Locality sensitive hashing

- Binary vectors

$K, L \rightarrow$ # repetitions

↳ projection

I_1
I_2
$\vdots$
I_L
} sets of K positions each

$h_{I_j}(p) = p_{i_1} p_{i_2} \dots p_{i_K}$

$i_j \in I_j$

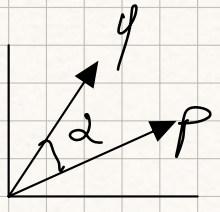
COSINE SIMILARITY

Given vectors $p, q \in \mathbb{R}^m_{\geq 0}$, we define cosine similarity as

$$\cos\theta = \frac{p \cdot q}{\|p\| \cdot \|q\|} \rightarrow [0,1]$$

$$p \cdot q = \sum_{i=1}^{m} p_i q_i$$

$$\|q\| = \sqrt{\sum_{i=1}^{m} p_i^2}$$



BIG VECTORS

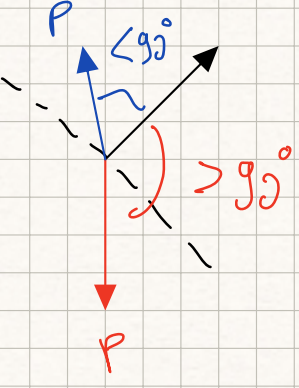
SKETCHING FOR COSINE SIMILARITY

pick random $z \in \mathbb{R}^m$, define $h_z(p) = \text{sign}(p \cdot z)$

± 1

"distant vector"

$$P(h_z(p) = h_z(q)) = 1 - \frac{\theta}{\pi}$$



random vectors $r_1, r_2, \dots, r_k$

$$\forall p \in D: \quad SKETCH(p) = \langle h_{r_1}(p), \dots, h_{r_k}(p) \rangle \in \{-1, +1\}^k$$
$$\in \mathbb{R}^m_{\geq 0}$$

$$\frac{SKETCH(p) \setminus SKETCH(q)}{k} \quad \frac{\#\text{ equal components}}{k} \simeq \frac{k \cdot P(h_r(p) = h_r(q))}{k} \simeq 1 - \frac{d}{\pi}$$

MIN HASHING

$A = \{1, 4, 6, 8, 9\}$ → SHINGLING + CARPE DIEM

$B = \{2, 3, 4, 6, 7, 8\}$

$C = \{2, 5, 11, 12\}$

use MIN-HASHING to estimate Jaccard similarity by using a sketch of size 2, assuming $U = \{0, \dots, 13\}$

permutations: $\pi_1(x) = 2x + 1 \bmod 13$

$\pi_2(x) = x + 2 \bmod 13$

$$J(A,B) = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

	π_1	π_2	
A	3, 9, **0**, 4, 6	**3**, 6, 8, 10, 14 → SKETCH(A) = ⟨0, 3⟩	$\frac{1}{2}$
B	5, 7, 9, **0**, 2, 4	**4**, 5, 6, 8, 9, 10 → SKETCH(B) = ⟨0, 4⟩	

C [5], 11, 10, 12 | 4, 7, [0], 1 → sketch(c) = ⟨5, 0⟩

$\tilde{S}(A,B) = 1/2$

$\tilde{S}(A,C) = 0/2 = 0$

$\tilde{S}(B,C) = 0/2 = 0$

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{8}$$

T_1 = white xmas

T_2 = xmas xmas happy

T_3 = happy white

T_4 = red red

Compute IL by encoding the porting list with gap of

happy → 2, 3

red → 4_2

white → $1, 3$

xmas → $1, 2_2$

GAP encoding

2, 1
4
1, 2
1, 1

1-1 "0"

2-1 "0" 2 1

0101
00100
1010
11

cdl		t_1	T_2	T_3	T_4	Q
$l_{y_2}\frac{4}{2}=1$	happy	0	1.1	1.1	0	0

$$idl_{t,d} = f_{req} \cdot l_{y_2} \frac{4}{\mu_t}$$

t,d

$\log_2 \frac{n}{n_1} = 2$	real	0	0	0	$2 \cdot 2$	$1 \cdot 2$
1	white	$1 \cdot 1$	0	$1 \cdot 1$	0	0
1	X mas	$1 \cdot 1$	$2 \cdot 1$	0	0	$1 \cdot 1$

