INFORMATION RETRIEVAL

Week 9 – Vector Space Model

02.05.2025

Severin Mills

Today

1

Exercise Recap

- Discussion
- Questions

2

Theory

- Recap: Ranked Retrieval
- Vector Space Model
- SMART Notation

3

Kahoot

- BONUS: Vector Space Model

Recap

Heaps' Law: $\#terms = k \times \#tokens$

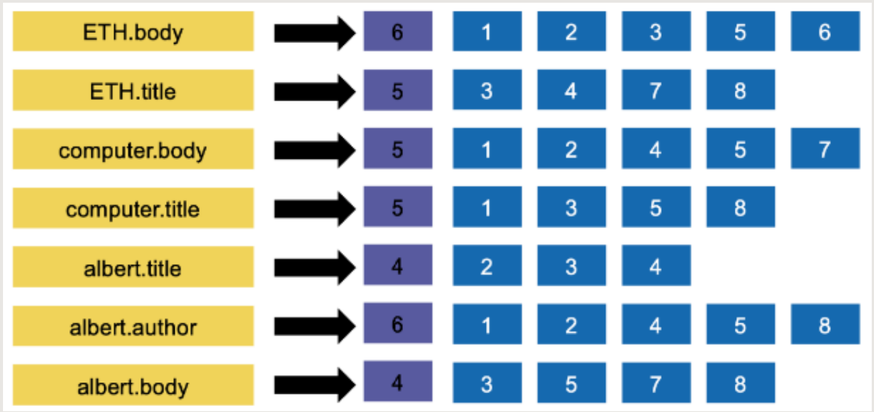
Zipf's Law: $Frequency = \frac{k}{Rank}$

Previously

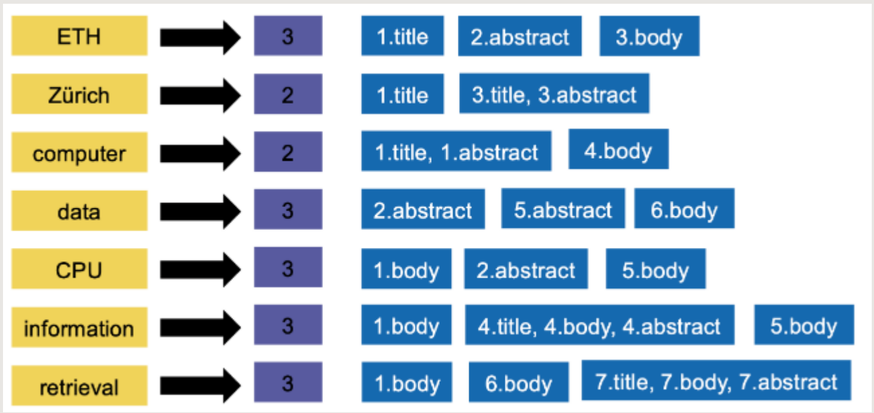
- We want to rank documents based on a score
- Measure similarity between documents

