Finding similar items

Uvod

- Mnogi zadaci mogu da se definišu kao pronalaženje sličnih skupova
  - Najbliži susjed
- Primjeri
  - Dokumenti sa sličnim riječima, detekcija duplikata, klasifikacija po tematici
  - Klijenti koji kupuju slične proizvode, proizvodi sa zajedničkim kupcima
  - Korisnici koji posjete slične sajtove

Distance measure

- Cilj, naći najbližeg susjeda u više-dimenzionom prostoru
- Jakardovo rastojanje/sličnost
  - Za dva skupa definiše se kao količnik veličine presjeka i veličine unije ta dva skupa
  - similarity(A, B) = 1 – distance(A, B)

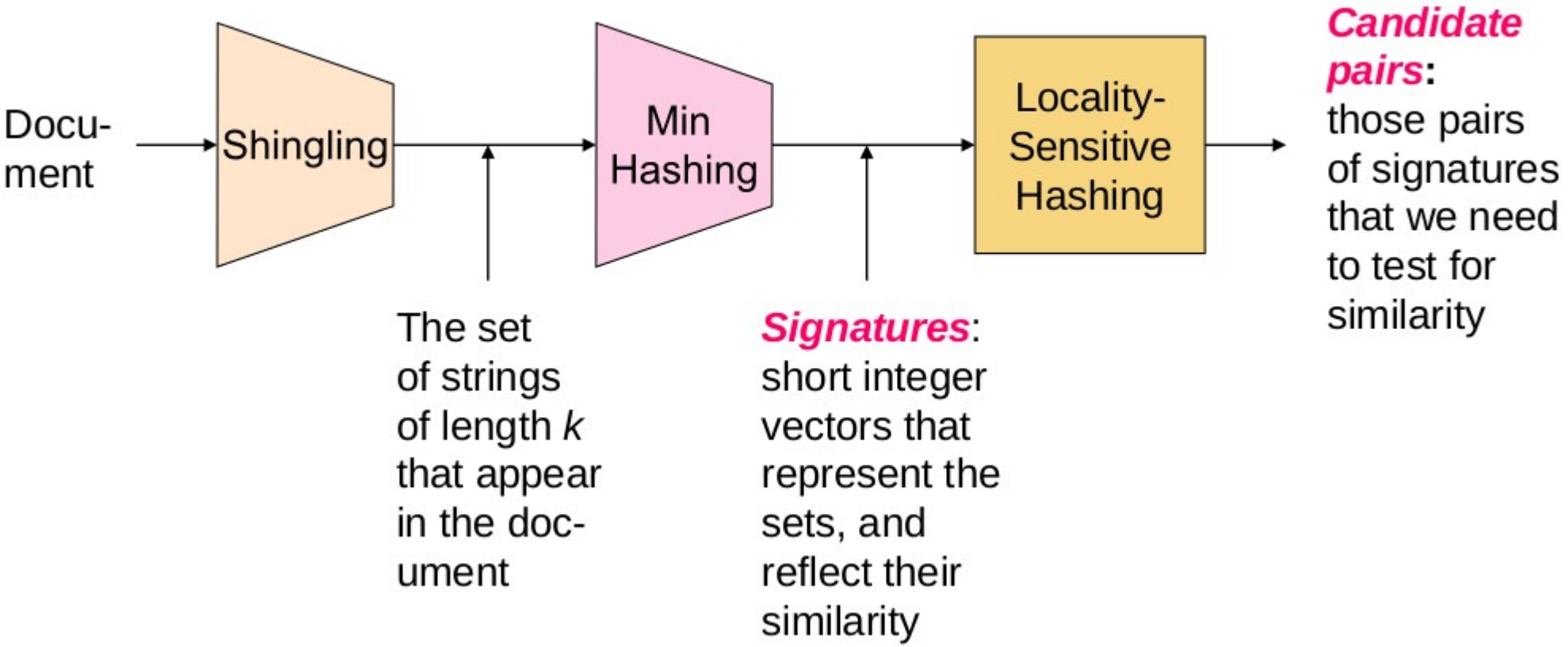
Identifikovanje sličnih dokumenata

- Zadatak, za dati skup od N (nekoliko miliona) tekstualnih dokumenata pronaći parove koji su skoro duplikati - “near duplikates”
- Primjene
  - Mirror websites, ne prikazivati oba u rezultatima pretrage
  - Slični novinski članci, klasterisati članke po temi
- Problemi
  - Slučajne pojave dijelova jednog dokumenta u drugom
  - Previše parova za poređenje
  - Ograničenja operativne meorije (broj dokumenata ili veličina)

