Uvod u složeno pretraživanje podataka

Istodobno otkrivanje zajedničkih i diskriminativnih tema putem zajedničke nenegativne matrične faktorizacije

1. veljače 2019.

Ana Parać
Sara Pužar
Petra Rožić
Paula Vujasić

Sadržaj

Sažetak

Velike zbirke podataka postale su naša svakodnevica te je često potrebno razumijevanje i obrada njih samih. Želimo uočiti pravilnosti, zajedničke teme i teme koje su specifične samo za neku određenu zbirku dokumenata. Kako bi se suočili s problemom modeliranja zajedničkih i diskriminativnih tema različitih skupova podataka, koristimo se raznim pristupima. U ovom radu obradit ćemo pristup kojim simultano određujemo zajedničke i diskriminativne teme dva različita skupa temeljen na blokovskom spustu. Potom provodimo algoritam na umjetno konstruiranim podatcima te podatcima iz stvarnog svijeta. U konačnici iznosimo rezultate testiranja.

1 Opis problema

1.1 Uvod

Analiziranje i razumijevanje velikih zbirki dokumenata provodi se kako bismo odredili zajedničke karakteristike dokumenata iz zajedničkih ili različitih područja. Primjerice, u realnom svijetu htjeli bismo usporediti zajedničke ključne riječi istih tema pisanih iz ženske ili muške perspektive. Također, nerijetko želimo promotriti razvoj specifične domene kroz vrijeme te usporediti zajedničke i diskriminativne teme radova o određenom području.

Za analizu velikih zbirki dokumenata često se koristi nenegativna matrična faktorizacija, koju ćemo i mi koristiti. Naš rad se temelji na metodi modeliranja tematskih cjelina bazirane na nenegativnoj matričnoj faktorizaciji koja istovremeno otkriva zajedničke i diskriminativne teme skupova dokumenata. Radi jednostavnosti koristit ćemo samo dva skupa dokumenata te na njima obraditi dvije metode. Kako bi istovremeno otkrili zajedničke i specifične teme, uz standardnu ciljnu funkciju za nenegativnu matričnu faktorizaciju, uvodimo dva kaznena izraza. Jednim izrazom nagrađujemo sličnost zajedničkih tema dok drugim kažnjavamo različitost tema koje nisu zastupljene u oba skupa podataka.

1.2 Matematičko modeliranje problema

1.2.1 Nenegativna matrična faktorizacija za modeliranje tema dokumenata

Matematičko modeliranje navedene teme započinjemo definicijom nenegativne matrične faktorizacije.

Definicija 1 (Nenegativna matrična faktorizacija). Neka su dani matrica $X \in \mathbf{R}_+^{m \times n}$, gdje je $\mathbf{R}_+$ skup svih nenegativnih realnih brojeva, prirodan broj $k \ll \min(m, n)$. Nenegativna matrična faktorizacija rješava aproksimaciju niskog ranga zadanu s:

$$X \approx WH^T$$

pri čemu su $W \in \mathbf{R}_+^{m \times k}$ i $H \in \mathbf{R}_+^{n \times k}$.

Aproksimacije matrica W i H mogu se pronaći pomoću različitih mjera udaljenosti, a mi ćemo koristiti Frobeniusovu normu.

Definicija 2 (Frobeniusova norma). Neka je dana matrica $A \in \mathbf{R}_+^{m\times n}$, tada Frobeniusovu normu označavamo sa $\|A\|_F$ te je jednaka sljedećem izrazu.

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i,j}^2|}$$

Cilj matrične faktorizacije jest pronaći matrice W i H takve da njihov produkt što bliže aproksimira zadanu matricu X. U našem slučaju još postoji zahtjev na nenegativnost svih matrica, pa u konačnici problem svodimo na rješavanje sljedećeg izraza.

$$\min_{W,H\geq 0} f(W,H) = \|X - WH^T\|_F^2 \tag{1}$$

U nastavku slijedi objašnjenje matrica i njihovih elemenata u kontekstu modeliranja tema.

