Big Data
Lista zadań

Jacek Cichoń, WiT, PWr, 2021/22

1 Wstęp

Zadanie 1 — Znajdź źródła swojej ulubionej książki. Zapisz je w formacie utf-8.

1. Zaimportuj bibliotekę io.Source (import scala.io.Source)
2. Skorzystaj z polecenia Source.fromFile(source, "UTF-8") do wczytania książki, zamień plik na łańcuch (mkString) i następnie podziel na wyrazy (split("\\s+")). Możesz to zrobić jednym poleceniem.
3. Usuń z tej listy stop-words (możesz ja znaleźć na stronie https://pl.wikipedia.org/wiki/Wikipedia:Stopwords (coś w stylu Book.filterNot(Stop.contains(_)))
4. Przekształć zbudowaną listę słów w kolekcję par typu (String,Int) postaci (słowo,1)
5. Pogrupuj listę par (coś w stylu Filtered.groupBy(x=¿x._1))
6. Zredukuj słowa (coś w rodzaju Groupped.mapValues(x=¿x.length))
7. Posortuj według drugiego parametru (np. reduced.toSeq.sortWith((x,y)=¿x._2¿y._2))
8. Zapisz wynik do pliku. Uwaga: możesz skorzystać z obiektu PrintWriter z bibliotek java.io
9. Wyświetl kilkadziesiąt pierwszych elementów. Usuń z niej kilkanaście początkowych elementów i zapisz listę do pliku tekstowego.
10. 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 książki "Pan Tadeusz"):

Zadanie 2 — To jest kontynuacja poprzedniego zadania.

1. Podziel swoją książkę na rozdziały.
2. Każdy rozdział potraktuj jako dokumenty.
3. Podziel dokumenty na słowa. Wyznacz indeksy TF.IDF wszystkich słów we wszystkich rozdziałach
4. Zbuduj chmury wyrazów dla wszystkich rozdziałów i jedną chmurę dla całego dokument.

Zadanie 3 — Załóżmy, że mamy dostęp do bazy zakupów klientów w sieci hurtowni środków chemicznych z poprzedniego roku. W ciągu roku 10^7 klientów odwiedza ją 10 razy i za każdym razem kupuje średnio 10 różnego typu produktów z puli 200 dostępnych typów produktów. Załóżmy że znaleźliśmy w tej bazie danych dwóch klientów którzy zakupili choć raz ten sam koszyk produktów. Czy jest to czysty przypadek?

2 Model MapReduce

Zadanie 4 — Wymień jakie aspekty działania systemu MapReduce są poza zasięgiem programisty. Które elementy kontroluje programista?

Zadanie 5 — Wracamy do zadania "'Word Count"'. Przepisz algorytmy które już napisałaś w języku Scala do programów działający w paradygmacie Map Reduce

Zadanie 6 — Zaprojektuj algorytm MapReduce który dostaje bardzo duży zbiór liczb całkowitych i produkuje na wyjściu jednocześnie:

1. Największą liczbę, najmniejszą liczbę.
2. Średnią wszystkich liczb.
3. Ten sam zbiór ale bez powtórzeń.
4. Liczbę różnych elementów bez powtórzeń.

Zadanie 7 — Niech $x \oplus y = x + y + 1$ oraz $x \otimes y = xy + x + y$ dla $x, y \in \mathbb{R}$. Pokaż, że są to działania łączne i przemienne na $\mathbb{R}$. Wskazówka: Spróbuj to zrobić z minimalną liczbą rachunków; rozważ funkcje $f(x) = x + 1$ oraz $g(x) = x - 1$; zauważ, że $g \circ f = Id$.

Zadanie 8 — Podaj kilka przykładów działań nieprzemiennych. Podaj kilka przykładów działań które nie są łączne.

Zadanie 9 — Pokaż, że operacje $\min(x, y)$ i $\max(x, y)$ są przemienne i łączne. Czy operacja $s(x, y) = \frac{x+y}{2}$ jest łączna?