Osnovni koraci rješenja

- Shingling – predstavljanje dokumenata u formi skupova
- Minhashing – konverzija velikih skupova u male *potpise*, uz zadržavanje sličnosti
- Locality-sensitive hashing – formiranje parova potpisa koji su sa velikom vjerovatnoćom iz sličnih skupova, kandidatski parovi

Kompletna slika



Shingling

- Konverzija dokumenata u skupove
- Pristupi
  - dokument = skup riječi koje se pojavljuju
  - dokument = skup “ključnih” riječi
  - Nijesu prihvatljiva rješenja, jer ne uzima u obzir redosljed riječi
- Drugačiji pristup: shingling

Definicija za shingles

- k-shingle je sekvenca k tokena koji se pojavljuju u dokumentu
  - Token je karakter, riječ ili nešto treće, zavisno od aplikacije, podrazumijevamo karakter
- Primjer: k = 2, dokument D_1 = abcab, skup 2 – shinglova: $S(D_1)$ = {ab, bc, ca}
  - Razmotriti varijantu sa multiskupom

Kompresija shinglova

- Za kompresiju dugih šinglova koristi se heširanje
- Reprezentacija dokumenta skupom heš vrijednosti njegovih k-shinglova
  - Moguće je (rijetko) da dva dokumenta prividno dijele shinglove, dok se ustvari dijele samo heš vrijednosti
- Primjer: k = 2, dokument D_1 = abcab, skup 2 – shinglova: $S(D_1)$ = {ab, bc, ca}, heširanje $h(D_1)$={1, 5, 7}

Mjera sličnosti za shinglove

- Dokument D_1 = skup k-shinglova $C_1 = S(D_1)$
- Ekvivalentno, svaki dokument je 0/1 vektor u prostoru k-shinglova
  - Svaki shingle je dimenzija
  - Vektori su rijetki (sa mnogo nula)
- Jakardova mjera za sličnost
  - $SIM(D_1, D_2) = |C_1 * C_2| / |C_1 + C_2|$

Pretpostavka

- Dokumenti sa velikim brojem zajedničkih shinglova imaju sličan tekst, iako je poredak riječi eventualno različit
- Napomena: potrebno je izabrati k dovoljno veliko, inače će “većina” dokumenata sadržati većinu shinglova
  - k = 5 za kraće dokumente
  - k = 10 za duže dokumente

Motivacija za LSH algoritam

- Zadatak: naći duplikate među N = milion dokumenata
- Algoritam grube sile, za svaku par dokumenata izračunati Jakardovo rastojanje
  - $N*(N-1)/2 = 5 * 10^{11}$ poređenja
  - Za 10^5 sekudni/danu i 10^6 poređenja/sekundi, zadatak bi trajao 5 dana
- Za N = 10 miliona, rješavanje je duže od godine dana

Minhashing

- Reprezentacija skupova bit vektorima
  - Jedna dimenzija za svaki element
- Formalizacija problema: identifikovanje skupova sa “značajno velikim” presjekom
- Realizacija
  - Presjek je AND
  - Unija je OR
  - Primjer, C_1 = 10111, C_2 = 10011, veličina presjeka = 3, veličina unije = 4, SIM = ¾, DISTANCE = 1 – 3/4

Formiranje matrice bita za skupove

- Redovi = elementi (shingles)
- Kolone = skupovi (dokumenti)
  - 1 u redu e i koloni s ako je e element s
  - SIM mjera za kolone je Jakardova mjera za sličnost odgovarajućih skupova (redovi sa 1)
  - Rijetka matrica
- Svaki dokument je predstavljen kolonom
  - Primjer: SIM(C_1, C_2) = ?