Q = real X mas

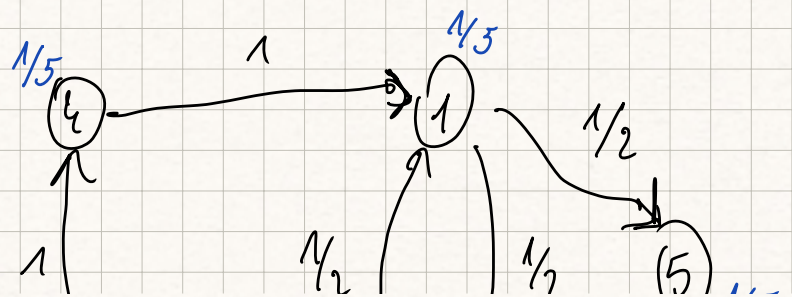
only color product

$\widehat{\cos}(\tau_1, Q) = 1$

$\widehat{\cos}(\tau_2, Q) = 2$

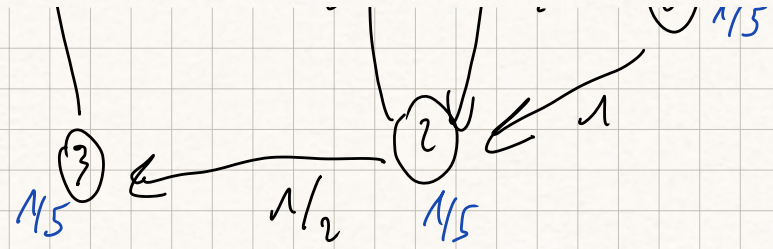
$\widehat{\cos}(\tau_3, Q) = 0$

$\widehat{\cos}(\tau_4, Q) = 8$ → most similar



uniform starting dist.

$\alpha = \frac{1}{2}$, $N = 5$



$$PR[1] = \frac{1}{2}\left(PR[4]\cdot 1 + PR[2]\cdot\frac{1}{2}\right) + \frac{1}{2}\cdot\frac{1}{5}$$

$$= \frac{1}{2}\left(\frac{1}{5}\cdot 1 + \frac{1}{5}\cdot\frac{1}{2}\right) + \frac{1}{2}\cdot\frac{1}{5} =$$

$$= \frac{1}{2}\cdot\frac{3}{10} + \frac{1}{10} = \frac{5}{20} = \frac{1}{4} = \frac{5}{20}$$

$$PR[2] = \frac{1}{2}\left(\frac{1}{5}\cdot\frac{1}{2} + \frac{1}{5}1\right) + \frac{1}{10} = \frac{1}{4}$$

$$= \frac{5}{20}$$

$$PR[3] = \frac{1}{2}\left(\frac{1}{5}\cdot\frac{1}{2}\right) + \frac{1}{10} = \frac{3}{20}$$

$$PR[4] = \frac{1}{2}\left(\frac{1}{5}\cdot 1\right) + \frac{1}{10} = \frac{2}{10} = \frac{4}{20}$$

$$PR[5] = \frac{1}{2}\left(\frac{1}{5}\cdot\frac{1}{2}\right) + \frac{1}{10} = \frac{3}{20}$$

Compute the Personalized PR with respect to $\{1,5\} = S$

$P_R(u) = \alpha \cdot$ [hatched] $+ (1-\alpha) \cdot$ [hatched] $\begin{cases} \rightarrow u \in S \quad 1/|S| \\ \rightarrow u \notin S \quad 0 \end{cases}$

$P_R[1] = 3/20 + \frac{1}{2} \cdot \frac{1}{2} = \frac{8}{20}$

$P_R[2] = 3/20 + \frac{1}{2} \cdot 0 = 3/20$

$P_R[3] = 1/20 + \frac{1}{2} \cdot 0 = 1/20$

$P_R[4] = 1/10 + \frac{1}{2} \cdot 0 = 1/10$

$P_R[5] = 1/20 + \frac{1}{2} \cdot \frac{1}{2} = 6/20$

- Compute the similarity of node ② wrt node ① & node ⑤
- Compute the PPR of G wrt S and then return the most similar node to S

