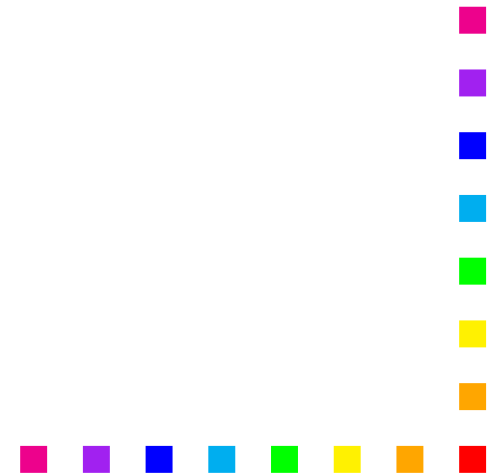


Prva hrvatska web tražilica

www.csc.unist.hr/trazilica

Festival znanosti 2006.

Ivan Slapničar
FESB Split



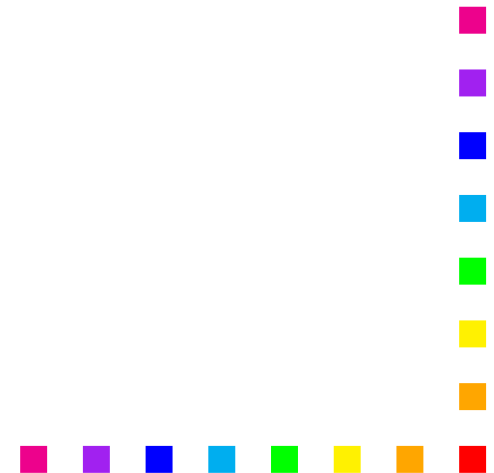
Projekti

- STIRP projekt CRO-GRID, 2004.-2007.
 - CRO-GRID Infrastruktura
 - CRO-GRID Posrednik
 - CRO-GRID Aplikacije
 - Tema Indeksiranje web/lokalnih mreža - Data Mining
- Točni i mrzi matrični algoritmi i primjene (MZOŠ)



Projektni tim

- prof. dr. sc. Ivan Slapničar (voditelj)
- Branko Dodig, inž.
- Nevena Jakovčević Stor, prof.
- mr. sc. Damir Krstinić
- mr. sc. Ivančica Mirošević



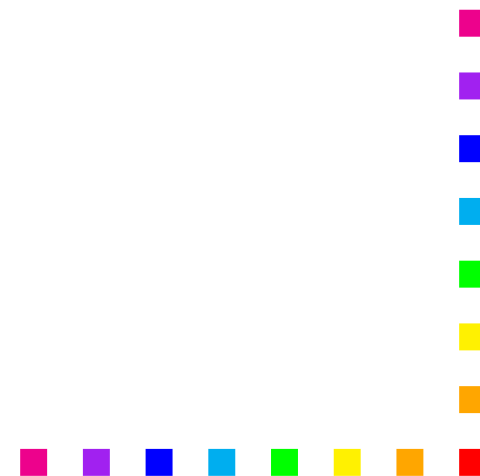
Suradnja

- Sveučilište u Splitu, Fakultete elektrotehnike, strojarstva i brodogradnje – **nositelj**
(www.fesb.hr)
- Institut *Ruđer Bošković* (www.irb.hr) – nositelj podprojekta *Aplikacije*
(www.cro-grid.hr)
- Sveučilišni računski centar (www.srce.hr) – nositelj podprojekta *Infrastruktura*
(www.srce.hr/crogrid/infrastruktura)
- Sveučilište u Splitu, Centar za znanstveno računanje (www.csc.unist.hr)
- Sveučilite u Zagrebu, Filozofski fakultet, Odsjek za lingvistiku – gramatika hrvatskog jezika (www.hnk.ffzg.hr)