Scoring

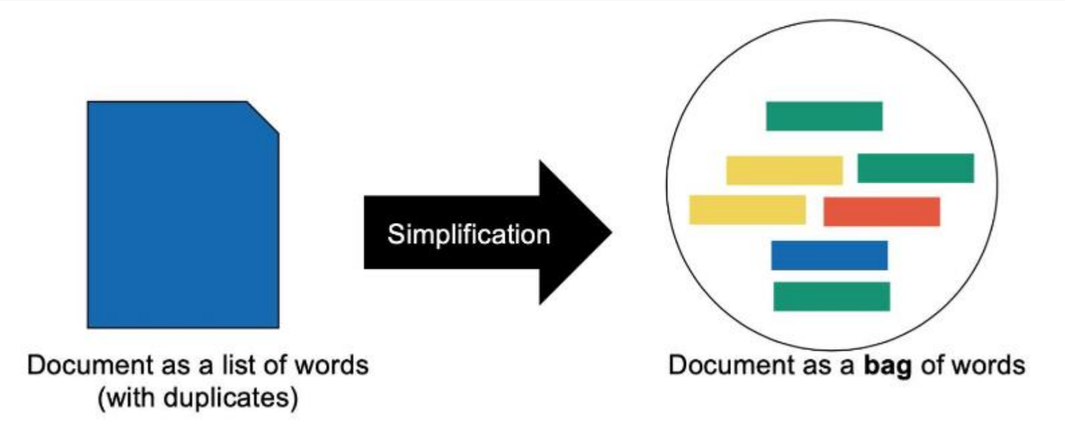
Zones in terms



Zones in postings



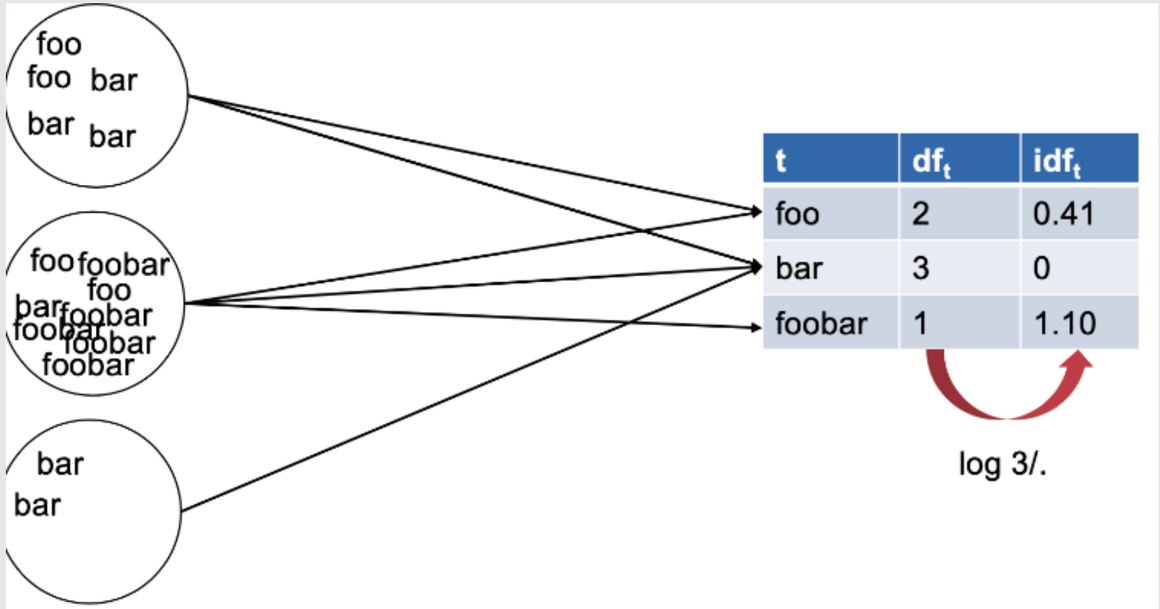
Scoring



t	$tf_{t,d}$
foo	3
bar	2
foobar	2

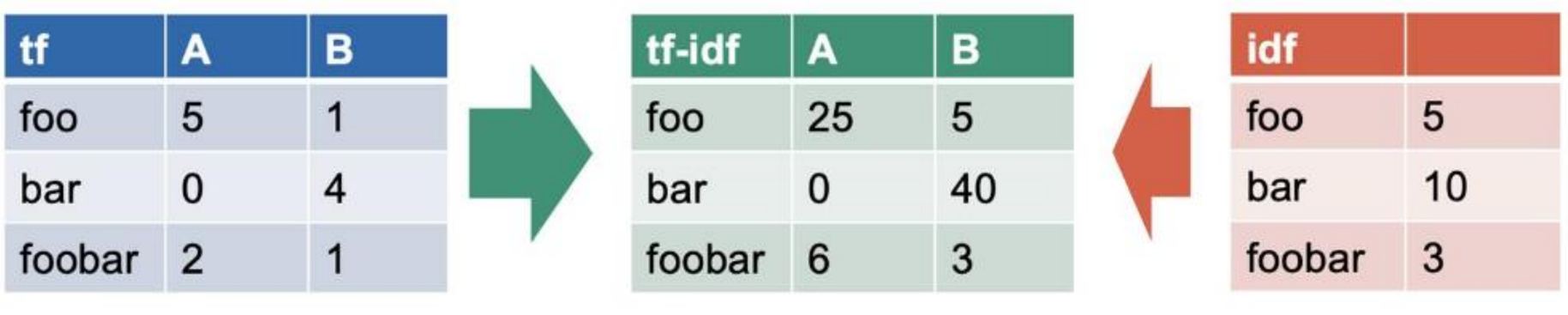
Scoring

If we want to measure rarity, we use the inverse document frequency



Putting it all together

tf-idf



tf	A	B
foo	5	1
bar	0	4
foobar	2	1

tf-idf	A	B
foo	25	5
bar	0	40
foobar	6	3

idf	
foo	5
bar	10
foobar	3

Now

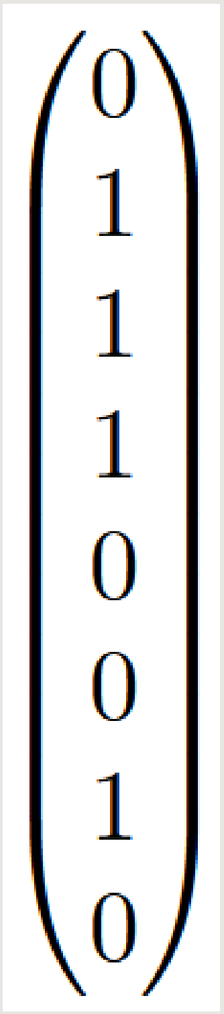
Main idea: represent documents as vectors

$$\begin{pmatrix} 1.5 \\ 1 \\ 3 \\ 5 \\ 0 \\ 0 \\ 10.4 \\ 0 \end{pmatrix}$$

Vectorization

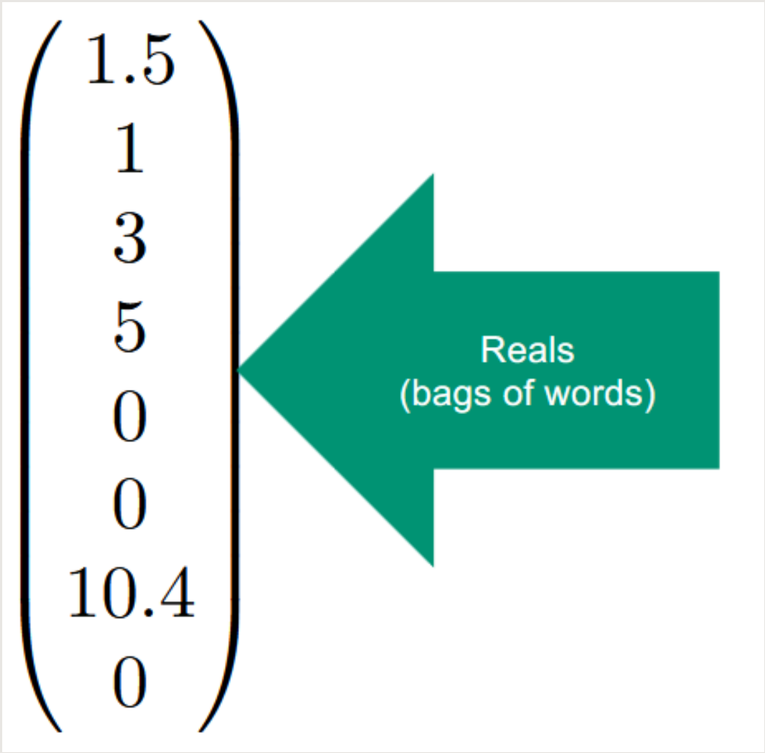
Boolean vector

- Not very usable for ranking, why?



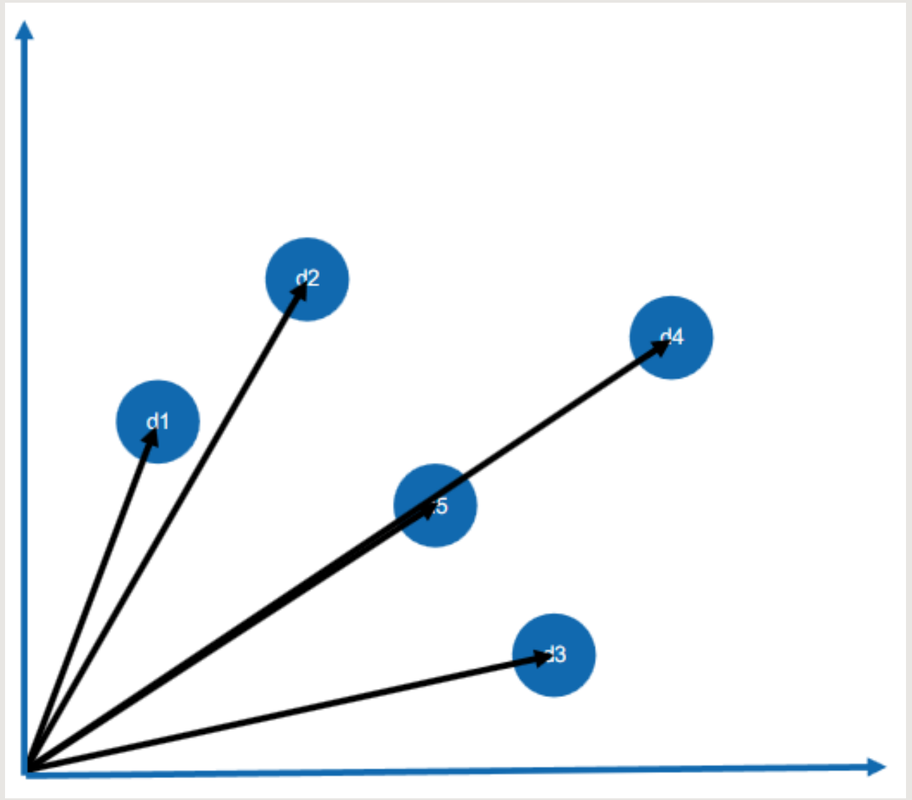
Vectorization

Use reals instead.
Could be term frequency, tf-idf (preferably), etc.



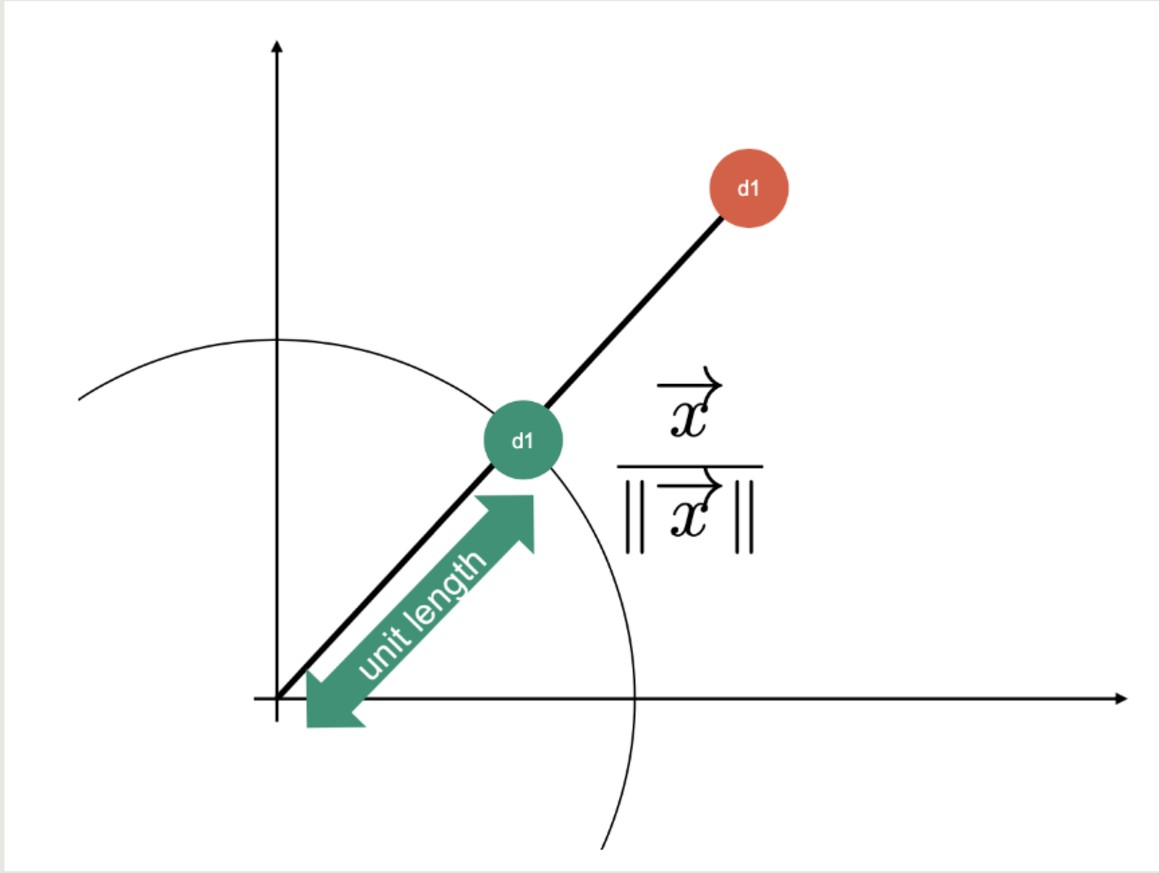
Vectorization

A document is then a vector in the first quadrant, $D \in \mathbb{R}^M$



Vectorization

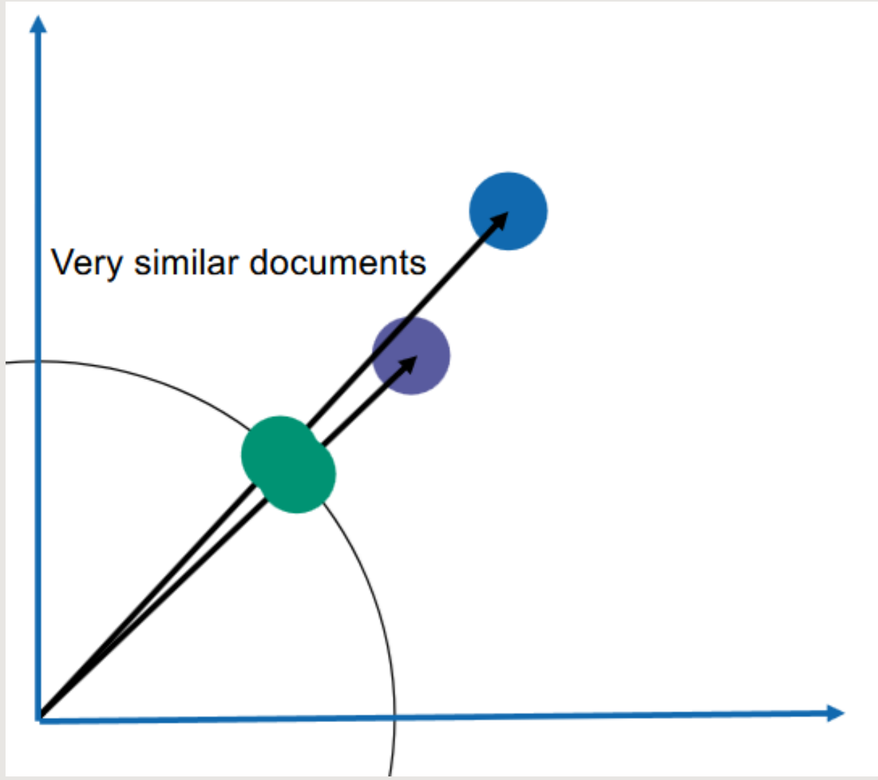
Normalize vectors



Inner Product (Renormalized)

$$\cos\theta = \frac{\vec{x}}{\|\vec{x}\|} \times \frac{\vec{y}}{\|\vec{y}\|}$$

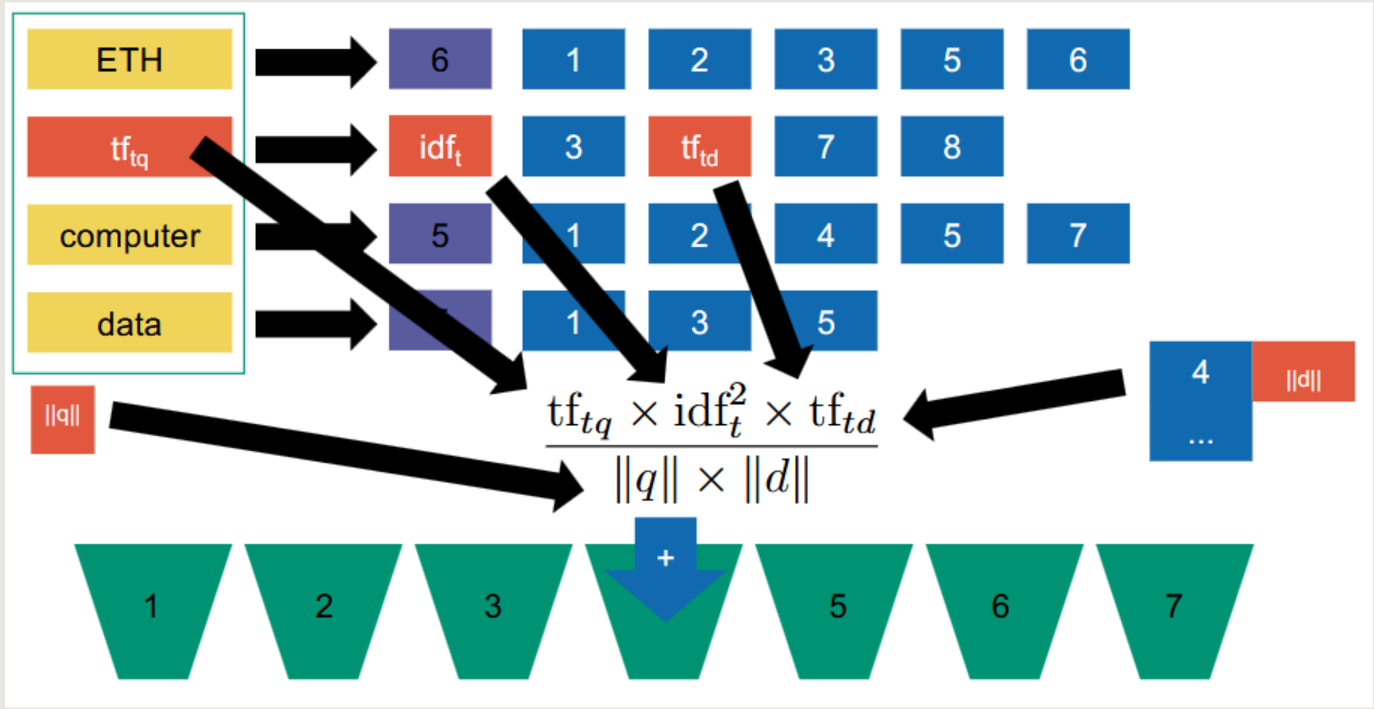
- Good quantification of similarity



Queries as vectors

- Build vector
- Calculate cosine similarity of documents and return the nearest ones
- What about a large collection of documents?
- Doesn’t matter, since query size is constant

Inverted index



SMART Notation

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log\frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0,\log\frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$ (Section 6.4.4)
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

BONUS TIME

- Same as previous
- Starts on 02.05 at 15:00, ends on 09.05 at 15:00
- If you already passed 2 / 3 quizzes, you got the bonus (you should get a mail soon)

Kahoot

https://create.kahoot.it/details/duplicate-of-information-retrieval-ex-07-vector-space-models-mschoeb/ef383953-b43a-4abd-af2a-d9ebf2ad1019

02.05.2025