1	1	1	0
1	1	0	1
0	1	0	1
0	0	0	1
1	0	0	1
1	1	1	0
1	0	1	0

Identifikovanje sličnih kolona

- Do sada:
  - Dokumenti → skupovi shinglova
  - Skupovi → Bulovi vektori u matrici
- Sljedeći korak: identifikovati slične kolone
- Algoritam
  - Signatura kolone
  - Razmatranje signatura
  - Provjera da su kolone sa sličnim signaturama zaista slične
  - Poređenje signatura: LSH algoritam (potencijalno sa false negative i false positive)

Heširanje kolona (signatura)

- Ideja: heširati kolonu C u signaturu h(C) tako da:
  - h(C) je dovoljno malo zbog ograničenja RAMa
  - $SIM(C_1, C_2) = SIM(h(C_1), h(C_2))$
- Potrebno je naći funkciju h(*) tako da:
  - Ako je $SIM(C_1, C_2)$ velika, onda je velika vjerovatnoća da $h(C_1) = h(C_2)$ i obratno
- Heširamo dokumente u bakete i očekujemo da duplikati budu heširani u iste bakete

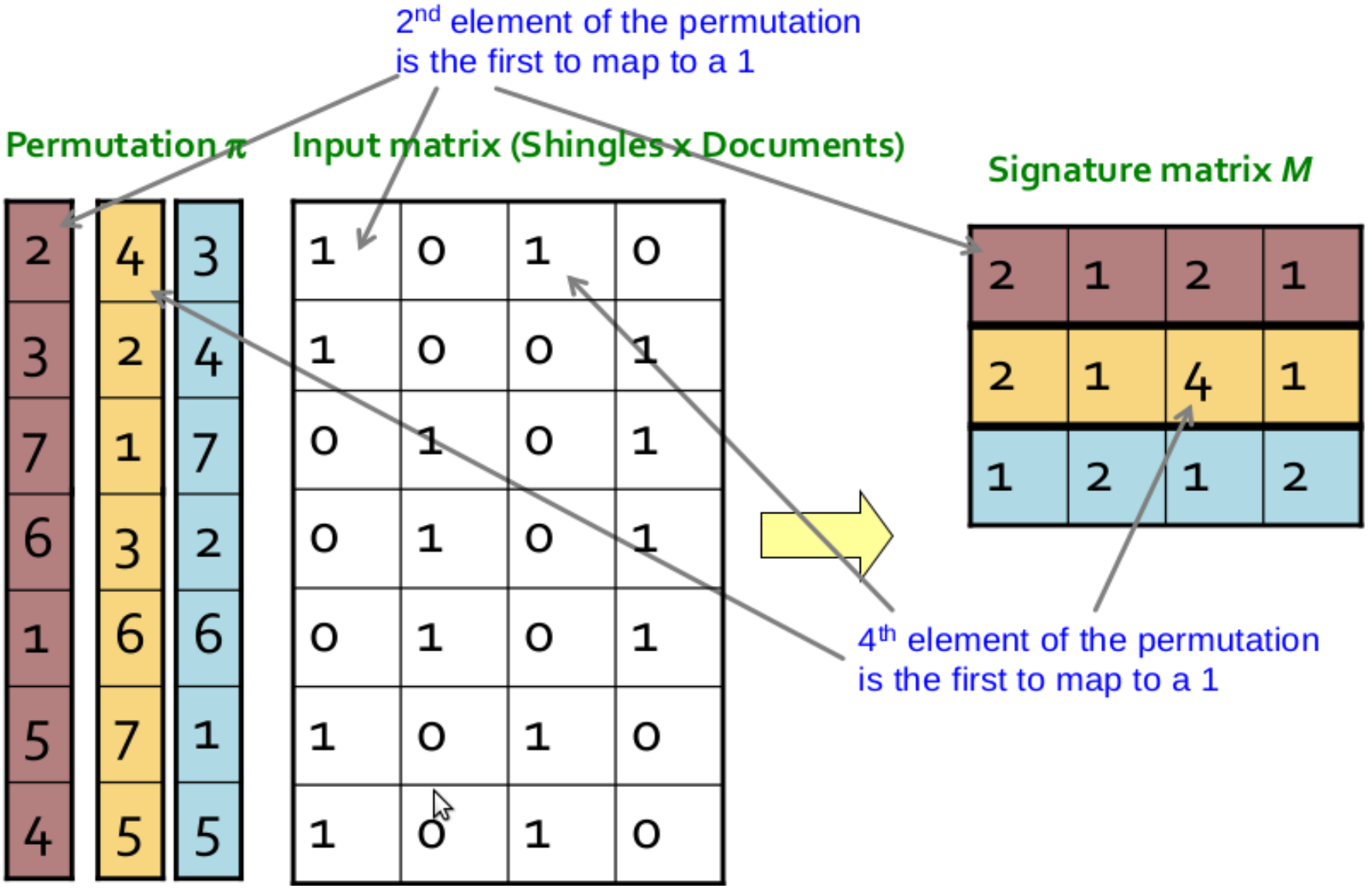
Min-hashing

- Heš funkcija zavisi od izabrane mjere sličnosti
  - Postoje mjere za koje nije moguće definisati odgovarajuću heš funkciju
- Min-hašing je odgovarajuća za Jakardovu mjeru

Min-hashing (2)

- Permutacija P redova u matrici dokumenata
- Definišimo h(C) = redni broj “prvog” reda u koloni C koji sadrži 1
