SPEKTRALNO PARTICIONIRANJE BIPARTITNOG GRAFA

Seminarski rad

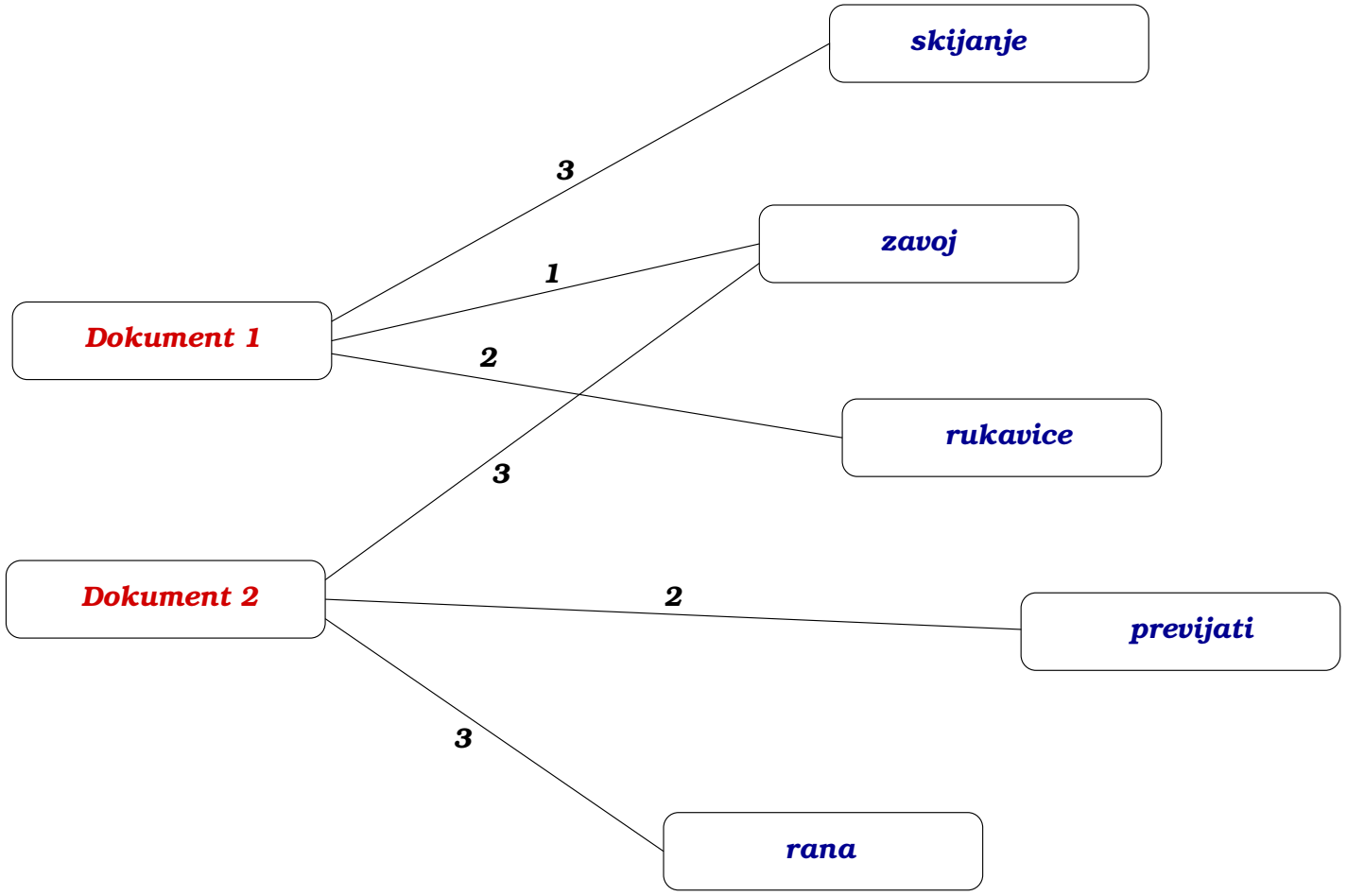
Ivančica Mirošević

Fakultet elektrotehnike, strojarstva i brodogradnje
Sveučilište u Splitu

Bipartitni graf

Neusmjereni bipartitni graf G je uređena trojka (R, D, B), gdje su $R = \{r_1, \cdots, r_m\}$ i $D = \{d_1, ..., d_n\}$ dva skupa vrhova i B skup bridova $\{\{r_i, d_j\} : r_i \in R, d_j \in D\}$. U kontekstu razvrstavanja dokumenata i riječi D je skup dokumenata, a R je skup riječi koje oni sadrže. Brid $\{r_i, d_j\}$ postoji ako se riječ r_i javlja u dokumentu d_j.

Primjer 1


skijanje
zavoj
rukavice
previjati
rana
Dokument 1
Dokument 2
3
1
2
3
2
3

Primjer 1

$$R = \{r_1, \cdots, r_5\} = \{skijanje, zavoj, rukavice, previjati, rana\}$$
$$D = \{d_1, d_2\} = \{dokument1, dokument2\}$$

$$W = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 3 & 0 & 2 & 3 & 0 & 0 \\ 3 & 1 & 2 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Matrica susjedstva

$$W = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix},$$

gdje je matrica $A \in \mathbb{R}^{m \times n}$ riječi $\times$ dokumenti matrica u kojoj su dokumenti prikazani kao vektori u prostoru bitnih riječi (svojstava).

Dualnost razvrstavanja