Zadanie 10 — Wymień jakie aspekty działania systemu MapReduce są poza zasięgiem programisty. Które elementy kontroluje programista?

Zadanie 11 — Zaprojektuj algorytm MapReduce który dostaje bardzo duży zbiór liczb całkowitych i produkuje na wyjściu:

1. Największą liczbę.
2. Średnią wszystkich liczb.
3. Ten sam zbiór ale bez powtórzeń.
4. Liczbę różnych elementów bez powtórzeń.

Zadanie 12 — Zaprojektuj algorytm MapReduce, który wyznacza złączenie dwóch relacji o schemacie R(A,B,C) i S(X,Y,Z) według połączenia B=X oraz C=Y, czyli wyznacz tabelę

$$\{(A, Y) : (\exists B, C)(R(A, B, C) \wedge S(B, C, Y)\} \ .$$

Zadanie 13 — (Odwrócenie grafu) Dany jest graf w postaci listy sąsiadów: $[w, [w_{i,1}, w_{i,2}, \ldots, w_{i,n_i}]]$ zapisany w zbiorze tekstowym, np.

```
[
  [1, [3,4,5]],
  [2, [1,3]],
  [3, [4,5]],
  [4, [1,2]],
  [5, [4,5]]
]
```

Zastosuj technologię MapReduce do zbudowania grafu z odwróconymi linkami.

Zadanie 14 — Niech $F : ((\mathbb{N} \setminus \{0\}) \times \mathbb{R})^2 \to (\mathbb{N} \setminus \{0\}) \times \mathbb{R}$ będzie funkcją określoną wzorem

$$F([c_1, x_1], [c_2, x_2]) = [c_1 + c_2, \frac{c_1 x_1 + c_2 x_2}{c_1 + c_2}]$$

1. Pokaż, że F jest działaniem przemiennym i łącznym.
2. Oznaczmy przez $\oslash$ działanie $x \oslash y = F(x, y)$. Znajdź zwartą formułę dla

$$[c_1, x_1] \oslash [c_2, x_2] \oslash \ldots \oslash [c_n, x_n] \; .$$

3. Zastosuj tę własność funkcji do zastosowania combainera dla problemu wyznaczania średniej i wariancji.

Zadanie 15 — Zastosuj metodę map-reduce do wyznaczenia średniej geometrycznej i harmonicznej.

Zadanie 16 — Zastosuj metodę map-reduce do wyznaczenia wszystkich anagramów występujących w zbiorze tekstowym.

Zadanie 17 — Multizbiorem o skończonym nośniku Ω nazywamy funkcję $F : \Omega \to \mathbb{N}$. Dla $F, G : \Omega \to \mathbb{N}$ określamy $(F \cup G)(\omega) = \max\{F(\omega), G(\omega)\}$, $(F \cap G)(\omega) = \min\{F(\omega), G(\omega)\}$, $(F \setminus G)(\omega) = \max\{F(\omega) - G(\omega), 0\}$. Zaprojektuj map-reduce algorytm do wyznaczania tych trzech operacji. Algorytm na wejściu dostaje listę elementów zbioru

$$\{(1, \omega, F(\omega)) : \omega \in \Omega \wedge F(\omega) > 0\} \cup \{(2, \omega, G(\omega)) : \omega \in \Omega \wedge G(\omega) > 0\}$$

Zadanie 18 — (To nie jest zadanie na MapReduce) W pliku TwoCollisions.csv, do którego link znajduje się na stronie wykładu, w każdej linijce znajduje się (NumerHotelu, NumerDnia,NumerOsoby). Znajdź takie osoby, które w dwóch różẏch dniach znajdowały się się w tym samym hotelu.

Zadanie 19 — Zastosuj dwukrotnie MapReduce do swojej książki którą używałeś do zadania z Word - Count zrób listę pięciu najczęściej powiązanych wyrazów z danym słowem. Przez powiązane słowa rozumiemy słowa występujące obok siebie (po usunięciu stop-words). Oczywiście oprogramować masz to zadanie w paradygmacie MapReduce.
Spróbuj zastosować otrzymaną listę do wygenerowania losowego paragrafu ze swojej książki.