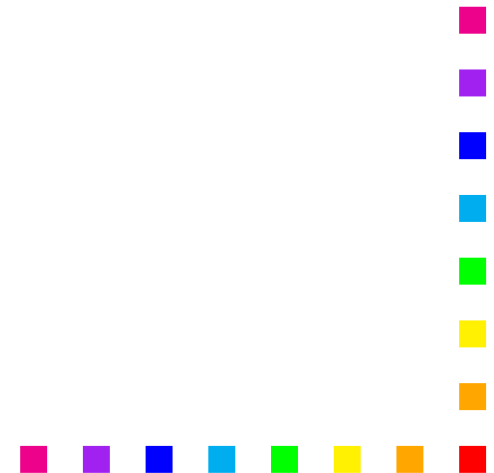
Podrška za hrvatski jezik

- 38.000 riječi → 300.000+ različenica
- 200 pravila za promjene imenica, glagola i pridjeva
- → [Novi hrvatski spelling checker](#)

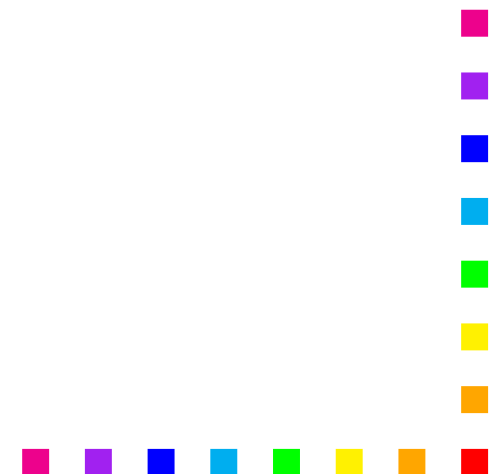
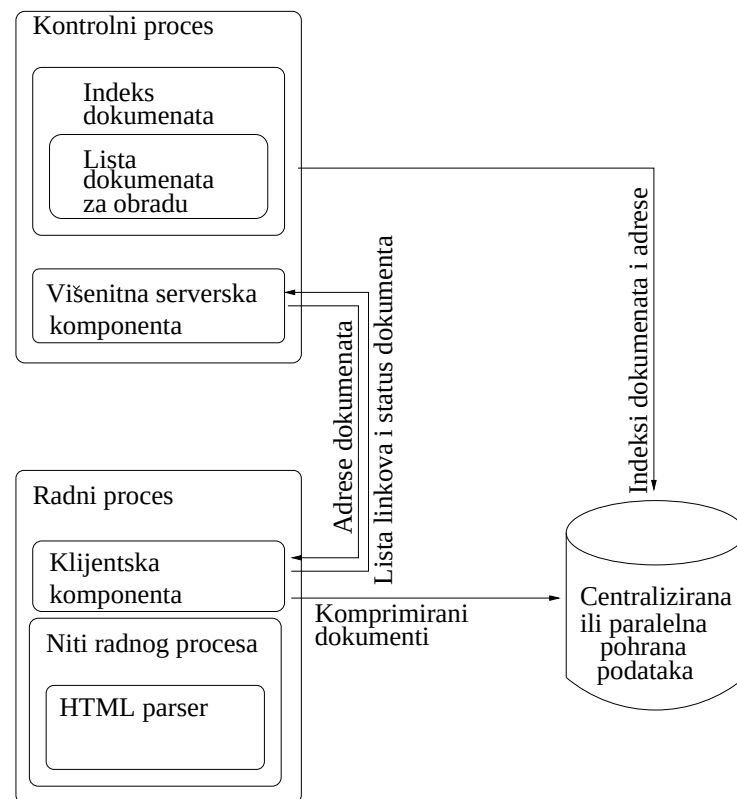


Spider

- Prikupljaju se (za sada) .html, .doc, .pdf, .ps, .xls, .txt i .ppt datoteke.
- Tekstualne datoteke se konvertiraju u *html* oblik.
- Pohranjuju se *html* veze radi rangiranja upita.