	UB
$t_1 \rightarrow 1, 5, 6, 7, 8, 11$	1
$t_2 \rightarrow 4, 13, 15$	0,2
$t_3 \rightarrow 5, 6, 8, 9$	0,6
$t_4 \rightarrow 2, 3, 5, 6$	2

current $\theta = 3.7$

- what's next pivot?
- is its score computed?

sort

	UB
$t_1 \rightarrow 1, 5, 6 \ldots$	1
$t_4 \rightarrow 2, 3, 5 \ldots$	2

e.g. $3,6 > 3.5 = \theta$

$t_2 \rightarrow a, 13 \ldots$	0,2
$t_3 \rightarrow 5, 6, 8 \ldots$ (pivot)	0,6

5 is a potential candidate to enter in the heap, thus we compute its full score

---

$D = \{ \underset{1}{\text{abcd}}, \underset{2}{\text{mamo}}, \underset{3}{\text{atom}}, \underset{4}{\text{too}}, \underset{5}{\text{one}} \}$

1) build a 2-gram index

2) show the candidate dictionary strings for Q = mom with $e=1$

$\$$abcd

$\$$mama

$\$a \rightarrow 1$	
$ab \rightarrow 1$	
$bc \rightarrow 1$	
$cd \rightarrow 1$	
$\$m \rightarrow 2$	
$ma \rightarrow 2$	

$L - K \cdot e = 3 - 2 \cdot 1 = 1$

0m → 2,3

mo → 2

$e → 3

et → 3

t0 → 3

0m → 3,5

$2 → 4

20 → 4

00 → 4

$0 → 5

$mom

$m → 2

m0 → 2

0m → 3,5

$D = \{abe, baa, babe\}$
1 2 3

$Q = bab$

1 - error match candidates

$D_1 = D$

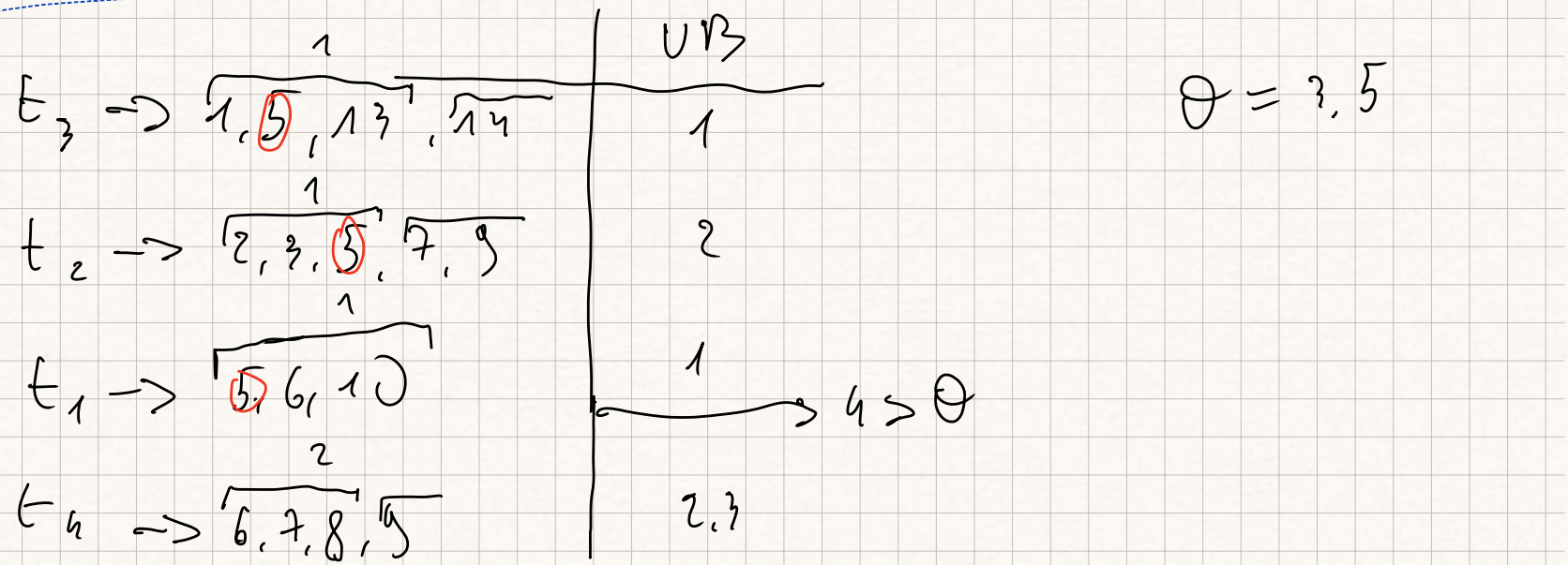
$D_2 = \{ba[1], ae[1], ab[1]$
$ae[2], ba[2]$
$abe[3], bbe[3], baa[3], bab[3]\}$