Zadanie 20 — Niech (G, E) będzie grafem skierowanym. Dla $g \in G$ oznaczamy $inDeg(g) = |\{y \in G : (y, g) \in E\}$ oraz $outDeg(g) = |\{y \in G : (g, y) \in E\}$. Zaprojektuj w MapReduce procedury służące do wyznaczania listy $\{(g, inDeg(g), outDeg(g)) : g \in G\}$. Napisz również procedurę do wyznaczania średnich stopni $\frac{1}{n} \sum\{inDeg(g) : g \in G\}$ oraz $\frac{1}{n} \sum\{outDeg(g) : g \in G\}$.
Zastosuj opracowane algorytmy to wyznaczenia średnich stopni dla grafu "'Stanford web graph"', który możesz znaleźć na stronie http://snap.stanford.edu/data/web-Stanford.html

Zadanie 21 — Niech (G, E) będzie grafem. Współczynnikiem klasteryzacji węzła $g \in G$ nazywamy liczbę

$$c(g) = \frac{2}{|N_g|(|N_g| - 1)} |N_g \cap E| \; ,$$

gdzie $N_g = \{a : \{g, a\} \in E\}$. Zaprojektuj w MapReduce procedury służące do wyznaczania listy $\{(g, c(g), outDeg) : g \in G\}$. Napisz również procedurę do wyznaczania średniej wartości $\frac{1}{n} \sum\{c(g) : g \in G\}$
Zastosuj opracowane algorytmy to wyznaczenia średnich stopni dla grafu "'Stanford web graph"' (musisz przerobić ten graf na graf nieskierowany).

Zadanie 22 — Mamy n serwerów (wartości liczbowe sprawdzaj dla $n = 3000$). Prawdopodobieństwo zepsucia się jednego serwera podczas wykonania zadania map-reduce przez pojedynczy serwer w czasie T wynosi p (przyjmij dla obliczeń numerycznych, że $p = \frac{1}{3000}$). Przyjmij, że zdarzenia polegające na zepsuciu się serwera w kolejnych slotach czasowych o długości T są niezależne.

1. Oblicz prawdopodobieństwo poprawnego zakończenia całego zadania

2. Dzielimy teraz serwery na 3 grupy po 1000 serwerów. Każde zadania wykonujemy na trzech serwerach (po jednym z każdej grupy). Czas wykonania zadania wynosie teraz $3T$. Jakie jest teraz prawdopodobieństwo poprawnego zakończenia zadania?

Zadanie 23 — Załóżmy, że dwie macierze A, B rozmiaru $2n \times 2n$ zostały podzielone na bloki

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

rozmiaru $n \times n$

1. Pokaż, że

$$A \cdot B = \begin{bmatrix} A_{11} \cdot B_{11} + A_{12} \cdot B_{21} & A_{11} \cdot B_{12} + A_{12} \cdot B_{22} \\ A_{21} \cdot B_{11} + A_{22} \cdot B_{21} & A_{21} \cdot B_{12} + A_{22} \cdot B_{22} \end{bmatrix}$$

2. Spróbuj sformułować i udowodnić bardziej ogólny fakt.

Zadanie 24 — Nadal zajmujemy się wybraną przez Ciebie książką podzieloną na rozdziały. Teraz dla każdego słowa oraz każdego rozdziału wyznacz współczynniki TF.IDF.

1. Dla każdego rozdziału wyznacz 20 wyrazów o najwyższymwpółczynniki TF.IDF
2. Napisz funkcję realizującą następujące zadanie: wprowadzasz do niej słowo, a ona zwraca listę najbardziej pasujących rozdziałów (czyli posortuj rozdziały wg parametryu FT.IDF.