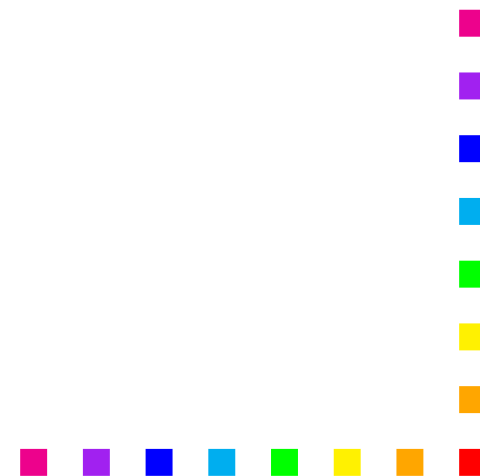


Arhitektura spidera



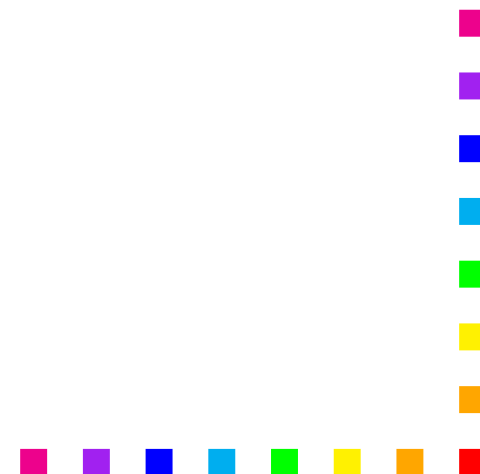
Spider na gridu

- Kontrolni proces se pokreće na bilo kojem računalu.
- Radni procesi se pokreću na računalima unutar CRO-GRID-a i spajaju se na kontrolni proces.
- Radni procesi mogu biti bilo gdje na gridu (uz Unix - Linux i softver).



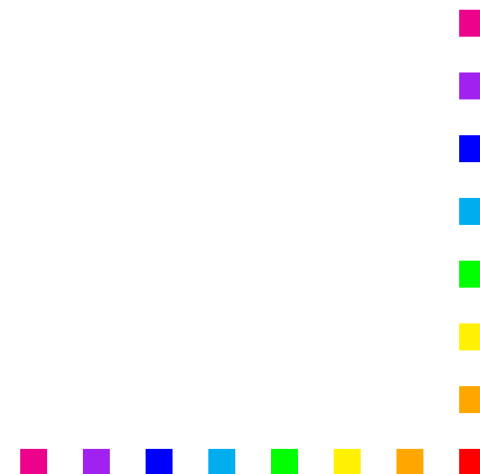
Hardver

- Klaster *grozd.fesb.cro-grid.hr*: 3 x 2 Intel Xeon, 3 x 2 Gb RAM, 1 Gb/s Ethernet.
- Slični klasteri u Rijeci i Zagrebu
- Operacijski sustav: *Linux* - SuSE i RedHat
- Klasterski sustavi: PBS i SGE
- Grid sustav: Condor i Globus



Rezultati

- 11 računala s 22 procesora i 22 Gb RAM
- 22 procesa
- 5 niti po procesu
- 36 sati rada
- 3.500.000 dokumenata
- 64 Gb podataka



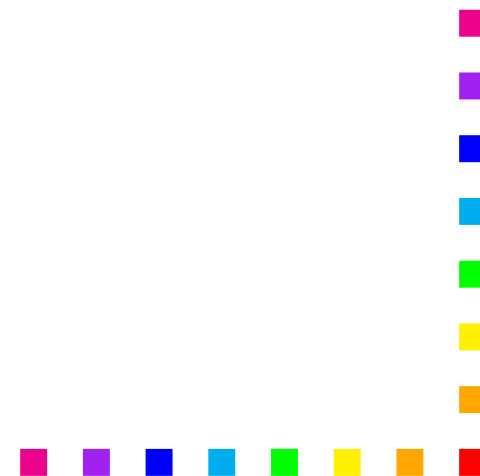
Plan daljnjeg rada

- Indeksiranje svih vrsta datoteka (slike, glazba, video)
- Bolje rangiranje odgovora
- Pretraživanje u vektorskom prostoru (LSI)
- Ažuriranje matričnih rastava
- Klasteriranje podataka - ekstrakcija znanja (Data Mining)
- Pretraživanje slika
- VELIKE matrice



Nevena Jakovčević Stor

Primjena ažuriranja SVD
rastava na tekstualne
pretraživače



Model vektorskog prostora

Reprezentacija informacija vektorskim prostorom
(VSM model)

- pojmovi i dokumenti prikazuju se kao vektori
- vrijednost pridružena komponenti vektora reflektira važnost pojma u dokumentu.

(standardno je funkcija frekvencije pojavljivanja pojma u dokumentu ili skupu dokumenata)

- baza podataka od d dokumenata opisanih s t pojmova reprezentira se sa $t \times d$ pojam-dokument matricom A .



Model vektorskog prostora

Matrica A je rijetka (svaka se riječ ne pojavljuje u svakom dokumentu).

Težina se elementima obično pridaje na slijedeći način:

$$A = [f_{ij} \times L(i, j) \times G(i)]$$

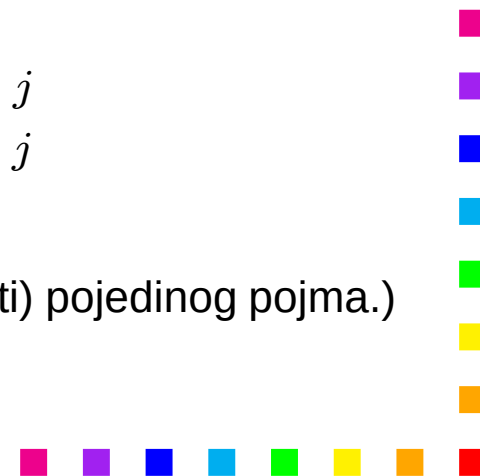
gdje je:

f_{ij} frekvencija pojavljivanja pojma i u dokumentu j

$L(i, j)$ lokalna težina poma i u dokumentu j

$G(i)$ globalna težina pojma i

(lokalna i globalna težina služe povećanju/smanjenju težine (važnosti) pojedinog pojma.)



LSI

(Latentno Semantičko Indeksiranje)

LSI je profinjenje standardnog VSM modela korištenjem RDR -a (reduced dimension representation).

RDR je zadana najboljom aproksimacijom ranga k (A_k) pojam dokument matrice $A \in R^{m \times n}$ i može se izvesti iz osnovnog SVD rastava koristeći k najvećih svojstvenih trojki.

$$A_k = U_k \Sigma_k V_k^T$$

gdje su nam elementi dijagonalne matrice $\Sigma_k \in R^{k \times k}$ - k najvećih singularnih vrijednosti originalne matrice A , a stupci matrica $U_k \in R^{m \times k}$ i $V_k \in R^{n \times k}$ pripadni singularni vektori.

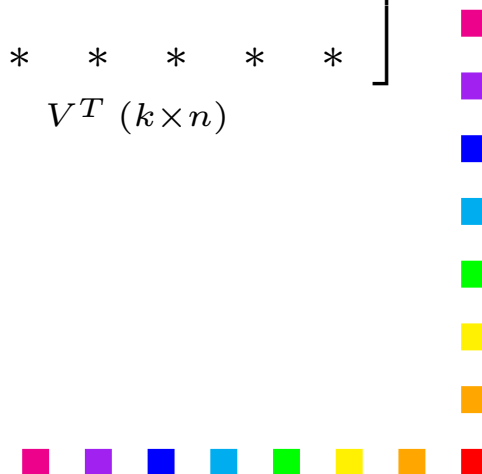


SVD reducirani rastav

$m > n : (m = 8, n = 6, k = 3)$

$$\begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \begin{bmatrix} * & & \\ & * & \\ & & * \end{bmatrix} \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix}$$

A_k
 $U (m \times k)$
 $\Sigma (k \times k)$
 $V^T (k \times n)$



Ažuriranje pojam dokument matrice

Zadanu pojam-dokument matricu $A \in R^{m \times n}$ nužno je ažurirati nakon što smo je na određeni način promijenili. Te promjene mogu biti:

- dodavanje stupca (ili retka)
- brisanje stupca (ili retka)
- promjene vrijednosti pojedinih elemenata



Dodavanje stupaca (1)

Pretpostavimo da novih s stupaca dodajemo u postojeću matricu, tj. dodajemo novu $m \times s$ matricu D , na zadani SVD rastav matrice A_k . Neka je:

$A \in R^{m \times n}$ originalna pojam dokument matrica

$A_k = U_k \Sigma_k V_k^T$ najbolja aproksimacija matrice A ranga k

D matrica novih stupaca, $D \in R^{m \times s}$

$B = \begin{pmatrix} A_k & D \end{pmatrix}$ matrica $m \times (n + s)$

Rang matrice B je po konstrukciji veći ili jednak rangu k matrice A_k .

Naša aproksimacija SVD rastava od B daje aproksimaciju ranga k SVD rastava matrice B .



Dodavanje stupaca (2)

Cilj nam je izračunati komponente SVD rastava

$$B = U_B \Sigma_B V_B^T.$$

Formiramo QR rastav matrice $(I - U_k U_k^T) D$

$$(I - U_k U_k^T) D = \hat{U}_k R$$

$(\hat{U}_k (m \times s))$ je ortonormalna, a R je $(s \times s)$ gornje trokutasta

Vrijedi:

$$B = \begin{pmatrix} A_k & D \end{pmatrix} = \begin{pmatrix} U_k & \hat{U}_k \end{pmatrix} \begin{pmatrix} \Sigma_k & U_k^T D \\ 0 & R \end{pmatrix} \begin{pmatrix} V_k^T & 0 \\ 0 & I_s \end{pmatrix}$$

Dodavanje stupaca (3)

Sada računamo SVD rastav od

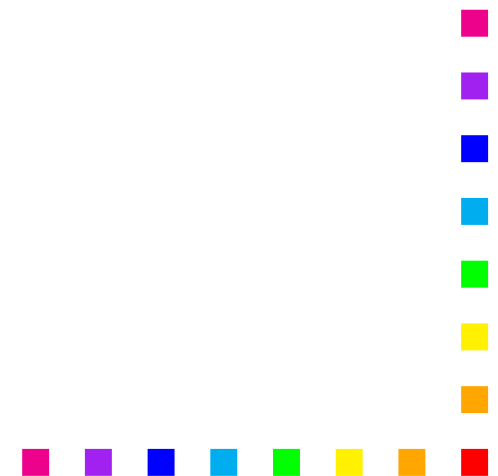
$$\overline{B} = \begin{pmatrix} \Sigma_k & U_k^T D \\ 0 & R \end{pmatrix} = \left(\overline{U}_k, \overline{U}_k^\perp \right) \begin{pmatrix} \overline{\Sigma}_k & 0 \\ 0 & \overline{\Sigma}_s \end{pmatrix} \left(\overline{V}_k, \overline{V}_k^\perp \right)^T$$

gdje su:

$\overline{U}_k$ $(k + s) \times k$ matrica ortonormalnih stupaca

$\overline{V}_k$ je $(k + s) \times k$ matrica ortonormalnih stupaca

$\overline{\Sigma}_k$ je $k \times k$ dijagonalna matrica.



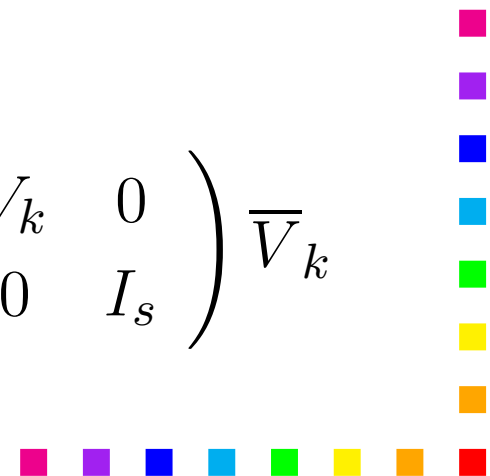
Dodavanje stupaca (4)

Najbolja aproksimacija od B ranga k sada je:

$$B_k = \left(\begin{pmatrix} U_k & \hat{U}_k \end{pmatrix} \bar{U}_k \right) \bar{\Sigma}_k \left(\begin{pmatrix} V_k & 0 \\ 0 & I_s \end{pmatrix} \bar{V}_k \right)^T$$

pa su komponente SVD rastava matrice B

$$U_B = \begin{pmatrix} U_k & \hat{U}_k \end{pmatrix} \bar{U}_k, \quad \Sigma_B = \bar{\Sigma}_k, \quad V_B = \begin{pmatrix} V_k & 0 \\ 0 & I_s \end{pmatrix} \bar{V}_k$$



Algoritam (1)

1. ULAZNI PODACI:

$$U_k \in R^{m \times k}, \Sigma_k \in R^{k \times k}, V_k \in R^{n \times k}, D \in R^{m \times s}.$$

2. Izračunamo projekciju: $\hat{D} = (I_m - U_k U_k^T) D.$

3. Izračunamo QR rastav: $\hat{D} = (I_m - U_k U_k^T) D = \hat{U}_k R,$

gdje je $\hat{U}_k \in R^{m \times s}, R \in R^{s \times s}.$

4. Izračunamo SVD rastav matrice

$$\overline{B} \equiv \begin{pmatrix} \Sigma_k & U_k^T D \\ 0 & R \end{pmatrix} \in \mathbb{R}^{(k+s) \times (k+s)}$$

Algoritam (2)

i napišemo ga u formi:

$$\overline{B} \equiv \left(\overline{U}_k, \overline{U}_k^\perp \right) \text{diag} \left(\overline{\Sigma}_k, \overline{\Sigma}_s \right) \left(\overline{V}_k, \overline{V}_k^\perp \right)^T$$

gdje su $\overline{U}_k, \overline{V}_k \in R^{(k+s) \times k}$ i $\overline{\Sigma}_k \in R^{k \times k}$.

5. IZLAZNI PODACI: Najbolja aproksimacija ranga k od $B = \left(A_k, D \right)$ je dana sa:

$$B_k = \left[\left(U_k, \hat{U}_k \right) \overline{U}_k \right] \overline{\Sigma}_k \left[\begin{pmatrix} V_k & 0 \\ 0 & I_s \end{pmatrix} \overline{V}_k \right]^T$$

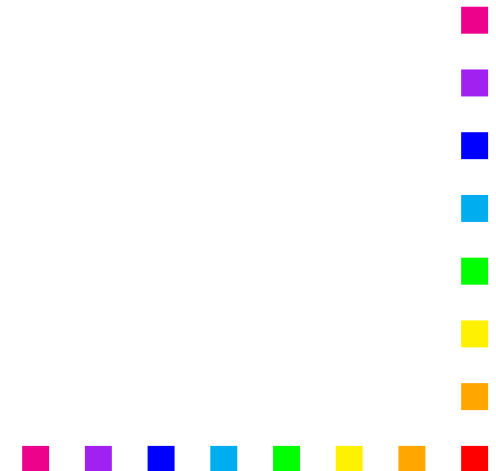
Primjer (1)

TABLICA1: ORIGINALNI DOKUMENTI

<i>D1 :</i>	<i>Zdravlje životinja</i>
<i>D2 :</i>	<i>Domaće riječi</i>
<i>D3 :</i>	<i>Hrvatski domaći kolači</i>
<i>D4 :</i>	<i>Kućne Mačke</i>

TABLICA 2: POJMOVI

<i>P1 :</i>	zdravlje
<i>P2 :</i>	domaće
<i>P3 :</i>	mačka
<i>P4 :</i>	pas
<i>P5 :</i>	životinje
<i>P6 :</i>	kućni

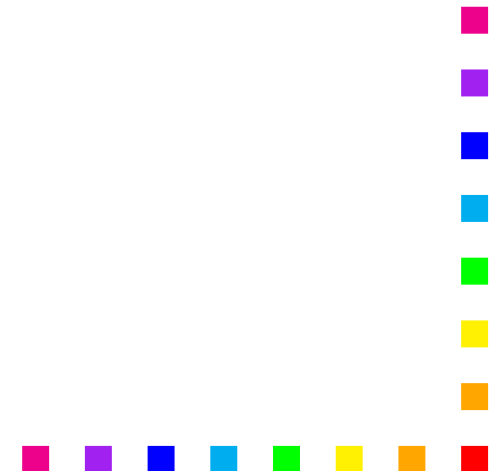


Primjer (2)

Tada je matrica A , pripadna pojam dokument matrica:

POJAM DOKUMENT MATRICA

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

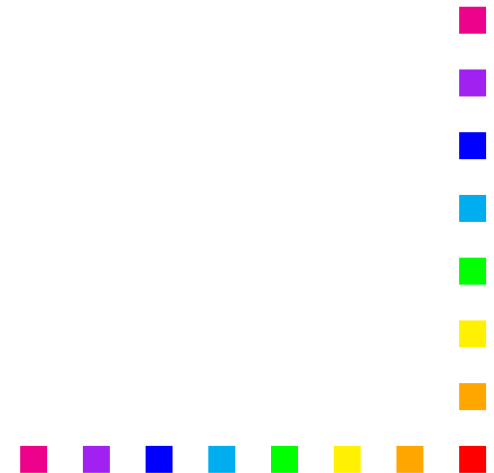


Primjer (3)