- Koristimo nekoliko nezavisnih heš funkcija da formiramo signaturu kolone

Primjer za Min-hashing



Svojstvo

- $Pr(h(C_1) = h(C_2)) = SIM(C_1, C_2)$
- Dokazati

Četiri tipa za redove

- Date su kolone C_1 i C_2

	C_1	C_2
A	1	1
B	1	0
C	0	1
D	0	0

- a = broj redova tipa A, itd.
- $SIM(C_1, C_2) = a / (a + b + c)$
- Važi: $Pr[h(C_1) = h(C_2)] = SIM(C_1, C_2)$

Sličnost za signature

- Mjera sličnosti za dvije signature je procenat heš funkcija koje se poklapaju na tim signaturama
- Sličnost za kolone je ista kao sličnost za signature

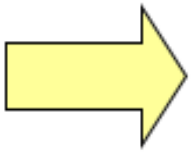
Min-hashing signature

Permutation π

2	4	3
3	2	4
7	1	7
6	3	2
1	6	6
5	7	1
4	5	5

Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0



Signature matrix *M*

2	1	2	1
2	1	4	1
1	2	1	2

Similarities:

	1-3	2-4	1-2	3-4
Col/Col	0.75	0.75	0	0
Sig/Sig	0.67	1.00	0	0

Implementacija

- Bira se n=100 permutacija redova, umjesto permutacija može se uzeti n heš funkcija
- Formira se matrica signatura SIG(*i, c*) – element matrice za heš funkviju *i*, a kolonu *c*, inicijalno su svi postavljeni na beskonačno
- Za svaki red *r*
  - Ako red sadrži 1, za svako i=1, 2, …, n postavi SIG(i, c) = min(SIG1(i, c), h_i(r))