$Q = bab$

- Q in D_1 = exact
- Q^- in $D_1 = \{ab, bb, ba\}$ in D_1

- Q in $D_2^- = \{ab, bb, be\}$ in $D_2 = S_1, S_2 = aba, bee$

- Q in $D_2 = S_3 = baba$



5 is discarded because $3 < 3,5$

$t_2 \rightarrow$ 1, 5, 8, 18 | 1, 3

$t_1 \rightarrow$ 2, 5, 8, 16 | 1

$t_4 \rightarrow$ 5, 8, 10, 21 | 2

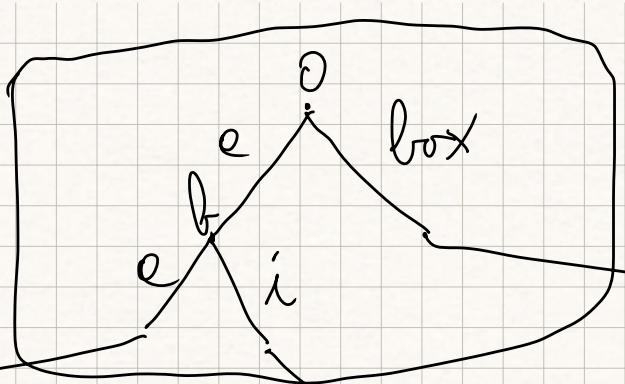
$t_3 \rightarrow \leq 16$

$\theta = 4$

D = {abe, abete, abort, abri, boss, bot, box, brun, cat}

1) Design a two level index which uses a block of 3 strings and Front coding on each block

stored alphabetically sorted



internal memory

disk: [<0, abe>, <3, te>, <3, rt>] [<0, abi>, <0, box>, <2, t>]

[<0, box>, <1, um>, <0, cat>]

2) Search for q = { acc, aa, abb, to }

∧ exact

acc, aa : 1 I/O → access to disk; no disk access

abb : access block three

3) Prefix search for Q = e : scan block 1 and 2

Web Graph

w=4

- 13 → 1, 2, 13, 14, 26
- 14 → 3, 10, 11, 13, 14, 17, 19, 21, 25
- 15 → 4, 8, 9, 32
- 16 → 5, 10, 11, 12, 14, 17, 19, 20, 21, 24, 33

compress 16 in the best way

id	ref	outd	copy blocks	extra nodes
14	--	--		
15	--	--		
16	2	11	0 1 1 0 1 1 1 1 0	5, 12, 20, 24, 33

≤ 9

# blocks	copy block
5	~~0, 1, 2, 4, 6, 1~~
	→ 0, 0, 1, 0, 3

$S = (1, 4, 6, 10, 17, 14)$ Elias-Fano

$w = 4$

$$\ell = \left\lceil \log_2 \frac{u}{n} \right\rceil = \left\lceil \log \frac{15}{6} \right\rceil = 2$$

1	1
1	00,01

4	0100
6	0110
10	1010
12	1100
14	1110
	k : l

$n = (\max S) + 1 = 15$

$L = \overbrace{0100101000010}^{l \cdot n}$

$H = 10110101010$

$\#1 = n$

$\#0 = 2^k$

now let's do the reverse

$L = 100|0|10|10|00$

$H = \underbrace{10}\,\underbrace{11}\,\underbrace{0}\,\underbrace{10}\,\underbrace{10}$ (00, 01, 10, 11)

	k=2	l=2
2	00	10
4	01	00

S ?