Promatrat ćemo k-particiju π vrhova grafa

$$\begin{aligned} R &= R_1 \cup R_2 \cup \cdots \cup R_k, \\ D &= D_1 \cup D_2 \cup \cdots \cup D_k. \end{aligned}$$

Očigledno vrijedi

$$R_m = \{r_i : \sum_{d_j \in D_m} a_{ij} \geq \sum_{d_j \in D_l} a_{ij}, \forall l = 1, ..., k\}, \text{ i}$$

$$D_m = \{d_i : \sum_{r_j \in R_m} a_{ij} \geq \sum_{r_j \in R_l} a_{ij}, \forall l = 1, ..., k\}.$$

Veza s SVD-om

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}, \quad L = \begin{bmatrix} D_1 & -A \\ -A^T & D_2 \end{bmatrix}$$

Vrijedi

$$L_n = D^{-\frac{1}{2}} \begin{bmatrix} D_1 & -A \\ -A^T & D_2 \end{bmatrix} D^{-\frac{1}{2}} = \begin{bmatrix} I & -D_1^{-\frac{1}{2}} A D_2^{-\frac{1}{2}} \\ -D_2^{-\frac{1}{2}} A^T D_1^{-\frac{1}{2}} & I \end{bmatrix}$$

Veza s SVD-om

Neka je

$$\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \quad \mathbf{u} \in \mathbb{R}^m, \quad \mathbf{v} \in \mathbb{R}^n,$$

svojstveni vektor normaliziranog Laplacijana.

Iz $D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \mathbf{w} = \lambda \mathbf{w}$ dobije se

$$\begin{aligned} D_1^{-\frac{1}{2}} A D_2^{-\frac{1}{2}} \mathbf{v} &= (1-\lambda)\mathbf{u}, \\ D_2^{-\frac{1}{2}} A^T D_1^{-\frac{1}{2}} \mathbf{u} &= (1-\lambda)\mathbf{v}. \end{aligned}$$