3 Funkcje haszujące

Zadanie 25 — Załóżmy, że $\mathcal{H} = \{h_1, \ldots, h_n\}$ jest rodziną k-niezależnych funkcji haszujących ze zbioru D w zbiór skończony R, czyli, że dla dowolnych $x_1, \ldots, x_k \in D$ oraz $y_1, \ldots, y_k \in R$ mamy

$$\Pr_{h \leftarrow \mathcal{H}}(h(x_1) = y_1 \wedge \ldots \wedge h(x_k) = y_k) = \frac{1}{|R|^k}\ .$$

Pokaż, że dla dowolnego $m \leqslant k$ jest ona również m-niezależna.

Zadanie 26 — Rozważmy funkcje haszującą zadaną wzorem $h(x) = x \mod 21$. Stosujemy ją do liczb podzielnych przez pewną stałą c. Dla jakich stałych c jest to odpowiednia funkcja haszująca, czyli dla jakich stałych c można się spodziewać, że rozkład załadowania kubełków $\{0, \ldots, 20\}$ będzie jednostajny?

Zadanie 27 — Znajdź wzór na rząd elementu $k \in \{0, \ldots, n-1\}$ w grupie $C_n = (\{0, \ldots, n-1\}, \oplus_n)$? Jaki jest związek tego zadnia z poprzednim zadaniem?

Zadanie 28 — Mamy n kubełków. Rzucamy do niech k kul.

1. Oszacuj k taki aby z dużym prawdopodobieństwem doszło do 3-kolizji, czyli aby a jakimś kubełku znalazły się 3 kulki.
2. Sprawdź eksperymentalnie otrzymany wynik
3. Uogólnij zadanie na a - kolizje
4. Wyznacz w podobny sposób wartość oczekiwaną liczby pustych urn. Kiedy (dla jakich k) ta liczba staje się mniejsza od 1.

Zadanie 29 — Dwóch studentów ma dzban wypełniony 8 litrami napoju. Mają do dyspozycji dzbanek o pojemności 5 litrów oraz drugi dzbanek o pojemności 3 litrów. Chcą podzielić się równo napojem. Jak mogą to zrobić? Zagadnie to można potraktować jako system przepisujący o stanie początkowym $(8, 0, 0)$. Możemy to zadanie próbować rozwiązać tak: losowo wybieramy dwie różne liczby i, j ze zbioru $\{1, 2, 3\}$ i próbujemy przelać napój z i-tego pojemnika do j-tego pojemnika; iterujemy to błądzenie losowe tak długo aż dojdziemy do stanu $(4, 4, 0)$. Jednak jest to kiepskie rozwiązanie - algorytm taki wpada bardzo często w pętle. Zastosuj technikę śledzenia przebiegu do unikania zapętleń. Oprogramuj to w języku Python.

4 Podobieństwo tekstów

Zadanie 30 — Pokaż, że funkcja $d(A,B) = |A \triangle B|$ jest metrykę na przestrzeni niepustych skończonych podzbiorów ustalonego zbioru X.

Zadanie 31 — Niech $f : [0,\infty) \to [0,\infty)$ będzie funkcją rosnącą i wklęsłą.

1. Pokaż, że dla $a, b \geqslant 0$ mamy $f(a+b) \leqslant f(a) + f(b)$.
   Wskazówka: Zauważ, że możemy założyć, że $a + b > 0$; następnie zauważ, że $a = (a+b)\frac{a}{a+b}$ oraz $b = (a+b)\frac{b}{a+b}$ i zastosuj nierówność Jensena dla funkcji wklęsłych.
2. Załóżmy dodatkowo, że $f(0) = 0$. Niech d będzie metryką na zbiorze X. Pokaż, że funkcja $\rho(x,y) = f(d(x,y))$ jest również metryką na zbiorze X.
3. Pokaż, że jeśli $\epsilon \in (0,1)$ oraz d jest metryką na zbiorze X, to funkcja $\rho(x,y) = d(x,y)^{\epsilon}$ jest metryką na zbiorze X.
4. Pokaż, że jeśli d jest metryką na zbiorze X, to funkcja $\rho(x,y) = \frac{d(x,y)}{1+d(x,y)}$ jest metryką na zbiorze X.