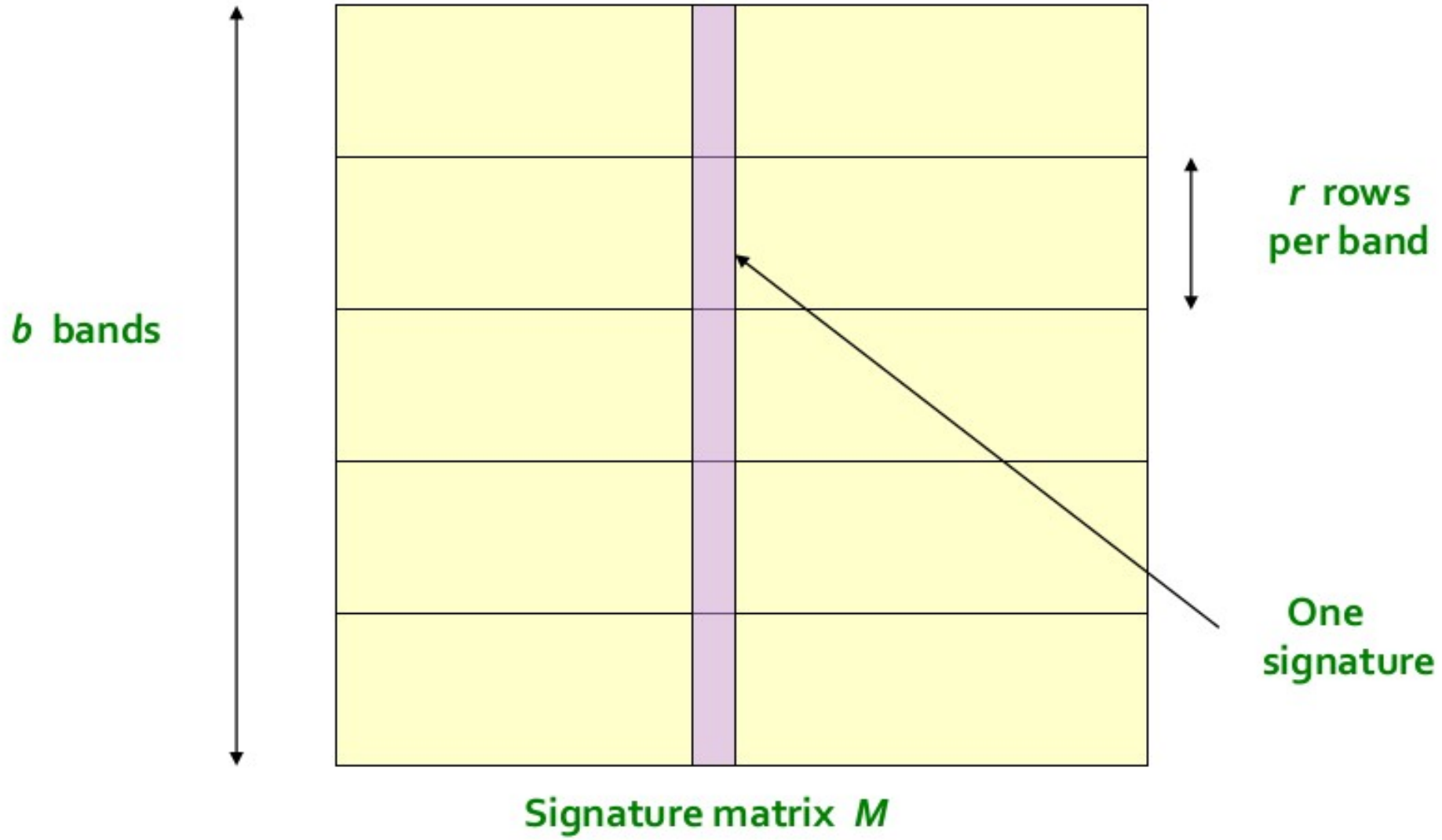
LSH

- Ostvareno je predstavljanje dokumenata sa “malim” signaturama
- Dalji cilj je naći dokumente sa SIM > s, s je ulazni parametar
- LSH ideja, koristi se funkcija f(x, y) koja određuje da li su x i y kandidatski par
  - Heširaju se kolone matrice sa signaturama, dokumenti koji pripadaju istim baketima formiraju kandidatske parove

Kandidati

- Izabere se s, $0 < s < 1$
- Kolone x, y iz matrice signatura su kandidatski par ako se poklapaju u s%
- Očekuje se da sličnost originalnih dokumenata bude ista kao sličnost signatura