$$X = W \cdot H^T$$

- Stupac $x_l \in \mathbf{R}_+^{m\times 1}$ predstavlja odgovarajući skup od m riječi iz l-tog dokumenta
- Skalar k odgovara broju tema
- l-ti nenegativni stupac vektor matrice W predstavlja l-tu temu kao težinsku linearnu kombinaciju svojih m ključnih riječi
- Velika vrijednost u stupčanom vektoru matrice W daje naslutiti blisku povezanost teme s odgovarajućom riječi
- l-ti stupac matrice H^T prikazuje l-ti dokument kao težinsku kombinaciju k tema

1.2.2 Simultano modeliranje zajedničkih i specifičnih tema

Dane su dvije zbirke dokumenata, gdje jedna predstavlja n_1 dokumenata, a druga predstavlja n_2 dokumenata. Naš cilj je naći k tema iz oba skupa podataka tako da vrijedi $k = k_c + k_d$, pri čemu je k_c broj zajedničkih tema, dok k_d predstavlja broj specifičnih tema.

Kao ulazi u algoritme dani su nam:

- dvije nenegativne matrice $X_1 \in \mathbf{R}_+^{m\times n_1}$, $X_2 \in \mathbf{R}_+^{m\times n_2}$ koje predstavljaju dva skupa dokumenata
- $k_c \in \mathbf{N}$, $k_d \in \mathbf{N}$

Kao izlaz iz algoritma dajemo NMF aproksimacije ulaznih matrica. Odnosno,

$$W_1, W_2, H_1, H_2$$

tako da vrijedi

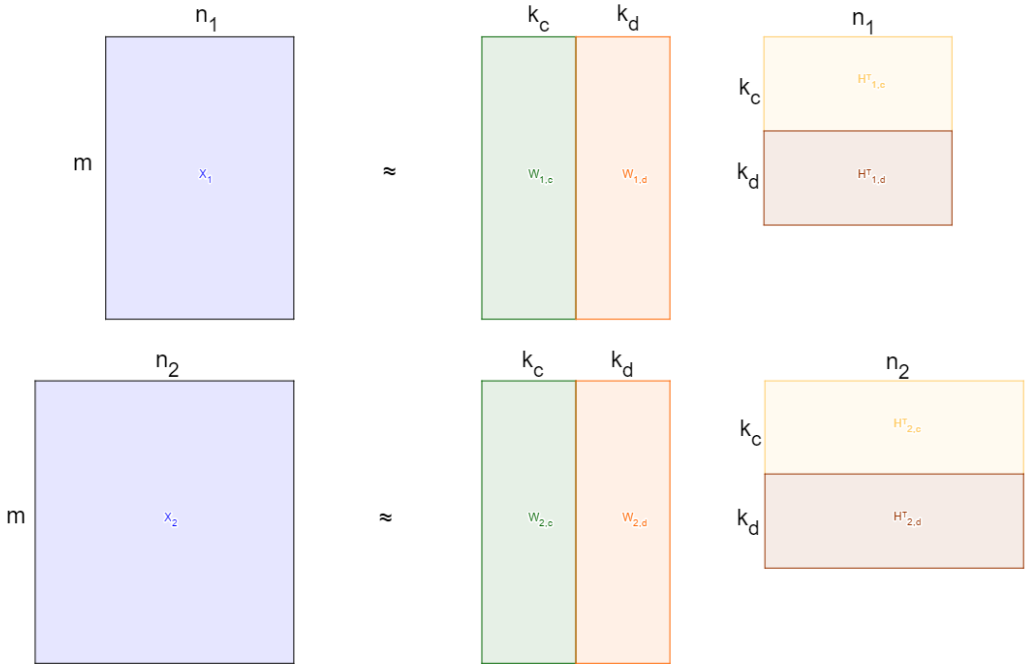
$$X_1 \approx W_1 H_1^T \ , \ X_2 \approx W_2 H_2^T,$$

pri čemu je:

$$W_i = [W_{i,c} \ \ W_{i,d}] \in \mathbf{R}_+^{m\times k}, \quad W_{i,c} \in \mathbf{R}_+^{m\times k_c}, \quad W_{i,d} \in \mathbf{R}_+^{m\times k_d}$$

$$H_i = [H_{i,c} \ \ H_{i,d}] \in \mathbf{R}_+^{n_i\times k} \text{ za } i = 1,2 \quad H_{i,c} \in \mathbf{R}_+^{n_i\times k_c}, \quad H_{i,d} \in \mathbf{R}_+^{n_i\times k_d}$$

Želimo osigurati da su skupovi zajedničkih tema reprezentirani stupcima matrica $W_{i,c}$, a diskriminativne teme stupcima matrica $W_{i,d}$.



