np. $F^{\leftarrow}$ jest niemalejąca, $F^{\leftarrow}(F(x)) \leqslant x$, $F(F^{\leftarrow}(p)) \geqslant p$) oraz pokaż, że $F^{\leftarrow} \circ U$ ma taki sam rozkład co zmienna X, gdzie U, podobnie jak w poprzednim punkcie, ma rozkład jednostajny na odcinku $(0,1)$.

Zadanie 25 — Zaimplementuj podstawową wersję algorytmu Bravermana, Ostrovsky'iego, Zaniolo z pracy *Optimal sampling from sliding windows*.

1. Sprawdź poprawność działania implementacji generując odpowiedni histogram (możesz użyć polecenia `plt.hist(sample, density=True, bins=N)` języka Python) ze wskazywanych przez ten algorytm elementów.
2. Przetestuj swoją implementację dla okna długości 5 i po zaobserwowaniu 10000 elementów. Po wczytaniu każdego elementu zapamiętaj w jakiejś strukturze pozycję wskazywanego elementu. Zastosuj test χ^2 p-wartością $p = 0.01$ dla hipotezy zerowej

   $$H_0 = \text{próbka pochodzi z rozkładu jednostajnego}$$

   (wartość krytyczna dla tej wartości p oraz 4 stopni swobody wynosi 11.345). Możesz też posłużyć się biblioteką `scipy.stats` do przeprowadzenia tego testu.

c.d.n.
Powodzenia,
Jacek Cichoń