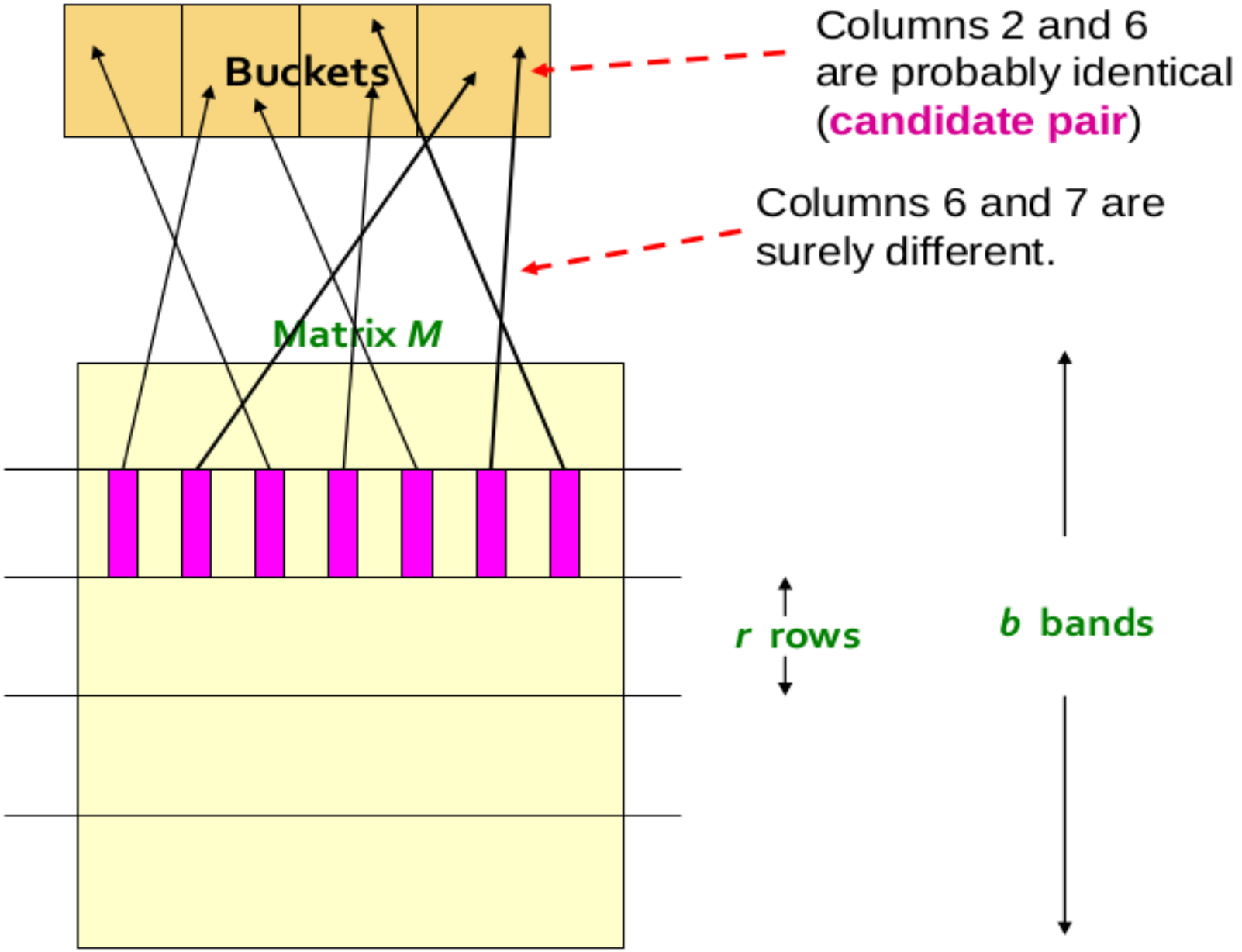
LSH za matricu signatura



Objašnjenje

- Matrica signatura M podijeli se na b baketa sa po r redova
- Za svaki baket se sprovodi heširanje po dijelu kolone koji pripada tom baketu (r redova)
- Kandidati su oni koji imaju istu vrijednost heš funkcije u makar jednom baketu

Primjer


Buckets
Columns 2 and 6
are probably identical
(candidate pair)
Columns 6 and 7 are
surely different.
Matrix M
r rows
b bands