Veza s SVD-om

Umjesto računanja svojstvenih vektora normaliziranog Laplacijana koji odgovaraju najmanjim svojstvenim vrijednostima, možemo računati lijeve i desne singularne vektore koji odgovaraju najvećim singularnim vrijednostima matrice A_n,

$$\begin{aligned} A_n \mathbf{v}^{[i]} &= \sigma_i \mathbf{u}^{[i]}, \\ A_n^T \mathbf{u}^{[i]} &= \sigma_i \mathbf{v}^{[i]}, \end{aligned}$$

gdje je $\sigma_i = 1 - \lambda_i$

Primjer 1

Lijevi i desni normalizirani singularni vektori matrice susjedstva grafa koji odgovaraju drugoj najvećoj singularnoj vrijednosti su

$$D_1^{-\frac{1}{2}}\mathbf{u}^{[2]} = \begin{bmatrix} 0.45 & -5.6 \cdot 10^{-17} & 0.37 & -0.12 & -0.15 \end{bmatrix}^T$$

i

$$D_2^{-\frac{1}{2}}\mathbf{v}^{[2]} = \begin{bmatrix} -0.11 & 0.38 \end{bmatrix}^T.$$

Njihovim komponentama očigledno je određena particija

$$V = C_1 \cup C_2, \qquad C_1 = \{zavoj, previjati, rana, dokument1\},$$
$$C_2 = \{skijanje, rukavice, dokument2\}.$$

Biparticijski algoritam

1. Za zadanu matricu A izračunaj $A_n = D_1^{-\frac{1}{2}} A D_2^{-\frac{1}{2}}$;
2. Izračunaj singularne vektore matrice A_n, $\mathbf{u}^{[2]}$ i $\mathbf{v}^{[2]}$; formiraj vektor

$$\mathbf{z}^{[2]} = \begin{bmatrix} D_1^{-\frac{1}{2}} \mathbf{u}^{[2]} \\ D_2^{-\frac{1}{2}} \mathbf{v}^{[2]} \end{bmatrix};$$

3. Komponente vektora $\mathbf{z}^{[2]}$ podijeli u dvije skupine C_1 i C_2 prema

$$C_1 = \{i : \mathbf{z}_i^{[2]} < 0\}, \quad C_2 = \{i : \mathbf{z}_i^{[2]} \geq 0\}.$$

Rekurzivni biparticijski algoritam

1. Biparticijskim algoritmom odredi optimalnu biparticiju skupa V; Postavi brojač $k_{pom} = 2$;
2. Provjeri je li $kpom = k$. Ako nije,
   - za svaku podskupinu skupa V odredi optimalnu biparticiju;
   - među dobivenim $(k_{pom} + 1)$-particijama odaberi onu s najmanjom vrijednošću ciljne funkcije;
   - postavi $k_{pom} = k_{pom} + 1$ i ponovi korak 2.
3. Stani.

Multiparticijski algoritam

1. Za zadanu matricu A izračunaj $A_n = D_1^{-\frac{1}{2}} A D_2^{-\frac{1}{2}}$;
2. Izračunaj k singularnih vektora matrice A_n, $\mathbf{u}^{[1]}, ..., \mathbf{u}^{[k]}$ i $\mathbf{v}^{[1]}, ..., \mathbf{v}^{[k]}$, i formiraj matricu

$$Z = \begin{bmatrix} D_1^{-\frac{1}{2}} U \\ D_2^{-\frac{1}{2}} V \end{bmatrix},$$

gdje je

$$U = [\mathbf{u}^{[1]}, ..., \mathbf{u}^{[k]}] \quad \text{i} \quad V = [\mathbf{v}^{[1]}, ..., \mathbf{v}^{[k]}].$$

3. Pokreni algoritam k-means nad retcima matrice Z.

Multiparticijski algoritam

K-means algoritam će na izlazu dati središta $\mathbf{c}_1, ..., \mathbf{c}_k$ skupina, te vektor $[sk_1, sk_2, \cdots, sk_{m+n}]$ gdje $sk_i \in \{1, 2, \cdots, k\}$, $i = 1, \cdots, m+n$, označava redni broj skupine kojoj $Z(i)$ pripada. To znači da je skupina i-te riječi, za $i = 1, \cdots, m$, dana s sk_i, i da je skupina j-tog dokumenta, $j = 1, \cdots, n$, dana s sk_{j+m}.

Primjer 2

Na stranici

ftp://ftp.cs.cornell.edu/pub/smart

nalazi se kolekcija od 1033 sažetka članaka iz područja medicine (MEDLINE), 1400 sažetaka članaka o zrakoplovnim sustavima (CRANFIELD) te 1460 sažetaka članaka o ekstrakciji znanja (CISI).