$k = 2$ count the #0 in H

$n = 5$

$l = \frac{|L|}{n} = 2$

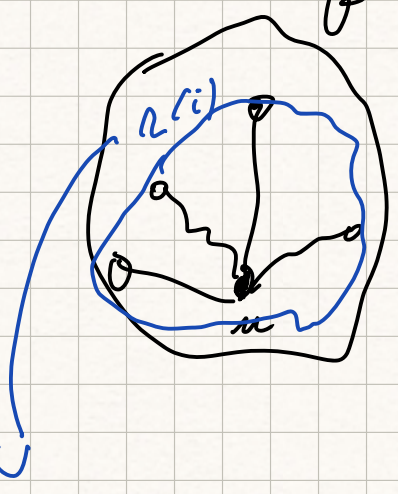
6	01	10
10	10	10
12	11	00

$$S = 2, 4, 6, 10, 12$$

$$r[i] = \alpha \sum_{j \to i} \frac{r[j]}{\#\text{out}\{j\}} + (1-\alpha)\frac{1}{N} \begin{cases} (1-\alpha)\frac{1}{|S|} & i \in S \\ 0 & \end{cases}$$

contribution depends on G's structure

teleportation step



$r[i]$ = similarity between nodes i and u

$\Downarrow$

volume of paths btw $u \rightsquigarrow i$

Personalized / topic-based

S

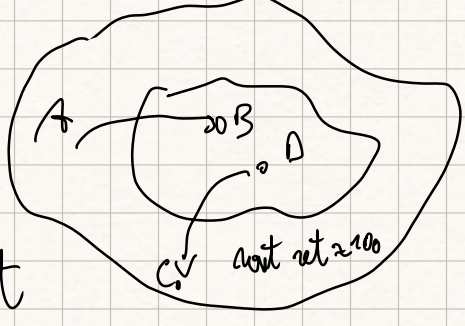
$|S|=1$

$(1-\alpha).1$ for u

0 otherwise

$H(+)$ (query dependent)

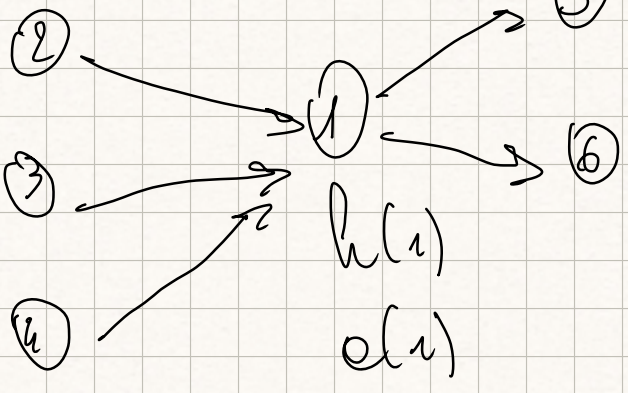
- Compute "root set" = list of pages answering Q
- expand to base set $\Rightarrow$ root set
- Random walk over the base set



base set web

authority score / hubness score \

- good authority is a page pointed by good hub pages
- "hub pages is a page that points to good authority pages



$h(1) = a(5) + a(6)$ (w(1,5), w(1,6))

$a(1) = h(2) + h(3) + h(4)$

$\vec{h}$ = vector of all hubness score

$\vec{a}$ = // // // authority score

$$\begin{cases} \vec{a} = A^T \cdot \vec{h} = A^T \cdot A \cdot \vec{a} \\ \vec{h} = A \cdot \vec{a} = A \cdot A^T \cdot \vec{h} \end{cases}$$