Uvest ćemo dvije kaznene funkcije f_c i f_d za sličnost i diskriminativnost. Manja povratna vrijednost upućivat će na veću sličnost odnosno veću diskriminativnost. Koristeći i navedene dvije kaznene funkcije naš problem se svodi na:

$$\min_{W_1,H_1,W_2,H_2 \geq 0} \frac{1}{n_1}\|X_1 - W_1 H_1^T\|_F^2 + \frac{1}{n_2}\|X_2 - W_2 H_2^T\|_F^2 + \alpha f_c(W_{1,c}, W_{2,c}) + \beta f_d(W_{1,d}, W_{1,d}) \tag{2}$$

pri čemu koeficijenti $\frac{1}{n_1}$ i $\frac{1}{n_2}$ održavaju ravnotežu između različitog broja podataka u X_1 i X_2 te parametri α i β kontroliraju težine kaznenih funkcija.

2 Metode rješavanja

Problem istovremenog određivanja zajedničkih i diskriminativnih tema između dva velika skupa podataka može se riješiti na više načina. Mi ćemo prikazati :

- Batch-Processing pristup
- Pseudo-deflation pristup

2.1 Batch-Processing pristup

U ovom pristupu dane kaznene funkcije definiramo kao:

$$f_c(W_{1,c}, W_{2,c}) = \|W_{1,c} - W_{2,c}\|_F^2 \tag{3}$$

$$f_d(W_{1,d}, W_{2,d}) = \|W_{1,d}^T W_{2,d}\|_{1,1} \tag{4}$$

pri čemu $\|\cdot\|_{1,1}$ predstavlja apsolutni zbroj svih matričnih unosa. Pomoću kaznene funkcije (3) minimizirat ćemo kvadratnu sumu razlika elemenata između $W_{1,c}$ i $W_{2,c}$. U kaznenoj funkciji (4) (i,j)-ta komponenta od $W_{1,d}^T W_{2,d}$ odgovara produktu između i-tog vektora $w_{1,d}^{(i)}$ matrice $W_{1,d}$ i j-tog vektora $w_{2,d}^{(j)}$ matrice $W_{2,d}$. Minimizirajući izraz (4) i uvjetovanjem jedinične norme stupaca matrice W potičemo pojavljivanje nula. Na taj način jednu ključnu riječ uparujemo s točno jednom temom, osiguravajući veću diskriminativnost tema.

Uvrštavanjem kaznenih funkcija (3) i (4) u (2), ciljna funkcija glasi (5):

$$\min_{W_1,H_1,W_2,H_2\geq 0} \frac{1}{n_1}\|X_1 - W_1 H_1^T\|_F^2 + \frac{1}{n_2}\|X_2 - W_2 H_2^T\|_F^2 + \alpha\|W_{1,c} - W_{2,c}\|_F^2 + \beta\|W_{1,d}^T W_{2,d}\|_{1,1} \tag{5}$$

S uvjetom $\|(W_1)_{\cdot l}\|_2 = 1$, $\|(W_2)_{\cdot l}\|_2 = 1$ za $l = 1, \ldots, k$.