Zadanie 32 — Wybierzmy dwa losowe m-elementowe podzbiory A, B n elementowego zbioru X. Jaka jest wartość oczekiwana podobieństwa Jaccarda $J(A,B)$?

Zadanie 33 — Korzystając z Twierdzenia o Gęstości Liczb Pierwszych (Prime Numbers Theorem) oszacuj liczbę liczb pierwszych z przedziału $[2^{64}, 2^{64} + 1000]$ i następnie wyznacz te liczby.

Zadanie 34 — (Twierdzenie Steinhausa) Niech d będzie metryką na zbiorze X. Ustalmy element $a \in X$ i zdefiniujmy funkcję

$$\rho(x,y) = \frac{2d(x,y)}{d(x,a) + d(y,a) + d(x,y)}$$

Celem tego zadania jest pokazanie, że ρ jest metryką na zbiorze X.

1. Pokaż najpierw, że jeśli $0 < p \leqslant q$ oraz $r \geqslant 0$ to $\frac{p}{q} \leqslant \frac{p+r}{q+r}$.
2. Wprowadź oznaczenia $p = d(x,y)$, $q = d(x,y) + d(x,a) + d(y,a)$ oraz $r = d(x,z) + d(y,z) - d(x,y)$ i zastosuj obserwację z poprzedniego punktu do pokazania nierówności trójkąta dla funkcji ρ.

Zadanie 35 — Zastosuj Twierdzenie Steinhausa do metryki $d(X,Y) = |X \triangle Y|$ na zbiorze skończonych podzbiorów zbioru Ω do pokazania, że funkcja $d(X, Y0 = 1 - S(X,Y)$ (odległość Jaccarda) jest metryką.

Zadanie 36 — Załóżmy, że S jest takim podobieństwem obiektów przestrzeni Ω, że istnieje rodzina funkcji haszujących $\mathcal{H}$ oraz prawdopodobieństwo na rodzinie $\mathcal{H}$ takie, że dla dowolnych dwóch obiektów $A, B \in \Omega$ mamy

$$P_h[h(A) = h(B)] = S(A,B)$$

Pokaż, że wtedy funkcja $d(A,B) = 1 - S(A,B)$ jest metryką na zbiorze Ω.

Zadanie 37 — Uzupełnij szczegóły dowodu tego, że jeśli $\Omega = \{\omega_i : 1 \leqslant N\}$, π jest losową permutacją zbioru $\{1,\ldots,N\}$ (wybieraną zgodnie z rozkładem jednostajnym), oraz $h_\pi(X) = \min\{k : \omega_{\pi(k)} \in X\}$ dla $X \subseteq \Omega$ to

$$P_\pi[h_\pi(A) = h_\pi(B)] = S(A,B)\ .$$

Zadanie 38 — Napisz procedurę o specyfikacji jaccard(f1:String,f2:String,k:Integer):Double, która dla plików o nazwach f1, f2 wyznacza ich k-gramy i następnie wylicza ich odległość Jaccarda. Przed wyznaczeniem k-gramów pliki powinne być oczyszczone (minimum to usunięcie znaków nowej linii, tabulatorów oraz podwójnych spacji)

1. Zastosuj tę procedurę do kilku wariantów swoich plików z algorytmami (zastosuj 4-gramy)
2. Zastosuj tę procedurę do porównania kolejnych rozdziałów analizowanej w Zadaniu 2 książki (zastosuj 7-gramy)

Zadanie 39 — Zastosuj metodę minhash do poprzedniego zadania. Twoja procedura powinna zależeć od paramertu H który określa liczbę funkcji haszujących stosowanych do budowania sygnatury tekstu.

1. Przetestuj tę procedurę na danych z poprzedniego zadania dla $H \in \{50, 100, 250\}$ - porównaj aproksymację odległości Jaccarda z jej dokładnymi wartościami.

Pamiętaj o wygenerowaniu wspólnej rodziny funkcji haszujących dla wszystkich analizowanych tekstów.