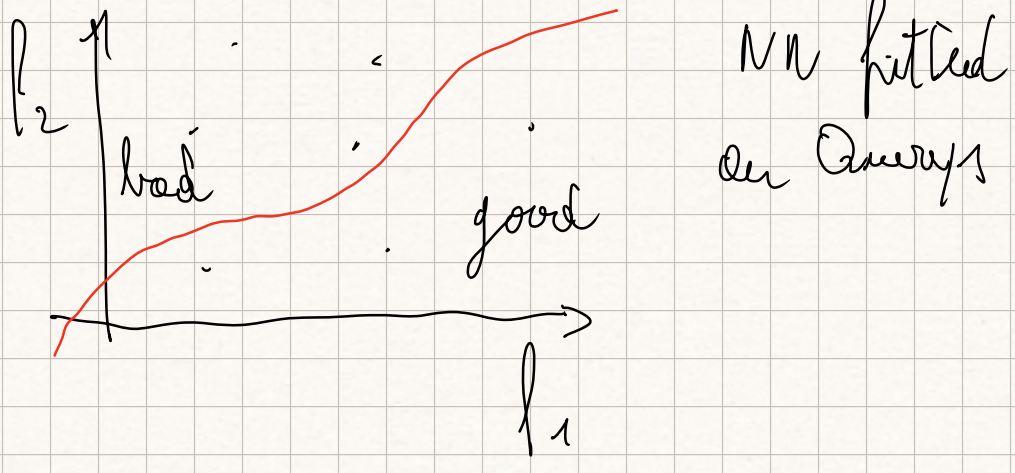
A: adjacency matrix of the bare set

eigenvalue

$$\lambda \cdot e = (A^T \cdot A) \cdot e \quad \text{eigenvector}$$

$$1 \cdot h = (A \cdot A^T)\, h$$

- TF
- Page Rank (HITS) / ...
- presence Q's terms in URL
- " " in titles of pages
- " " in the first lines
- Terms of Q are close to each other

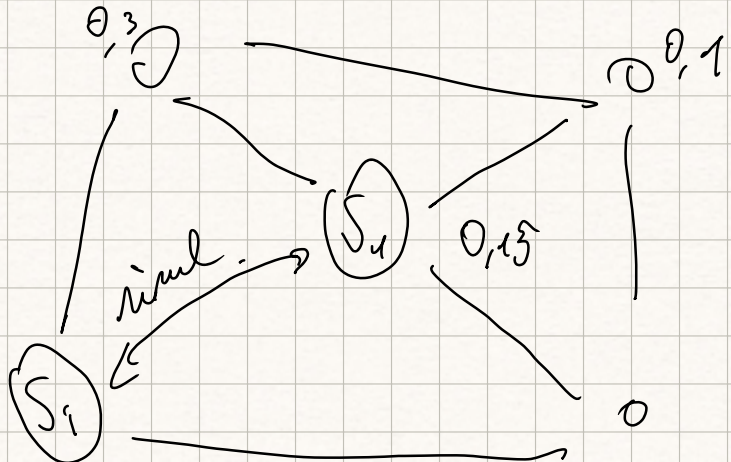


S_i = sentence

$$\text{relevancy}(S_i) = \sum_{w \in S_i} \frac{TF\text{-}IDF(w)}{|S_i|}$$

Text rank

$$Sim(S_i, S_j) = \frac{(S_i \cap S_j)}{\log|S_i| + \log|S_j|}$$



① $Sim(S_i, S_j)$ = cosine sim

$(\vec{s}_i, \vec{s}_j)$

tf-idf

② Pruning

01/12/22

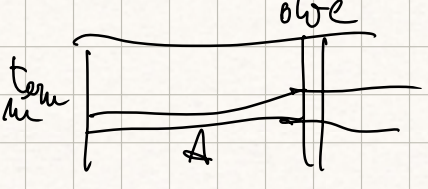
Vector Space model m dim space → k dim space