Nakon uvrštavanja funkcija f_c i f_d u (2), dobiveni izraz izjednačimo s 0 te deriviramo po vektorima $h_i^{(l)}, w_{i,c}^{(l)}, w_{i,d}^{(l)}$. Dobivamo sljedeća pravila ažuriranja :

$$w_{1,c}^{(l)} \leftarrow \left[\frac{(H_1^T H_1)_{ll}\ w_{1,c}^{(l)} + X_1 h_{1,c}^{(l)} - W_1 (H_1^T H_1)_{\cdot l} + n_1 \alpha w_{2,c}^{(l)}}{(H_1^T H_1)_{ll} + n_1 \alpha}\right]_+ \tag{6}$$

$$w_{1,d}^{(l)} \leftarrow \left[w_{1,d}^{(l)} + \frac{X_1 h_{1,d}^{(l)} - W_1 (H_1^T H_1)_{\cdot l} - n_1 \frac{\beta}{2} \sum_{p=1}^{k_d} w_{2,d}^{(p)}}{(H_1^T H_1)_{ll}}\right]_+ \tag{7}$$

$$h_1^{(l)} \leftarrow \left[h_1^{(l)} + \frac{(X_1^T W_1)_{\cdot l} - (H_1 W_1^T W_1)_{\cdot l}}{(W_1^T W_1)_{ll}}\right]_+, \tag{8}$$

gdje je $[x]_+ = \max(x, 0)$ i $(\cdot)$ predstavlja (l,l)-tu komponentu pomnoženih matrica unutar zagrada. Nakon ažuriranja, vektor $w_1^{(l)}$ normaliziramo kako bi imao jediničnu normu te $h_1^{(l)}$ pomnožimo s odgovarajućim $\|w_1^{(l)}\|_2$ za $l = 1, ..., k$.

Analogno ažuriramo vektore $w_{2,c}^{(l)}$, $w_{2,d}^{(l)}$, $h_2^{(l)}$

2.2 Pseudo-deflation

Pseudo-deflation pristup razmatra samo najreprezentativnije ključne riječi u svakoj temi te otkriva diskriminativne teme jednu po jednu.

Kao u Batch-Processing pristupu, i ovaj pristup će se voditi kaznenim funkcijama f_c i f_d. U kaznenoj funkciji (4) promatrali smo sve ključne riječi, a u ovom pristupu ograničavamo se samo na najreprezentativnije riječi. Stoga, neka je dan fiksan broj t te neka $R_{1,d}^{(i)}$ i $R_{2,d}^{(j)}$ predstavljaju skup od t najreprezentativnijih indeksa ključnih riječi iz dvaju vektora $w_{1,d}^{(i)}$ i $w_{2,d}^{(j)}$. Tada kaznenu funkciju f_d možemo definirati sljedećim izrazom:

$$\big(f_d(W_{1,d}, W_{2,d})\big)_{ij} = \Big(w_{1,d}^{(i)}\Big)^T I_m\Big(R_{1,d}^{(i)} \cup R_{2,d}^{(j)}\Big) w_{2,d}^{(j)}, \tag{9}$$

pri čemu je dijagonalna matrica $I_m(S) \in \mathbf{R}_+^{m\times m}$ definirana kao:

$$\Big(I_m(S)\Big)_{pp} = \begin{cases} 1, p \in S \\ 0, p \notin S \end{cases} . \tag{10}$$

Uočimo da je $S \subset 1, ..., m$ skup indeksa ključnih riječi. Stoga, S odabiremo kao $R_{1,d}^{(i)} \cup R_{2,d}^{(j)}$ iz čega slijedi da u kaznenoj funkciji f_d koristimo samo najreprezentativnije indekse ključnih riječi iz oba skupa dokumenata.

Glavni problem ovakvog pristupa je da se skupovi $R_{1,d}^{(i)}$ i $R_{2,d}^{(j)}$ dinamički mijenjaju kako se $w_{1,d}^{(i)}$ i $w_{2,d}^{(j)}$ ažuriraju tijekom iteracija. To uzrokuje da se naša ciljna funkcija (2) mijenja tijekom iteracija i zato nema više garancije da algoritam monotono poboljšava vrijednost ciljne funkcije.

Kako bismo prevladali taj problem koristimo pseudo-deflation pristup koji rješava ciljnu funkciju (2) ugrađujući kaznenu funkciju (9). Osnovna ideja je naći diskriminativne teme na pohlepan način kako bi se zadržao fiksirani skup najreprezentativniji indeksa ključnih riječi za svaku temu. Odnosno, nalazimo jedan diskriminativni tematski par po etapi te je nakon k_d etapa otkriveno cijelo rješenje. U svakoj etapi odabrane teme $w_{1,d}^{(l)}$ i $w_{2,d}^{(l)}$ moraju se razlikovati od tematskih parova prethodnih etapa, $\{w_{2,d}^{(l)}, ..., w_{2,d}^{(l-1)}\}$ i $\{w_{1,d}^{(l)}, ..., w_{1,d}^{(l-1)}\}$ respektivno te jedna od druge.

