Metoda Saltona

Zadanie1

Przedstaw graficznie reprezentację pni i grup w wyszukiwaniu strukturalnym. Masz do dyspozycji:

P_1=[0, 10, 11, 2, 7, 3, 4, 0, 1, 0]
P_2=[10, 0, 0, 9, 2, 11, 2, 10, 7, 0] poziom pni
P_3=[1, 0, 1, 12, 10, 3, 10, 4, 6, 11]
Oraz

G_{11}=[0, 6, 7, 2, 3, 4, 0, 1, 3, 2]
G_{12}=[1, 8, 9, 1, 4, 0, 2, 3, 2, 2, 1]
G_{21}=[0, 1, 3, 12, 1, 6, 0, 9, 10, 1]
G_{22}=[1, 3, 2, 6, 2, 5, 0, 10, 9, 3] poziom grup
G_{31}=[1, 0, 3, 9, 10, 0, 12, 2, 4, 0]
G_{32}=[2, 1, 3, 10, 8, 1, 13, 1, 2, 1]
G_{33}=[3, 4, 2, 11, 7, 0, 12, 2, 1, 3]
Omów sposób wyszukiwania dla pytania:

q=[5, 3, 0, 10, 6, 1, 13, 2, 0, 1]

Wykorzystaj odpowiedni wzór na korelację (f. podobieństwa) w omawianym procesie wyszukiwania.

Zadanie2

Wiedząc, że dwa dokumenty d_1 oraz d_2 mają następujące opisy:

d_1 =The old lady said that her daughter disliked John

d_2 =John got crazy because of the old lady's daughter

Zbuduj macierz termin-dokument dla tych dwóch dokumentów i oblicz podobieństwo między nimi stosując miarę cosinus i ovlap. Skomentuj uzyskane wyniki.

Zadanie 3

Mając dane opisy dokumentów: D_1=*abc,* D_2=*aef,* D_3=*bcd,* D_4=*abd,* D_5=*ab,* D_6=*af,* D_7=*afg,* D_8=*dec* dokonaj grupowania dokumentów o podanych opisach stosując algorytm Roccio'a.

Parametry: $N_1 = 2$, $N_2 = 4$,$p_1 = 0{,}6$, $p_2 = 0{,}24$ $N_{1c} = 2$, $N_{2c}=4$, $p_{1c} = 0{,}75$, $p_{2c} = 0{,}5$.

Zadanie 4

Przedstaw graficznie reprezentację pni i grup w wyszukiwaniu strukturalnym. Masz do dyspozycji:

P_1 = [0,5,5,1,3,1]
P_2 = [5,0,0,4,1,5]
P_3 = [0,0,0,6,5,1]
G_{11} = [0,3,3,1,1,2]
G_{12} = [0,4,4,0,2,0]
G_{21} = [0,0,1,6,0,3]
G_{22} = [0,1,1,3,1,2]
G_{31} = [0,0,1,4,5,0]
G_{32} = [1,0,1,4,5,0]
G_{33} = [1,2,1,5,3,0]

Gdzie P_1-P_3 – poziom pni zaś G_{11}-G_{33} – poziom grup.

Omów sposób wyszukiwania dla pytania q=[2,1,0,5,3,0]. Wykorzystaj w tym celu wzór na korelację w omawianym procesie wyszukiwania.

Omówić wpływ parametru T na moc wiązania dokumentów w grupy dla

$\alpha = 0{,}25$, $\alpha = 0{,}5$ i $\alpha = 0{,}75$.

Przedstaw różne metody ustalania wartości T dla algorytmu Doyle'a, gdy dane są wartości funkcji punktujących g (D_i , P_j).

Zadanie 5

Dany jest zbiór obiektów X = {x_1,...,X_{10}}, które są opisane pojęciami:

x_1: adfg
x_2: bcdfhij
x_3: aeij
x_4:defgh
x_5: cehij
x_6:adf
x_7: bcgj
x_8:afghi
x_9:dejf
x_{10}:ghij
Dla w/w zbioru obiektów dokonać podziału na 3 grupy algorytmem Doyle`a.