- Latent Semantic Indexing (LSI) - data dependent

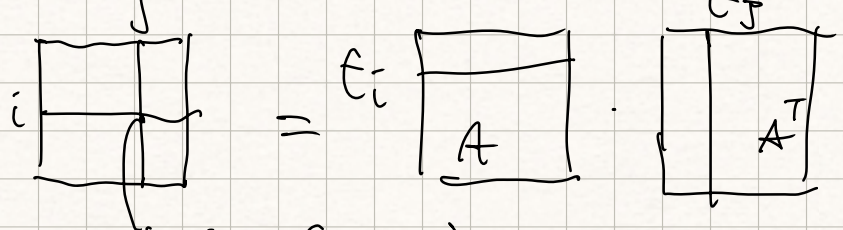
- Random Projections - data independent (with high prob.)

preserve the distances btw pairs of vectors in the reduced space

Singular-value decomposition

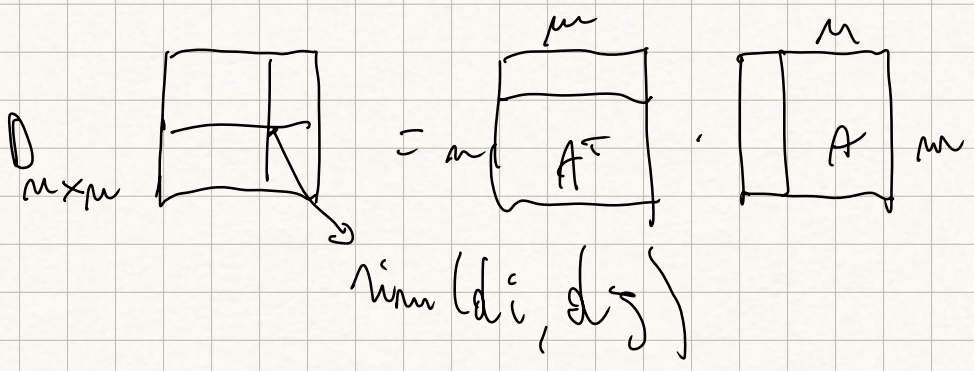
$A_{m \times n}$ 

· define $T = A \cdot A^T$ term-term matrix

$T_{m \times m}$ 

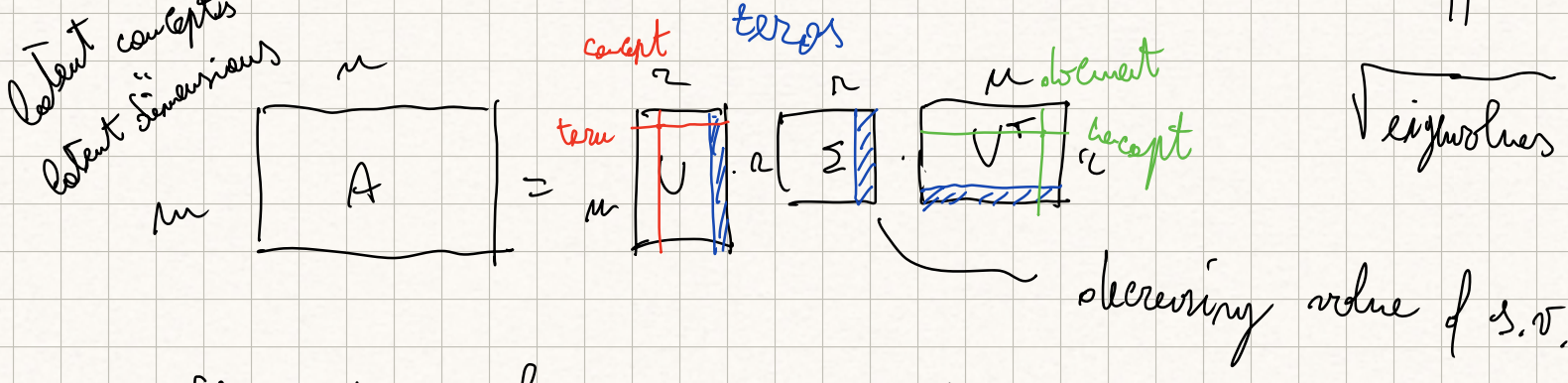
$= \dim(t_i, t_j)$

- define $D = A^T \cdot A$ doc-doc matrix