Na stranici

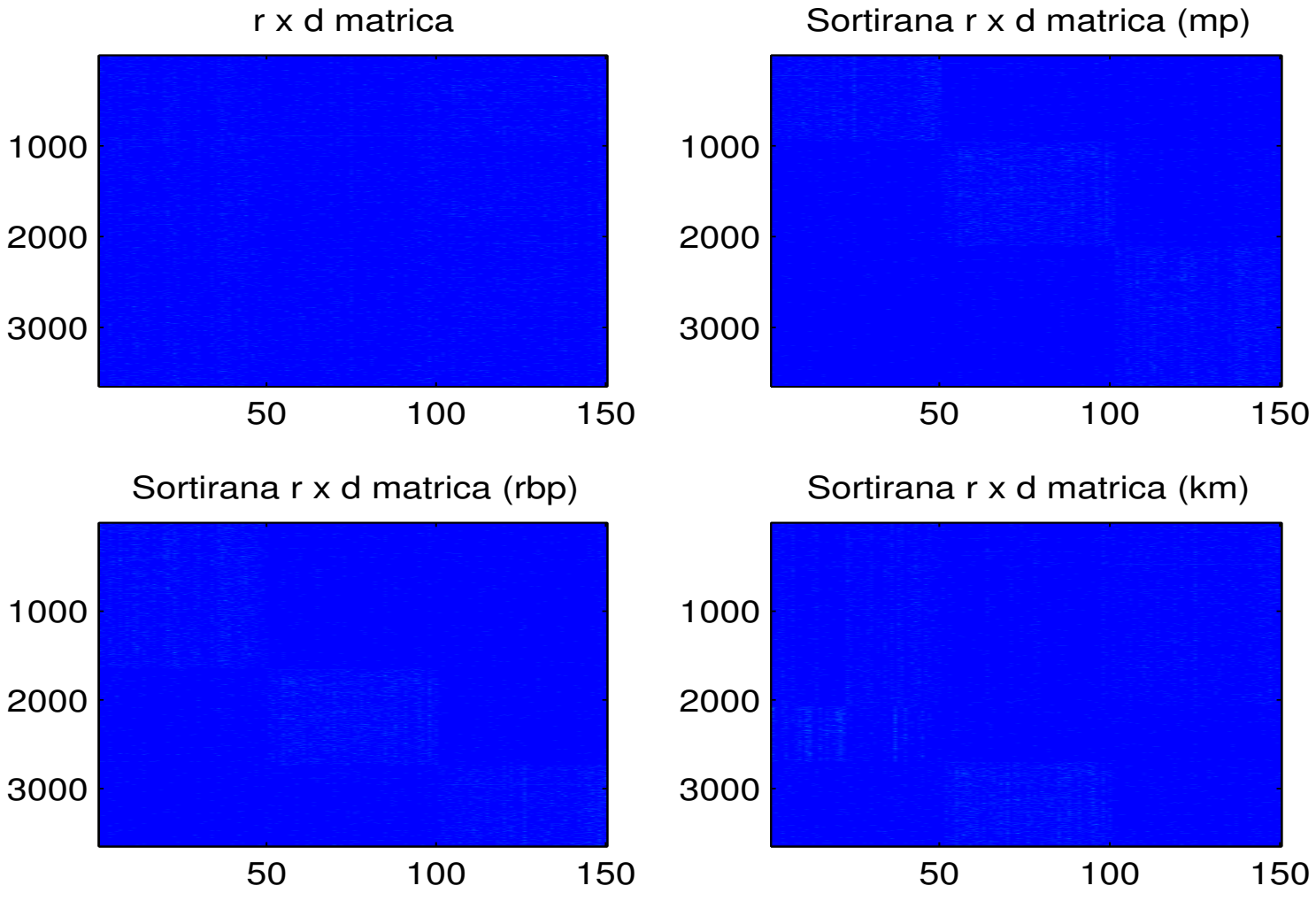
http://www.cs.utexas.edu/ftp/pub/inderjit/Data/Text/small_data_set

dane su matrice susjedstva za tri testna skupa od 30, 150 i 300 dokumenata, generirana slučajnim odabirom iz kolekcije sažetaka.