Zadanie 6

Dla systemu zawierającego opisy płyt głównych dla zestawów komputerowych przedstawionego za pomocą tabeli:

Model	Socket	Chipset	RAM	Szyna	VGA
Asus M4A785D-M Pro	AM2	AMD785G	DDR2	1200	tak
Gigabyte GA-M68M-S2P	AM2	GF7025	DDR2	1066	tak
MSI KA790GX	AM2	AMD790G	DDR2	1066	tak
Asus M4A78LT-M	AM3	AMD760G	DDR3	1800	tak
Gigabyte GA-870A-UD3	AM3	AMD870	DDR3	1866	nie
MSI 770-C45	AM3	AMD770	DDR3	1600	nie
Asus P7H55-M	s1156	Intel H55	DDR3	2200	tak
Gigabyte GA-H55-UD3H	s1156	Intel H55	DDR3	2200	nie
MSI H55-G43	s1156	Intel H55	DDR3	2133	nie
Asus P6T Deluxe V2	s1366	Intel X58	DDR3	2000	nie
Gigabyte GA-X58A-UD3	s1366	Intel X58	DDR3	2200	nie
MSI X58A-GD65	s1366	Intel X58	DDR3	2133	nie

przeprowadź jedną iterację grupowania płyt głównych dla zestawów komputerowych algorytmem Doyle'a dla w/w systemu. Utwórz 2 grupy (rozdzielając płyty z obsługą standardu graficznego VGA od tych, które takiego wyposażenia nie posiadają). Jako współczynnik skalujący przybierz wartość optymalną, zaś wartość parametru T wyznacz za pomocą formuły: T = 1/2(min(g(di.Pj)) +max(g(di.Pj))), gdzie g(di.Pj) określa wartość funkcji punktującej dokumentu di w profilu Pj.

Zadanie 7

Dany jest następujący zbiór dokumentów:

D_1: The King University College
D_2: King College Site Contents
D_3: University of King College
D_4: King County Bar Association
D_5: King County Government Seattle Washington
D_6: Martin Luther King
Oraz zbiór termów: T={*The, King, University, College, Site, Contents, of, County, Bar, Association, Government, Seattle, Washington, Martin, Luther*}Stwórz macierz termin-dokument oraz oblicz podobieństwo między dokumentami D_2 oraz D_4 stosując miarę COSINUS.Wiedząc, że na pytanie o dokumenty zawierające termin „king" znaleziono następujące dokumenty: D_1,D_3,D_4 oraz D_6 zbadaj parametry efektywności takiego systemu wyszukiwawczego i zinterpretuj je odpowiednio.

Zadanie 8

Przedstaw graficznie reprezentację pni i grup w wyszukiwaniu strukturalnym. Masz do dyspozycji:

$P_1 = [0,5,5,1,3,1]$
$P_2 = [5,0,0,4,1,5]$
$P_3 = [0,0,0,6,5,1]$
$G_{11} = [0,3,3,1,1,2]$
$G_{12} = [0,4,4,0,2,0]$
$G_{21} = [0,0,1,6,0,3]$
$G_{22} = [0,1,1,3,1,2]$
$G_{31} = [0,0,1,4,5,0]$
$G_{32} = [1,0,1,4,5,0]$
$G_{33} = [1,2,1,5,3,0]$
Omów sposób wyszukiwania dla pytania q=[2,1,0,5,3,0]. Wykorzystaj w tym celu wzór na korelację w omawianym procesie wyszukiwania.

Zadanie 9

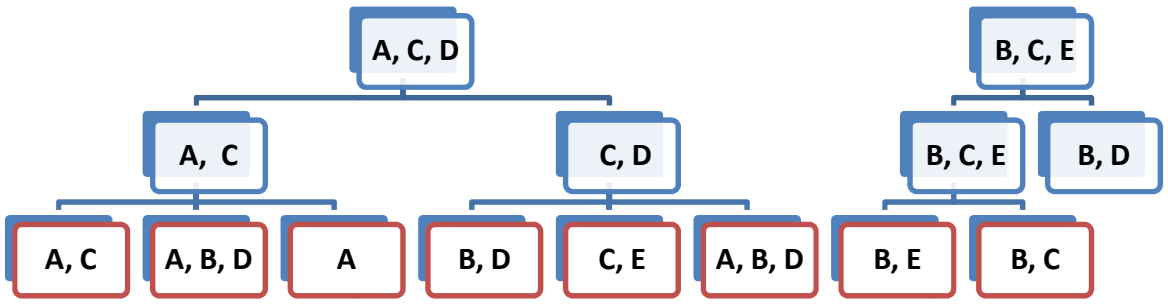
W kartotece wtórnej znajdują się dokumenty o opisach:

c_i	1	2	3	4	5	6	7	8	9	10
1				1		1	1	1	1	
2	1			1	1			1	1	1
3	1	1			1		1			
4		1	1		1			1		
5			1			1				
6			1			1	1		1	1

Przedstawić proces zakładania bazy danych zgodnie z algorytmem Doyle`a. Omówić parametry bazy danych następujących wartości parametru α:$\alpha_1 = 0,2$; $\alpha_2 = 0,5$; $\alpha_3 = 0,8$. Założona liczba grup m = 3.