$A = U \cdot \Sigma \cdot V^T$ U, V orthogonal, Σ diagonal → singular values

m x r r x r r x m



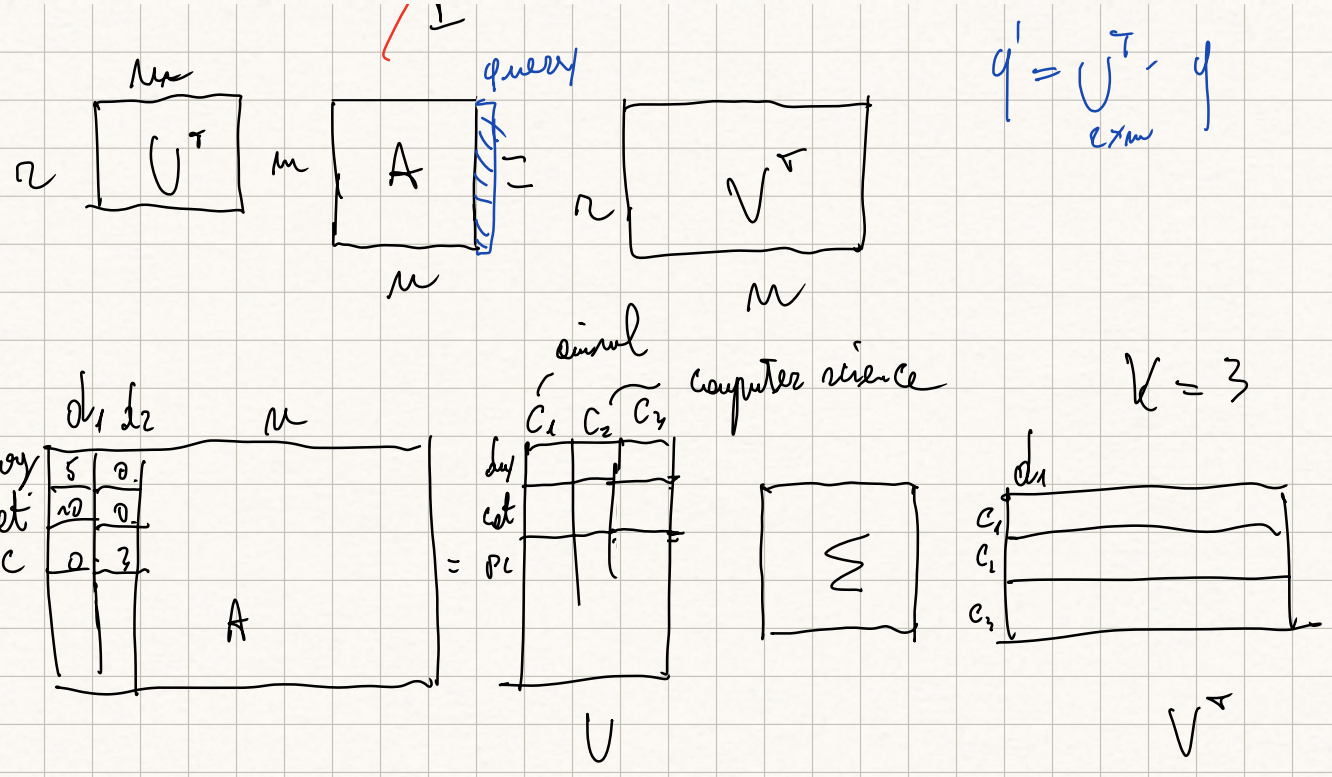
U = matrix of r eigenvalues of T

V = " " " D

per confrontare due doc. guardo le colonne di V^T

$$A = U \cdot \Sigma \cdot V^T$$

$$U^T \cdot A = U^T U \cdot \Sigma \cdot V^T$$


query
U^T
A
V^T
$q' = U^T \cdot q$
$r \times m$

dimensional computer science
$k = 3$
Es:
d_1 d_2
C_1 C_2 C_3
diary
cat
PC
A
Σ
U
V^T