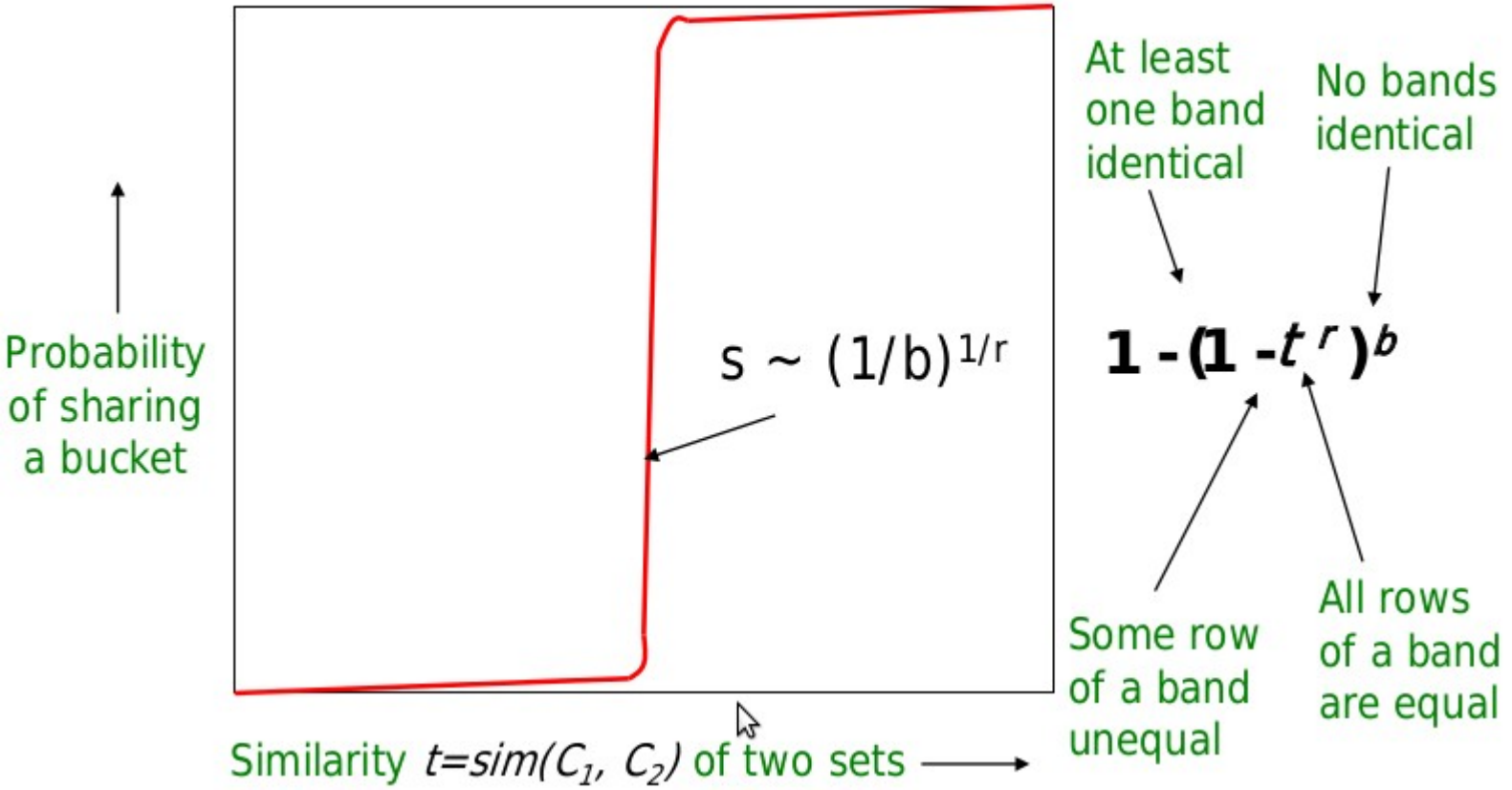
Pretpostavka

- Postoji dovoljan broj baketa tako da se kolone dodjeljuju istom baketu samo ako se poklapaju na r redova koji pripadaju posmatranom baketu

Primjer

- Neka je b=20, r=5, tražimo parove sa s>=0.8
- Ako je SIM(C_1, C_2) = 80%, oni treba da budu kandidatski par, tada C_1 i C_2 moraju da se poklapaju u bar jednom od 20 baketa
- Vjerovatnoća da se C_1, C_2 poklapaju u jednom baketu je $(0.8)^5 = 0.328$
- Vjerovatnoća da se C_1, C_2 ne poklapaju ni u jednom baketu $(1-0.328)^{20}$

Kako se biraju b i r?



Određivanje praga sličnosti

- Skupovi C_1 i C_2 imaju sličnost *t*
- Za proizvoljni baket sa *r* redova
  - Vjerovatnoća da su svi redovi u baketu isti = t^r
  - Vjerovatnoća da je bar jedan red različit = $1 – t^r$
- Vjerovatnoća da ne postoje dva ista baketa

  $= (1 – t^r)^b$
- Vjerovatnoća da su je bar jedan baket identičan

  $= 1 - (1 – t^r)^b$
- Prag je $(1/b)^{1/r}$

Algoritam

- Shingling, predstavljanje dokumenata skupovi, heširanje da svaki shingle dobije cjelobrojni ID
- Minhashing, predstavljanje velikih skupova preko malih signatura uz zadržavanje sličnosti polaznih dokumenata
- Locality-sensitive hashing, formiraju se parovi kandidatskih signatura koje potiču iz dokumenata sa sličnosti većom od zadatog praga