Kako bi kreirali algoritam, na početku inicijaliziramo pomoćne varijable k_c^s i k_d^s tako da vrijedi $k_c^s = k_c + k_d$ i $k_d^s = 0$, pri čemu k_c^s predstavlja privremeni broj zajedničkih, a k_d^s diskriminativnih tema u svakoj etapi. U svakoj sljedećoj etapi smanjujemo k_c^s i povećavamo k_d^s za jedan te rješavamo novu ciljnu funkciju:

$$\min_{W_{1,c}, W_{1,d}^{(k_d^s)}, H_1, W_{2,c}, W_{2,d}^{(k_d^s)} H_2 \geq 0} \frac{1}{n_1}\|X_1 - W_1 H_1^T\|_F^2 + \frac{1}{n_2}\|X_2 - W_2 H_2^T\|_F^2$$
$$+ \frac{\alpha}{k_c^s}\sum_{l=1}^{k_c^s}\|W_{1,c}^{(l)} - W_{2,c}^{(l)}\|_F^2 + \frac{\beta}{k_d^s - 1}\sum_{l=1}^{k_d^s-1}\|(W_{1,d}^{(l)})^T I_m(R_{1,d}^{(l)}) w_{2,d}^{k_d^2} + \gamma(w_{1,d}^{(k_d^s)})^T(w_{2,d}^{(k_d^s)}) \tag{11}$$

Izraz $\gamma(w_{1,d}^{(k_d^s)})^T(w_{2,d}^{(k_d^s)})$ iz prethodne jednadžbe ima ulogu povećanja razlike između para tema $w_{1,d}^{(k_d^s)}$ i $w_{2,d}^{(k_d^s)}$. Ova metoda sadrži različite elemenate koji pridonose kaznenim funkcijama kako se k_c^s i k_d^s mijenjaju. Stoga, za razliku od konstantnih parametara α i β iz prethodnog pristupa, u ovom pristupu parametre dijelimo s k_c^s i $k_d^s - 1$ respektivno.

U konačnici, pravila ažuriranja možemo definirati kao:

$$w_{1,c}^{(l)} \leftarrow \left[\frac{(H_1^T H_1)_{ll}\ w_{1,c}^{(l)} + X_1 h_{1,c}^{(l)} - W_1 (H_1^T H_1)_{\cdot l} + n_1 \frac{\alpha}{k_c^s} w_{2,c}^{(l)}}{(H_1^T H_1)_{ll} + n_1 \alpha} \right]_+ \tag{12}$$

$$w_{1,d}^{(k_d^s)} \leftarrow \left[w_{1,d}^{(k_d^s)} + \frac{X_1 h_{1,d}^{(k_d^s)} - W_1 (H_1^T H_1)_{\cdot k_d^s} - n_1 \frac{\beta}{2(k_d^s - 1)} \sum_{p=1}^{k_d^2 - 1} I(R_{2,d}^{(p)}) w_{2,d}^{(p)} + n_1 \frac{\gamma}{2} w_{2,d}^{(k_d^s)}}{(H_1^T H_1)_{k_d^s k_d^s}} \right]_+ \tag{13}$$

$$h_1^{(l)} \leftarrow \left[h_1^{(l)} + \frac{(X_1^T W_1)_{\cdot l} - (H_1 W_1^T W_1)_{\cdot l}}{(W_1^T W_1)_{ll}} \right]_+ \tag{14}$$

Nakon ažuriranja, vektor $w_1^{(l)}$ normaliziramo kako bi imao jediničnu normu te $h_1^{(l)}$ pomnožimo s odgovarajućim $\|w_1^{(l)}\|_2$ za $l = 1, ..., k$.

Analogno ažuriramo vektore $w_{2,c}^{(l)}$, $w_{2,d}^{(l)}$, $h_2^{(l)}$.