Reducirani SVD rastav matrice A (za $k = 2$) je sljedeći:

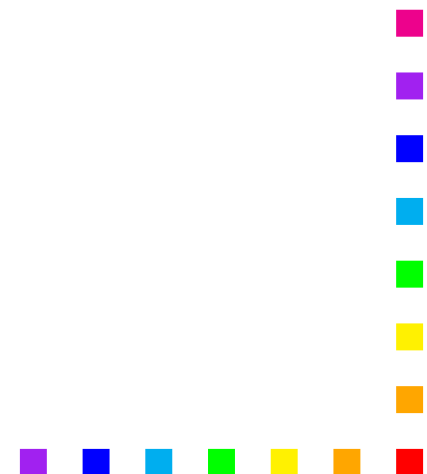
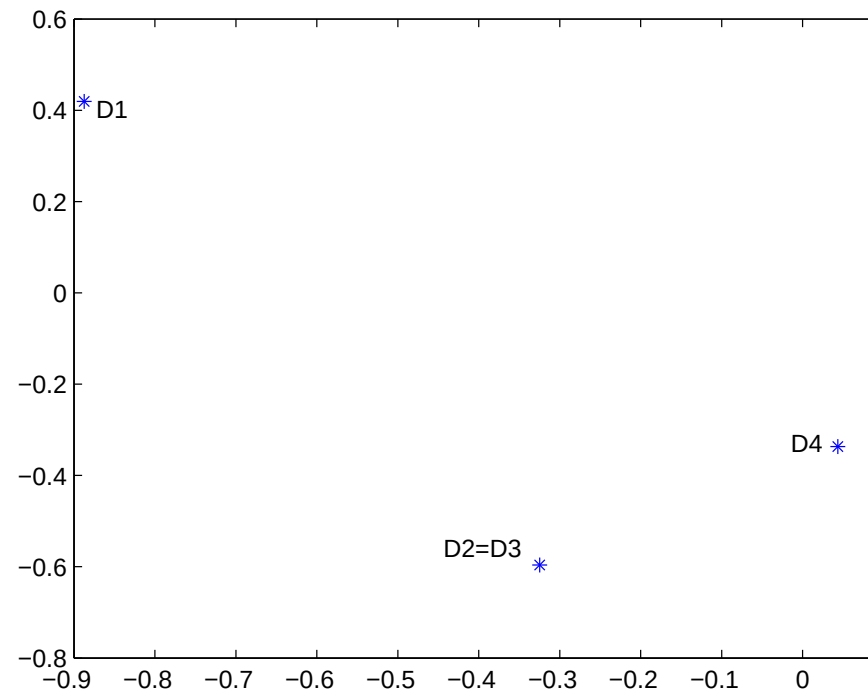
$$U_k = \begin{bmatrix} 0.6669 & -0.1944 \\ -0.2729 & -0.3180 \\ -0.1345 & -0.6416 \\ 0.0000 & -0.0000 \\ 0.6669 & -0.1944 \\ -0.1345 & -0.6416 \end{bmatrix}, \quad S_k = \begin{bmatrix} 1.4142 & 0 \\ 0 & 1.4142 \end{bmatrix},$$

$$V_k = \begin{bmatrix} 0.9431 & -0.2750 \\ -0.1929 & -0.2249 \\ -0.1929 & -0.2249 \\ -0.1901 & -0.9073 \end{bmatrix}.$$



Primjer (4)

Prikaz tih dokumenata u ravnini je (prvi stupac matrice V_k crtamo na x osi, a drugi na y osi):



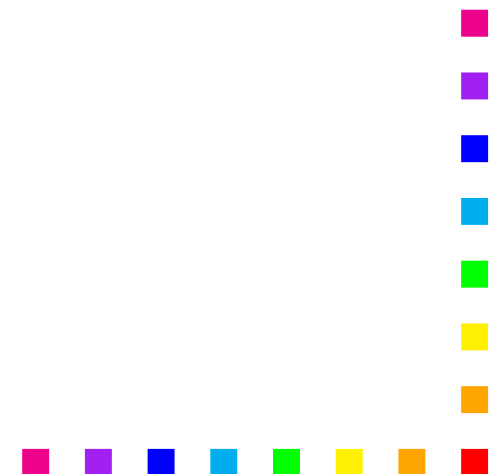
Primjer (5)

Dodajmo četiri nova dokumenta:

$D5 :$	<i>Bilje za zdravlje</i>
$D6 :$	<i>Sve o zdravlju kućnih ljubimaca</i>
$D7 :$	<i>Psi i mačke</i>
$D8 :$	<i>Mačka među golubovima</i>

matrica dodanih dokumenata D je:

$$D = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

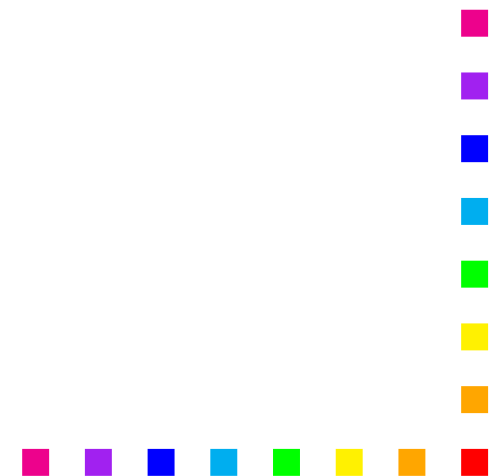


Primjer (6)

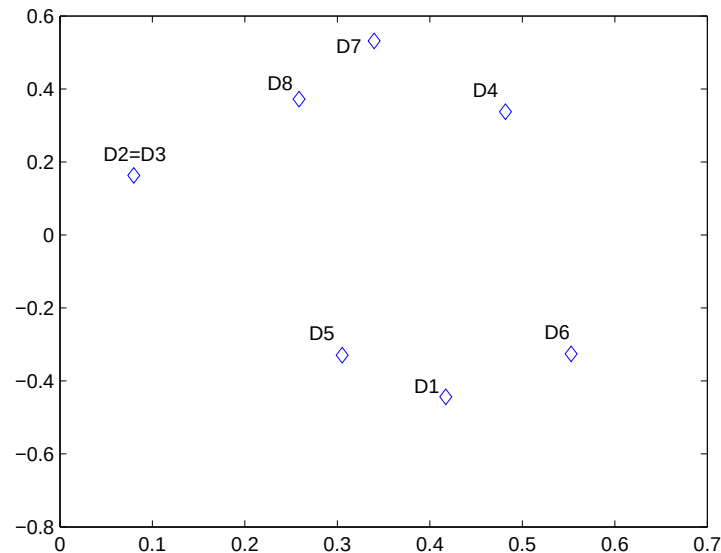
Nakon dodavanja matrice D algoritmom 1 dobivamo reducirani SVD rastav za ($k = 2$) matrice $B = \begin{pmatrix} A_k, D \end{pmatrix}$

$$U_B = \begin{bmatrix} 0.6238 & -0.6017 \\ 0.0782 & 0.1787 \\ 0.5285 & 0.6799 \\ 0.1663 & 0.2912 \\ 0.2041 & -0.2428 \\ 0.5061 & 0.0064 \end{bmatrix}, \quad S_B = \begin{bmatrix} 2.0441 & 0 \\ 0 & 1.8262 \end{bmatrix},$$

$$V_B = \begin{bmatrix} 0.4172 & -0.4434 \\ 0.0799 & 0.1632 \\ 0.0799 & 0.1632 \\ 0.4818 & 0.3376 \\ 0.3052 & -0.3295 \\ 0.5528 & -0.3260 \end{bmatrix}.$$



Primjer (7)



<i>D1 :</i>	<i>Zdravlje životinja</i>
<i>D2 :</i>	<i>Domaće riječi</i>
<i>D3 :</i>	<i>Hrvatski domaći kolači</i>
<i>D4 :</i>	<i>Kućne Mačke</i>
<i>D5 :</i>	<i>Bilje za zdravlje</i>
<i>D6 :</i>	<i>Sve o zdravlju kućnih ljubimaca</i>
<i>D7 :</i>	<i>Psi i mačke</i>
<i>D8 :</i>	<i>Mačka među golubovima</i>