Primjer 2

testni podaci	algoritam	broj krivo smještenih		trajanje (s)
		riječi	dokumenata	
	mp	92	2	0.078
30 dokumenata	rbp	89	2	0.172
(1073×30)	km	283	14	0.063
	mp	163	1	0.375
150 dokumenata	rbp	147	1	0.672
(3658×150)	km	396	51	1.172
	mp	399	5	0.67
300 dokumenata	rbp	372	6	1.906
(5577×300)	km	715	67	5.03

Primjer 2


r x d matrica
1000
2000
3000
50
100
150
Sortirana r x d matrica (mp)
1000
2000
3000
50
100
150
Sortirana r x d matrica (rbp)
1000
2000
3000
50
100
150
Sortirana r x d matrica (km)
1000
2000
3000
50
100
150

Primjer 3

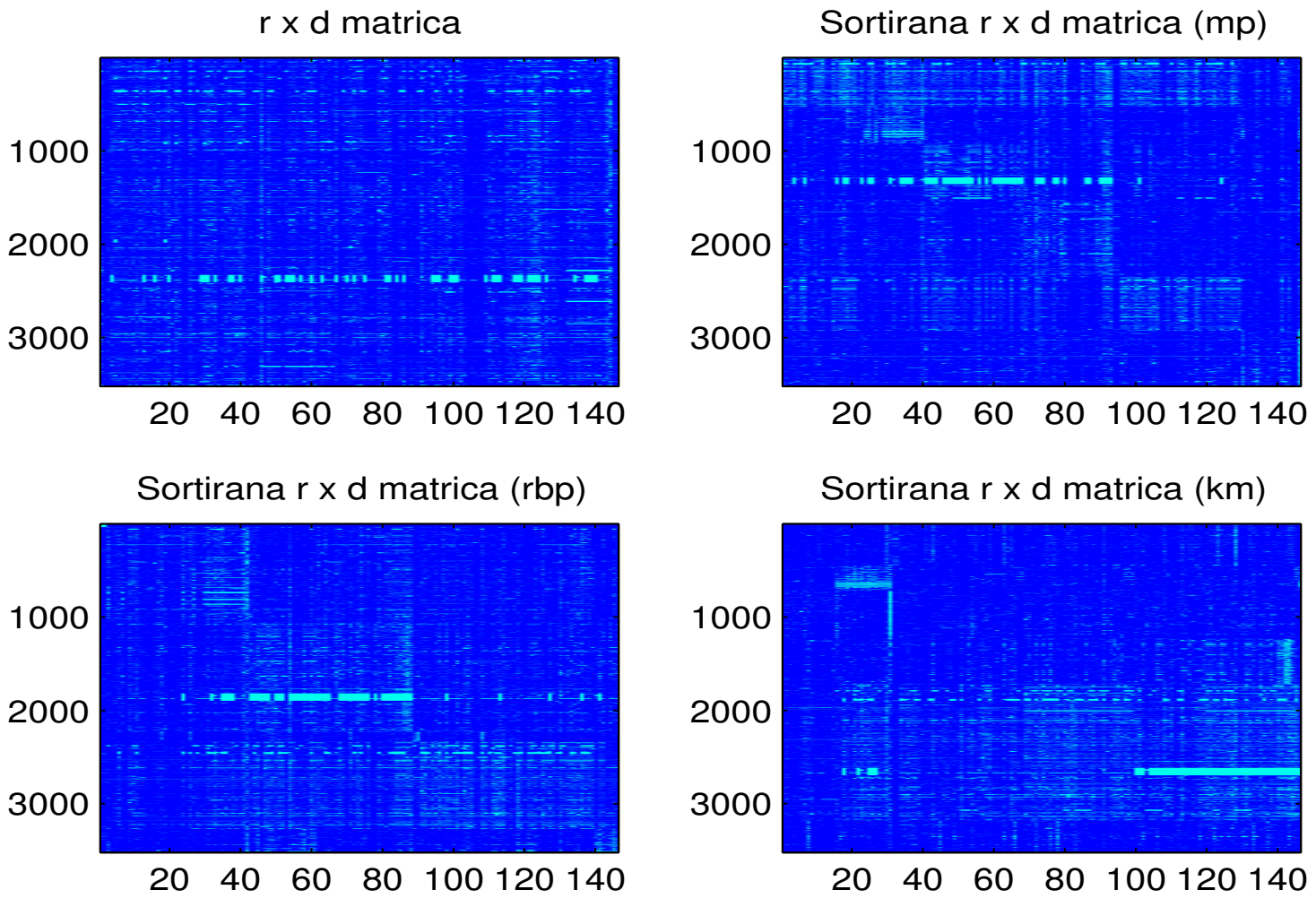
Elektronički udžbenik *Matematika 1* (http://lavica.fesb.hr/mat1) prof. Ivana Slapničara sastoji se 146 dokumenata podijeljenih u šest poglavlja:

Osnove matematike, Linearna algebra, Vektorska algebra i analitička geometrija, Funkcije realne varijable, Derivacije i primjene, Nizovi i redovi.

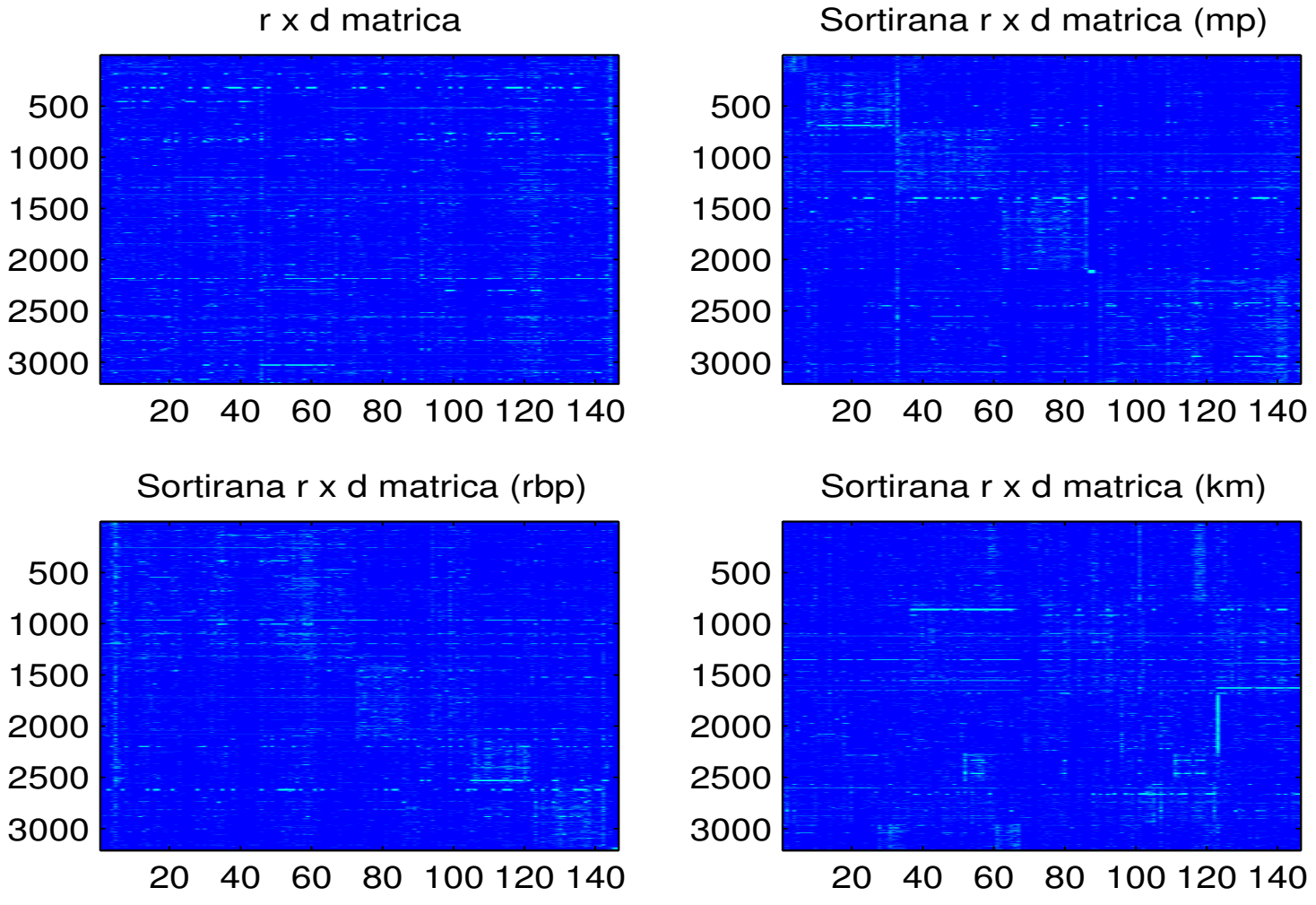
Rezultati spektralnog particioniranja dokumenata i riječi:

testni podaci	algoritam	norm. rez	trajanje (s)
(a)	mp	3.4254	0.672
(> 2 slova u riječi)	rbp	3.582	3.75
3522×146	km	5.3561	1.766
(b)	mp	3.0196	0.484
(> 4 slova u riječi)	rbp	3.115	3.875
3213×146	km	5.1435	2.797

Primjer 3a


r x d matrica
Sortirana r x d matrica (mp)
Sortirana r x d matrica (rbp)
Sortirana r x d matrica (km)
1000
2000
3000
20 40 60 80 100 120 140

Primjer 3b


r x d matrica
Sortirana r x d matrica (mp)
Sortirana r x d matrica (rbp)
Sortirana r x d matrica (km)
500
1000
1500
2000
2500
3000
20 40 60 80 100 120 140

Primjer 3b

Rezultat particioniranja multiparticijskim algoritmom:

Osnove matematike (21) - $[2\,6\,1\,3\,3\,3\,3\,4\,1\,1\,1\,1\,3\,1\,3\,3\,1\,3\,2\,2\,3]^T$

Linearna algebra (25) - $[4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,2\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4]^T$

Vektorska algebra i analitička geometrija (20) - $[2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2]^T$

Funkcije realne varijable (31) - $[6\,6\,2\,6\,6\,6\,3\,3\,3\,6\,6\,6\,6\,3\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6]^T$

Derivacije i primjene (27) - $[6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6\,6]^T$

Nizovi i redovi (21) - $[6\,3\,3\,3\,3\,3\,3\,6\,3\,3\,3\,3\,3\,3\,3\,3\,3\,6\,6\,2\,5\,5]^T$

Primjer 3b

Rezultat particioniranja rekurzivnim biparticijskim algoritmom:

Osnove matematike (21) - $[2\,1\,1\,3\,3\,3\,3\,5\,3\,3\,3\,3\,3\,3\,3\,3\,2\,3\,2\,2]^T$

Linearna algebra (25) - $[5\,3\,5\,3\,5\,5\,5\,3\,5\,5\,5\,5\,5\,5\,4\,5\,5\,5\,5\,3\,5\,5\,5\,5]^T$

Vektorska algebra i analitička geometrija (20) - $[4\,4\,4\,3\,3\,3\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,4\,3\,]^T$

Funkcije realne varijable (31) - $[2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,3\,2\,2\,2\,2\,2\,3\,2\,3\,3\,2\,3\,2\,3\,2\,2\,2\,2\,2]^T$

Derivacije i primjene (27) - $[2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,2\,6]^T$

Nizovi i redovi (21) - $[2\,2\,3\,3\,2\,2\,3\,2\,1\,3\,2\,2\,2\,2\,2\,2\,2\,1\,1\,6\,6]^T$