Kriterij po kojem odabiremo l-ti par tema iz $W_{1,c}$ i $W_{2,c}$ koji ćemo u sljedećem koraku preseliti na kraj $W_{1,d}$, odnosno $W_{2,d}$ glasi:

$$\arg \min_{w_{1,c}^{(l)}, w_{2,c}^{(l)}} \sum_{i=1}^{2} \|w_{i,c}^{(l)} (h_{i,c}^{(l)})^T - \max w_{i,c}^{(l)} (h_{i,c}^{(l)})^T - X_i, 0_{m \times n}\|_F^2, \tag{15}$$

pri čemu se max operacija primjenjuje po elementima.

3 Opis implementacije algoritma

Uz obrađivanje same teme simultanog pronalaska zajedničkih i diskriminativnih tema, implementirale smo algoritme koji se zasnivaju na *Batch processing* pristupu i *Pseudo deflation* pristupu. U nastavku navodimo pseudokodove implementiranih algoritama.

3.1 *Batch processing* pristup

Algorithm 1 NMF bazirana na batch-processing pristupu

Ulaz: Dvije ulazne matrice X_i i x_2, skalari k_c i k_d, i parametri α i β
Izlaz:
$\mathrm{W}_i = [W_{i,c}, W_{i,d}] \in \mathbf{R}_+^{m \times k}$
$\mathrm{H}_i = [H_{i,c}, H_{i,d}] \in \mathbf{R}_+^{n_i \times k}$ za $i = 1, 2$

Inicijalizacija W_i i H_i za $i = 1, 2$ pomoću SVD-dekompozicije
repeat
  Ažuriraj $W_{i,c}$ koristeći (6)
  Ažuriraj $W_{i,d}$ koristeći (7)
  Ažuriraj H_i koristeći (8)
  Normaliziraj stupce W_i-ova tako da imaju jediničnu normu i odgovarajućim skalarom pomnoži H_i
  Vrijednosti manje od 0 u ažuriranim $W_{i,c}$, $W_{i,d}$ i H_i postavi na 0
until;
dok kriterij zaustavljanja nije zadovoljen

3.2 *Pseudo deflation* pristup

Algorithm 2 NMF bazirana na pseudo-deflation pristupu

Input : Dvije ulazne matrice X_i i x_2, skalari k_c i k_d, i parametri α, β i γ
Output:
$\mathrm{W}_i = [W_{i,c}, W_{i,d}] \in \mathbf{R}_+^{m \times k}$
$\mathrm{H}_i = [H_{i,c}, H_{i,d}] \in \mathbf{R}_+^{n_i \times k}$ for $i = 1, 2$

Inicijaliziraj W_i i H_i za $i = 1, 2$ pomoću SVD-dekompozicije
for $k_d^s \leftarrow 0$ *to* k_d *do* **do**
  $\mathrm{k}_c^s \leftarrow k_c + k_d - k_d^s$;
  //Za k_c^s i k_d^s riješi (11)
  repeat
    Ažuriraj W_i koristeći (12),(13)
    Ažuriraj H_i koristeći (14)
    Normaliziraj stupce W_i-ova tako da imaju jediničnu normu i odgovarajućim skalarom pomnoži H_i
    Vrijednosti manje od 0 u ažuriranim $W_{i,c}$, $W_{i,d}$ i H_i postavi na 0
  until;
  Dok kriterij zaustavljanja nije zadovoljen
  Izaberi $w_{1,c}^l$ i $w_{i,d}$ tako da zadovoljava (15)
  // Ukloni $w_{i,c}^l$ iz W_i^c
  $W_{i,c} \leftarrow \mathrm{W}_{i,c} \backslash w_{i,c}^l$ za $i = 1, 2$
  // Nadodaj $w_{i,c}^l$ na $W_{i,d}$ sa desne strane
  $W_{i,d} \leftarrow [\ \mathrm{W}_{i,d}\ w_{i,c}^l]$ za $i = 1, 2$
end

4 Rezultati testiranja

4.1 Sintetizirani i stvarni podaci

Navedene algoritme razvijale smo na umjetno konstruiranim podatcima, a kasnije ih testirale na podatcima preuzetim sa stranice *https://www.cc.gatech.edu/gvu/ii/jigsaw/datafiles.html*.