Zadanie 10

Rysunek przedstawia system po grupowaniu dokumentów.



Omów sposób wyszukiwania dla pytania q=AC zgodny z metodą pnia najbardziej obiecującego. Zakładając, że w odpowiedzi system zwrócił dokumenty o opisach: [B,C], [A], [A,C] oblicz parametry efektywności wyszukiwania (kompletność, dokładność) w tym systemie.

Proces wyszukiwania dla pytania *q* będzie polegał na obliczeniu współczynnika korelacji dla każdego pnia i wybraniu pnia najbardziej obiecującego czyli tego który posiada największy współczynnik.
Współczynnik korelacji jest liczbą pojęć wspólnych(pytania i dokumentu) do wszystkich pojęć w dokumencie.

Zadanie 11

Jakie są najwyższe możliwe wartości parametrów N_1 oraz N_2 dla algorytmu Rocchia, aby dokument D_3 przeszedł poprawnie test gęstości dla następującej kartoteki wtórnej:
D_1=[ABC]
D_2=[BC]
D_3=[AEF]
D_4=[FCB]
D_5=[FAB]
D_6=[AEB]
D_7=[FEC]
D_8=[AE]
Przyjmij, że $p_1 = 0{,}43$, $p_2 = 0{,}58$. Jaka będzie wtedy wartość p_{min}?
Do określenia współczynnika korelacji należy użyć funkcji:

$$f(d_i, d_j) = \frac{card(p_i \cap p_j)}{card(p_i \cup p_j)}$$

gdzie p_i i p_j to zbiory pojęć opisujących dokumenty d_i oraz d_j.

Zadanie 12

Zakładając, że opisy obiektów (dokumentów) w systemie S są następujące:

t_{x1} = (TYP, DREWNO)(ROZ, SV)(CZAS, 5)(KOD, C)
t_{x2} = (TYP, METAL)(ROZ, TN)(CZAS, 2)(KOD, B)
t_{x3} = (TYP, METAL)(ROZ, MVA)(CZAS, 5)(KOD, C)
t_{x4} = (TYP, METAL)(ROZ, MVA)(CZAS, 2)(KOD, A)
t_{x5} = (TYP, METAL)(ROZ, TN)(CZAS, 12)(KOD, D)
t_{x6} = (TYP, METAL)(ROZ, SV)(CZAS, 5)(KOD, B)
t_{x7} = (TYP, DREWNO)(ROZ, SV)(CZAS, 2)(KOD, B)
t_{x8} = (TYP, DREWNO)(ROZ, MVA)(CZAS, 12)(KOD, C)
t_{x9} = (TYP, METAL)(ROZ, TN)(CZAS, 5)(KOD, D)
t_{x10} =(TYP, DREWNO)(ROZ, TN)(CZAS, 2)(KOD, D)
Dla przedstawionego zbioru obiektów wyznacz macierz termin-dokument.

Przeprowadź proces wyszukiwania sekwencyjnego, by uzyskać odpowiedź na pytanie: t = (TYP, METAL)(CZAS, 12). Wykorzystaj w tym celu korelację nakładania, oraz dobierz minimalny współczynnik podobieństwa pytania do dokumentu $p_{min} = 0{,}4$.

Przedstaw wady oraz zalety tego sposobu wyszukiwania w kontekście metody Saltona.

Zadanie 13

Zakładając, że opisy obiektów są następujące:

t_{x1} = (TYP, K)(MAT, SV)(CZAS, 5)(KOD, C)
t_{x2} = (TYP, M)(MAT, TN)(CZAS, 2)(KOD, B)
t_{x3} = (TYP, M)(MAT, MVA)(CZAS, 5)(KOD, C)
t_{x4} = (TYP, M)(MAT, MVA)(CZAS, 2)(KOD, A)
t_{x5} = (TYP, M)(MAT, TN)(CZAS, 12)(KOD, D)
t_{x6} = (TYP, M)(MAT, SV)(CZAS, 5)(KOD, B)
t_{x7} = (TYP, K)(MAT, SV)(CZAS, 2)(KOD, B)
t_{x8} = (TYP, M)(MAT, MVA)(CZAS, 12)(KOD, C)
t_{x9} = (TYP, M)(MAT, TN)(CZAS, 5)(KOD, D)
t_{x10} =(TYP, K)(MAT, TN)(CZAS, 2)(KOD, D)
Wykonaj jedną iterację algorytmu Doyle'a przyjmując następujące wartości parametrów: $\alpha = 0.5$, $T = 37$. Algorytm rozpocznij od wstępnego przydziału obiektów do 3 grup: $S_1 = \{x_1, x_2, x_3, x_4\}$

$S_2 = \{x_5, x_6, x_7\}$

$S_3 = \{x_8, x_9, x_{10}\}$

Co możesz powiedzieć o relacji między współczynnikiem skalującym α oraz wartością progową T? Na co te parametry mają wpływ i dlaczego?

Zadanie 14

Dany jest następujący zbiór dokumentów:

D_1: The Light Blue Sky
D_2:The Sky and The Moon
D_3: Sky stuff
D_4: The Moon on the sky
D_5: The Sky is wonderful
D_6: The Moon is like the sky
Oraz zbiór termów: T={*The, Sky, Moon, Blue, stuff, is, wonderful, like}*

Stwórz macierz termin-dokument oraz oblicz podobieństwo miedzy dokumentami D_2 oraz D_4 stosując miarę COSINUS.Wiedząc, że na pytanie o dokumenty zawierające termin „sky" znaleziono w systemie następujące dokumenty: D_1,D_3,D_4 oraz D_6 zbadaj parametry efektywności takiego systemu wyszukiwawczego i zinterpretuj je odpowiednio

Zadanie 15

W systemie zorganizowanym zgodnie z metodą Saltona występują dokumenty o następujących opisach:

d_1: t_1 t_2 t_5
d_2:t_1 t_3 t_5 t_6
d_3:t_1 t_2 t_5 t_3
d_4:t_1 t_2
d_5:t_3 t_4 t_5
d_6:t_4 t_5 t_6,
d_7:t_1 t_5 t_6
d_8:t_6
d_9:t_5 t_6 t_7
d_{10}:t_3 t_5 t_7.

Do systemu zadano pytanie $t=t_5 \cdot t_7+t_1 \cdot t_2$, na które odpowiedź systemu była następująca: $\{d_1,d_9,d_{10},d_5\}$. Określ wartości parametrów efektywności wyszukiwania.

Zadanie 16

W systemie zorganizowanym zgodnie z metodą Saltona występują charakterystyki zasobów księgarni internetowych o następujących opisach:

d_1: informatyka, elektronika, robotyka,
d_2: informatyka, nauki ścisłe, robotyka, bioniżynieria
d_3: informatyka, elektronika, robotyka, nauki ścisłe,
d_4: informatyka, elektronika,
d_5: nauki ścisłe, encyklopedie, robotyka,
d_6: encyklopedie, robotyka, bioinżynieria
d_7: informatyka, robotyka, bioinżynieria
d_8: bioinżynieria
d_9: robotyka, bioinżynieria

d_{10}: nauki ścisłe, robotyka, nauki przyrodnicze.
Do systemu zadano pytanie o książki z dziedzin: robotyki i bioinżynierii lub o książki z elektroniki, na które to pytanie odpowiedź systemu była następująca: $\{d_1,d_2,d_3,d_4,d_6,d_7\}$. Określ wartości parametrów efektywności wyszukiwania.

Zadanie 17

Dla systemu zawierającego opisy płyt głównych dla zestawów komputerowych przedstawionego za pomocą tabeli:

Model	Socket	Chipset	RAM	Szyna	VGA
Asus M4A785D-M Pro	AM2	AMD785G	DDR2	1200	tak
Gigabyte GA-M68M-S2P	AM2	GF7025	DDR2	1066	tak
MSI KA790GX	AM2	AMD790G	DDR2	1066	tak
Asus M4A78LT-M	AM3	AMD760G	DDR3	1800	tak
Gigabyte GA-870A-UD3	AM3	AMD870	DDR3	1866	nie
MSI 770-C45	AM3	AMD770	DDR3	1600	nie
Asus P7H55-M	s1156	Intel H55	DDR3	2200	tak
Gigabyte GA-H55-UD3H	s1156	Intel H55	DDR3	2200	nie
MSI H55-G43	s1156	Intel H55	DDR3	2133	nie
Asus P6T Deluxe V2	s1366	Intel X58	DDR3	2000	nie
Gigabyte GA-X58A-UD3	s1366	Intel X58	DDR3	2200	nie
MSI X58A-GD65	s1366	Intel X58	DDR3	2133	nie

przeprowadź jedną iterację grupowania płyt głównych dla zestawów komputerowych algorytmem Doyle'a dla w/w systemu. Utwórz 2 grupy (rozdzielając płyty z obsługą standardu graficznego VGA od tych, które takiego wyposażenia nie posiadają). Jako współczynnik skalujący przybierz wartość optymalną, zaś wartość parametru T wyznacz za pomocą formuły: T = 1/2(min(g(di.Pj)) +max(g(di.Pj))), gdzie g(di.Pj) określa wartość funkcji punktującej dokumentu di w profilu Pj.