Zadanie 40 — Napisz procedurę służącą do wyznaczania sygnatur kosinusowych plików tekstowych korzystających z 1024 losowych wektorów z $\mathbb{R}^n$ (n tutaj oznacza moc wspólnego zbioru słów występujących w badanych dokumentach). Dokumenty reprezentowane mają być przez wektor częstotliwości słów. Zastosuj tę metodę do plików z Zadanie 2.

5 Streaming

Zadanie 41 — Niech C_n będzie wartością licznika Morrisa po n krotnym wywołaniu procedury INCREMENT. Niech $L_n = 2^{C_n}$.

1. Wyznacz wariancję zmiennej L_n oraz oblicz $\frac{\sigma(L_n)}{E(L)n)}$.
2. Zbadaj eksperymentalnie dokładność zastawu czterech liczników Morrisa. Jako estymator liczby n przyjmij $2^{(C_1(n)+C_2(n)+C_3(n)+C_4(n))/4}$.
3. Dla jakich n mamy $4\log_2(\log_2(n)) < \log_2(n)$?

Zadanie 42 — Zaimplementuj Boyer–Moore'a Majority Algorithm.

1. Napisz najpierw funkcję której parametrem jest lista łańcuchów (List[String]).
2. Zaprojektuj następnie obiekt o dwóch metodach: add(x:String):Unit oraz get():String który realizuje ten algorytm.
3. Jaka jest złożoność obliczeniowa i pamięciowa tego algorytmu.

Zadanie 43 — Zaimplementuj Misra - Gries Algorithm.

1. Napisz najpierw funkcję której parametrami jest lista łańcuchów (List[String]) oraz liczba k określającą maksymalną liczbę śledzonych obiektów.
2. Zaprojektuj następnie obiekt o dwóch metodach: add(x:String):Unit oraz get():String który realizuje ten algorytm. Do utworzenia tego obiektu potrzebujesz jeden parametr k.
3. Jaka jest złożoność obliczeniowa i pamięciowa tego algorytmu.

Zadanie 44 — Algorytm HyperLogLog używa wartości $h(x) = (b_0 b_1 b_2 b_3 \ldots)$ do wyznaczenia numeru inkrementowanego licznika ($i = (b_0 \ldots b_{k-1})_2 + 1$) oraz do wyznaczenia z reszty ciągu bitów $(b_k b_{k+1} \ldots)$ do zwiększenia wartości licznika. Załóżmy, że $h(x)$ jest typy Int lub Long oraz, że $h(x) \geqslant 0$.

1. Jak można za pomocą operacji bitowych wyznaczyć z $h(x)$ ciąg $(b_0 \ldots b_{k-1})$?
2. Jak można za pomocą operacji bitowych wyznaczyć z $h(x)$ ciąg $(b_k k_{k+1} \ldots)$?
3. Załóżmy, że $n \geqslant 0$. Co robi operacja $n\&(-n)$. Jak tą operację możesz wykorzystać do inkrementacji licznika.

Zadanie 45 — Zaimplementuj algorytm HyperLogLog.

1. Pobierz ze strony http://ita.ee.lbl.gov/html/contrib/LBL-PKT.html plik lbl-pkt-4. Wypakuj z niego plik lbl-pkt-4.tcp. Oto format danych: timestamp, (przenumerowany) source host, (przenumerowany) destination host, source TCP port, destination TCP port, liczba bajtów danych(zero dla "pure-ack" pakietów).
2. Zastosuj HupeLogLog do wyznaczenia liczby różnych source hostów, liczby różnych destination hostów oraz liczby różnych par (source,destination).

Zadanie 46 — Pokaż, że wielomian $w(x) = 1+x+x^2$ jest nierozkładalny w pierścieniu $\mathbb{Z}_2[x]$. Rozważ ideał $(w) = \{\alpha \cdot w : \alpha \in \mathbb{Z}_2[x]\}$ w pierścieniu $\mathbb{Z}_2[x]$. Obliczenia będziemy prowadzić w pierścieniu ilorazowym $\mathbb{Z}_2[x]/(w)$.

1. Pokaż, że dla $\alpha, \beta \in \mathbb{Z}_2[x]$ mamy

$$((w) + \alpha = (w) + \beta) \equiv w|(\alpha - \beta)$$

2. Pokaż, że struktura $(\mathbb{Z}_2[x]/(w), +)$ jest izomorficzna z $(\mathbb{Z}_2^2, +)$.
3. Oznacz przez i zmienną x. Zauważ, że $\mathbb{Z}_2[x]/(w) = \{[0], [1], [i], [1+i]\}$.
4. Stosując oznaczenia: $\mathbf{0} = [0]$, $\mathbf{1} = [1]$, $\mathbf{i} = [i]$ oraz $\mathbf{1} + \mathbf{i} = [1+i]$ wyznacz tabliczki dodawania i mnożenia w $\mathbb{Z}_2[x]/(w)$
5. Pokaż, że $\mathbf{1} + \mathbf{i} + \mathbf{i}^2 = \mathbf{0}$.
6. Pokaż, że $\mathbb{Z}_2[x]/(w)$ jest czteroelementowym ciałem.

Zadanie 47 — Powtórz poprzednie zadanie dla wielomianu $w(x) = 1+x+x^3$. Skonstruujesz w ten sposób ciało 8 elementowe.

Zadanie 48 — Zbuduj ciało 9 elementowe.

Zadanie 49 — Niech F będzie ciałem. Niech $a_1, \ldots, a_4$ będą parami różnymi elementami ciała F. Niech $b_1, \ldots, b_4$ będą dowolnym elementami ciała F. Niech

$$w(x) = \sum_{i=1}^{4} b_j \prod_{j \neq i} \frac{x - a_j}{a_i - a_j}$$

(jest to wielomian interpolacyjny Legrange'a stopnia 4). Pokaż, że w jest jednym wielomianem stopnia trzeciego z pierścienia $F[x]$ takim, że $w(a_1) = b_1$, $w(a_2) = b_2$, $w(a_3) = b_3$ oraz $w(a_4) = b_4$.

Zadanie 50 — Wygeneruj listę $\{h_1, \ldots, h_{100}\}$ losowych wielomianów stopnia 3 nad ciałem $\mathbb{Z}_{11}$.
Wskazówka: W programie Mathematica można to zrobić za pomocą następującej funkcji:
LP[p_]:= Module[{},Function[x,Mod[RandomInteger[{0,p-1},4].{$1, x, x^2, x^3$},p]]]; .

1. Wyznacz moc zbioru $\{i \in \{1, \ldots, 100\} : h_i(1) = 2\}$. Powtórz ten eksperyment kilka razy. Pamiętaj aby obliczenia wykonywać w ciele $\mathbb{Z}_{11}$.
2. Wygeneruj histogram wartości $\{h_i(1) : i \in \{1, \ldots, 100\}\}$.

Zadanie 51 — Zapoznaj się z testem nierozkładalności wielomianów Rabina i pokaż, że wielomian $w(x) = x^{80}+x^9+x^4+x^2+1$ jest wielomianem nierozkładalnym nad ciałem $\mathbb{Z}_2$. Oblicz $[x^{40}]\cdot[x^{40}]$ w ciele $\mathbb{Z}_2[x]/(w)$.

Zadanie 52 — Oprogramuj Geometric Histogram Streaming Window Algorithm M. Datara, A. Gionisa, P. Indyka i R. Motwaniego z pracy http://www-cs-students.stanford.edu/˜datar/papers/sicomp˙streams.pdf

6 Page Rank

Zadanie 53 — Niech $A = (a_{ij})$ będzie kwadratową macierzą rozmiaru $n \times n$. Niech $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Pokaż, że

$$||Ax^T||_2 \leqslant \sqrt{\sum_{i,j} a_{ij}^2} \cdot ||x||_2 \ .$$

(gdzie $||(y_1, \ldots, y_n)||_2 = \left(\sum_i x_i^2\right)^{1/2}$). Wywnioskuj z tego, że jeśli λ jest wartością własną macierzy kolumnowo stochastycznej, to $|\lambda| \leqslant 1$.

Zadanie 54 — Załóżmy, że A jest macierzą kwadratową o współczynnikach rzeczywistych. Pokaż, że jeśli λ jest wartością własną macierzy A, to liczba $\overline{\lambda}$ (sprzężenie liczby λ) jest również wartością własną macierzy A.

Zadanie 55 — Załóżmy, że $A, B \in M_n(\mathbb{R})$ są kolumnowo stochastyczne.

1. Pokaż, że macierz $A \circ B$ jest kolumnowo stochastyczna.
2. Niech $\alpha \in [0, 1]$. Pokaż, że macierz $\alpha A + (1-\alpha)B$ jest kolumnowo stochastyczna.
3. Podaj interpretacje probabilistyczne powyższych faktów.

Zadanie 56 — Załóżmy, że w grafie nie ma 'wiszących wierzchołków'. Niech v będzie wierzchołkiem bez linków do tego wierzchołka. Pokaż, że $pagerank(v) = \frac{1-\alpha}{n}$.

Zadanie 57 — Rozważmy gwiazdę rozmiaru $n + 1$, czyli graf o krawędziach $\{i \to n + 1, i = 1 \ldots, n\}$. Wyznacz PageRank dla tego grafu.

Zadanie 58 — Wyznacz PageRank dla grafu $\{a \to b, b \to a, b \to c\}$.

Zadanie 59 — Napisz procedurę służącą do wygenerowania macierzy Google dla porządku liniowego rozmiaru n. Napisz procedurę 'naiwnego' obliczania PageRank metodą potęgową i zastosuj ją wyznaczenia PageRank dla tego liniowego porządku.

Zadanie 60 — Przeskanuj witrynę WWW któregoś z pracowników katedry. Wyznacz PageRank dla wszystkich wierzchołków tego grafu.

Zadanie 61 — Napisz pseudo-kod funkcji MAPPER i REDUCER służących do wykonania jednego kroku iteracyjnego wyznaczenia PageRank metodą polegającą na rozbiciu macierzy przejść M na k^2 bloków:

$$\begin{bmatrix} M_{11} & \ldots & M_{1k} \\ \vdots & \vdots & \vdots \\ M_{k1} & \ldots & M_{kk} \end{bmatrix} \circ \begin{bmatrix} V_1 \\ \vdots \\ V_k \end{bmatrix} = \begin{bmatrix} M_{11} \circ V_1 + \ldots + M_{1k} \circ Vk \\ \vdots \\ M_{k1} \circ V_1 + \ldots + M_{kk} \circ Vk \end{bmatrix}$$

1. Pokaż poprawność metody mnożenia macierzy przez wektor za pomocą podziału na bloki.
2. Niech R będzie pierścieniem łącznym. Pokaż, że

   $$(M_{n\times n}(M_{k\times k}(R)), \circ) \cong (M_{(nk)\times(nk)}(R), \circ)$$

   (gdzie $M_{m\times m}(R)$ oznacza zbiór macierzy kwadratowych wymiaru $m \times m$ o wyrazach z pierścienia R).

Zadanie 62 — Wybierz witrynę jednego z pracowników Katedry Informatyki. Sprawdź jej poprawność za pomocą narzędzi ze stron

•https://validator.w3.org/,

•https://jigsaw.w3.org/css-validator/

•https://search.google.com/search-console/mobile-friendly.

Sprawdź następnie poprawność meta-informacji i strukturę semantyczną głównej strony.

Zadanie 63 — Niech L oznacza zbiór wszystkich liniowych porządków na zbiorze $X = \{a, b\}$. Ile jest funkcji $F : L^n \to L$ spełniających zasadę Pareto?

Zadanie 64 — Niech L oznacza zbiór wszystkich liniowych porządków na zbiorze $X = \{a, b, c\}$. Ile jest funkcji $F : L^n \to L$ spełniających zasadę Pareto oraz zasadę uczciwości (niewrażliwości na trzecią możliwość)?

c.d.n.
Powodzenia,
Jacek Cichoń