Umjetne podatke smo konstruirale započevši s konstrukcijom matrica W_1, W_2, H_1 i H_2. Matrice W_1, W_2, H_1 i H_2 konstruirane su prateći sljedeća pravila :

$$\left(w_{i,c}^{(l)}\right)_p = \begin{cases} 1, & 100(l-1) < p \leq 100l \\ 0, & \text{inače} \end{cases} \tag{16}$$

$$\left(w_{i,d}^{(l)}\right)_p = \begin{cases} 1, & idx(i,l) < p \leq idx(i,l) + 100 \\ 0, & \text{inače} \end{cases} \tag{17}$$

pri čemu je: $idx(i,l) = 600 + 400(i-1) + 100(l-1)$

Množenjem odgovarajućih matrica dobivamo testne matrice X_1 i X_2, $W_1 \cdot H_1' = X_1$ i $W_2 \cdot H_2' = X_2$

Matrice koje reprezentiraju stvarne dokumente konsturirale smo koristeći podatke o dokumentima korištenima na InfoVis i Vast konferenciji.

4.2 Procjena točnosti

Kako bismo objektivno procijenile preciznost rezultata koje smo dobile navedenim algoritmima, implementirale smo tri funkcije greške koje procjenjuju neki aspekt točnosti pronalaska zajedničkih i diskriminativnih tema.

Algorithm 3 Rekonstrukcijska greška

Input : Šest ulaznih matrica X_1, X_2, W_1, W_2, H_1, H_2
Output:
Skalar procjenaGreške
return $\frac{1}{n_1}\|X_1 - W_1 H_1^T\|_F^2 + \frac{1}{n_2}\|X_2 - W_2 H_2^T\|_F^2$

Algorithm 4 Rezultat sličnosti

Input : Dvije ulazne matrice $W_{1,c}$, $W_{2,c}$ (prvih k_c vektora matrice W_i) i skalar k_c
Output:
Skalar procjenaGreške
return $\frac{1}{k_c}\|W_{1,c} - W_{2,c}\|_F^2$

Algorithm 5 Rezultat različitosti

Input : Dvije ulazne matrice $W_{1,d}$, $W_{2,d}$ (zadnjih k_d stupaca matrice W_i) i skalar k_d
Output:
Skalar procjenaGreške
return $\frac{1}{2k_d^2}\sum_{i=1}^{k_d}\sum_{j=1}^{k_d}[(W_{1,d}^{(i)})^T log(W_{1,d}^{(i)}) + (W_{2,d}^{(j)})^T log(W_{2,d}^{(j)}) - (W_{1,d}^{(i)})^T log(W_{2,d}^{(j)}) - (W_{2,d}^{(j)})^T log(W_{1,d}^{(i)})]$

4.3 Tablični prikaz rezultata

U sljedećim tablicama prikazujemo rezultate testiranja algoritama pokrenutih s paramterima:

- $\alpha = 100$
- $\beta = 10$
- $\gamma = 10$
- $k_c = 10$
- $k_d = 8$

Navedene parametre α i β smo preuzele iz literature. Parametre k_c i k_d smo odredile testiranjem.

4.3.1 Batch processing pristup

Batch-processing pristup			
Podatci	Rekonstrukcijska greška	Rezultat sličnosti	Rezultat različitosti
InfoVis - Vast	15.3477	2.014e-007	103.8173

4.3.2 Pseudo-Deflation pristup

Pseudo deflation pristup			
Podatci	Rekonstrukcijska greška	Rezultat sličnosti	Rezultat različitosti
InfoVis - Vast	14.9997	0.2003	148.8885

5 Literatura

Literatura

[1] H. Kim, Choo, J. Kim, Reddy, Park : *Simultaneous Discovery of Common and Discriminative Topics via Joint Nonnegative Matrix Factorization*,

[2] Kuang, Choo, Park: *Nonnegative matrix factorization for interactive topic modeling and document clustering*

[3] Kim, He, Park: *Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework*