Univerza v Ljubljani
Fakulteta za računalništvo
in informatiko

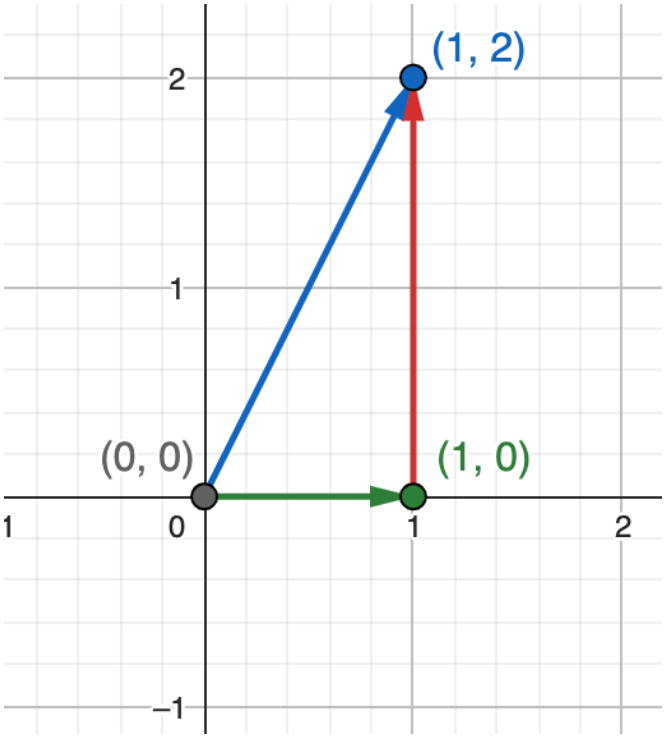
Predmet:
Osnove podatkovnih baz

Modul:
Vektorske podatkovne baze

Gradivo:
v.2025

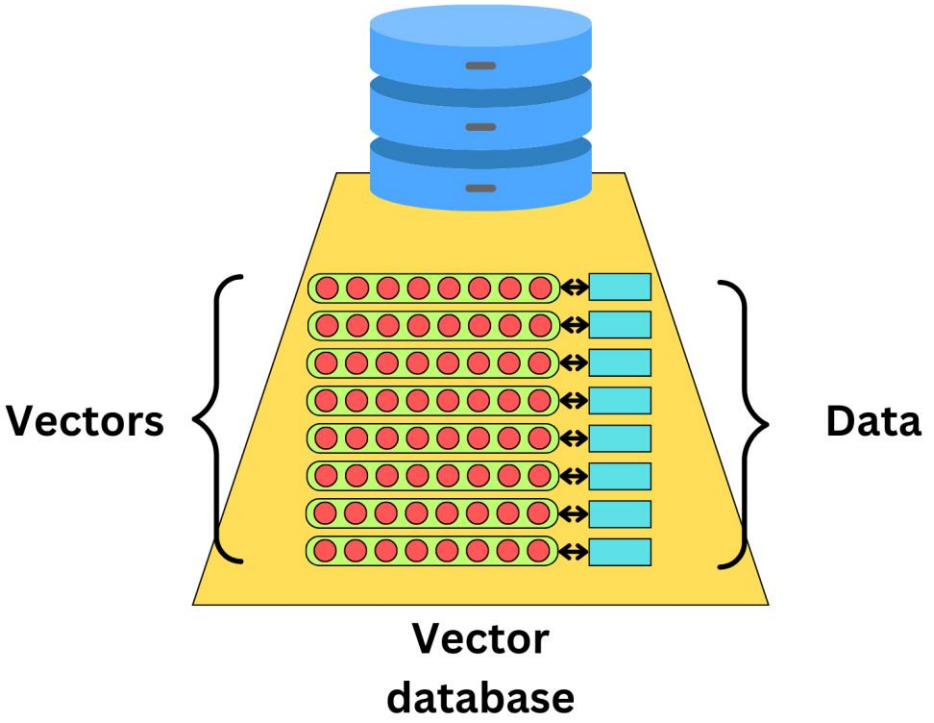
Vektorji

- Matematični objekti, ki imajo velikost in smer
- Predstavimo lahko s končnim zaporedjem števil
- Vizualiziramo lahko kot usmerjene črte



Vektorske podatkovne baze - VPB

- VPB je posebna vrsta PB, ki podatke hrani kot več-dimenzijske vektorje.
- Vektorji v VPB so matematične predstavitve lastnosti podatkov.
- Več lastnosti, kot jih lahko razberemo iz podatkov, več dimenzij ima vektor (tudi tisoč in mnogo več).



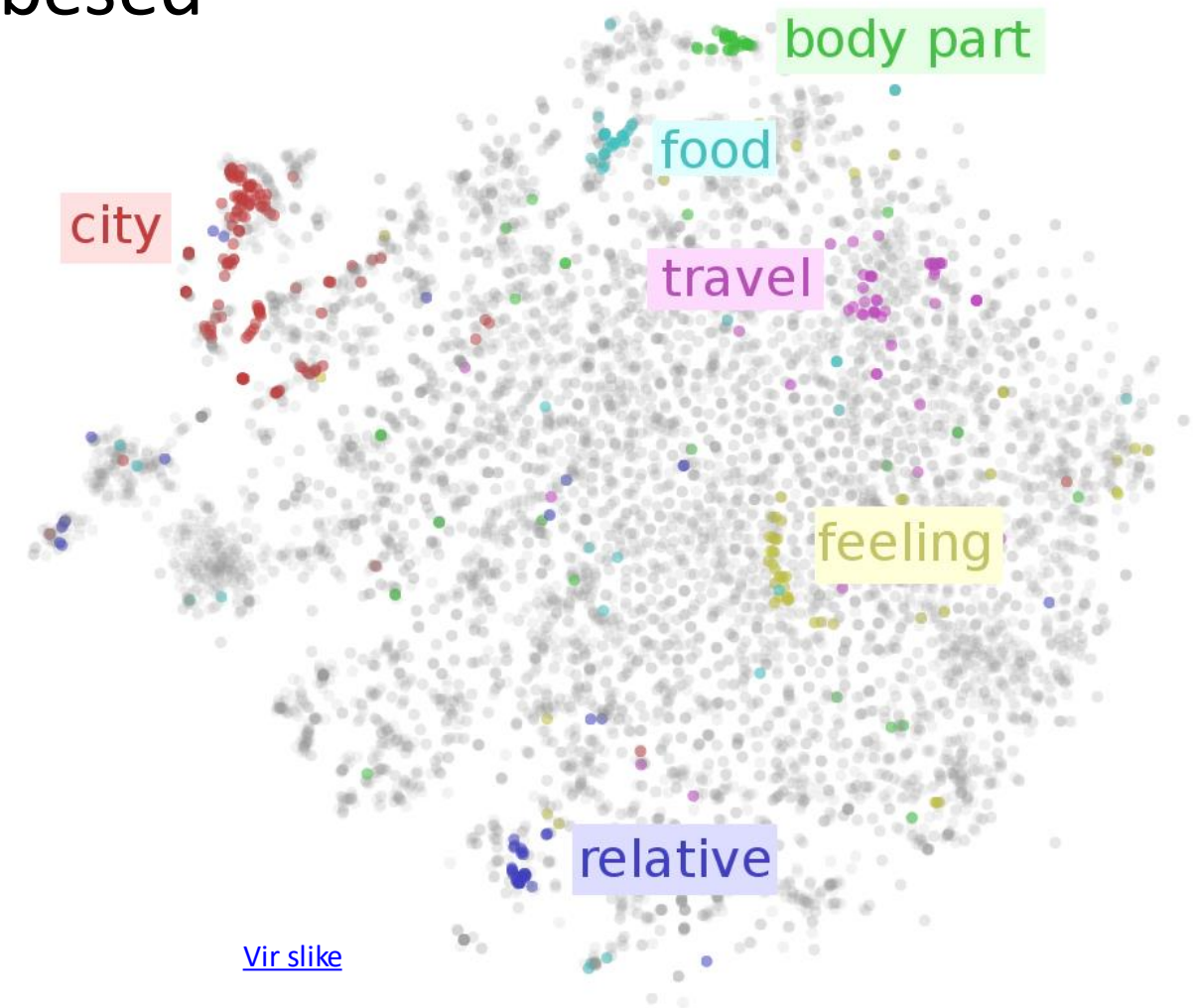
Vir slike

Vektorske podatkovne baze - VPB

- VPB se uporabljajo za hranjenje kompleksnih vektorjev, ki nastanejo s strojno obdelavo podatkov in iz podatkov izluščijo številne (tudi skrite) lastnosti.
- Različne metode za generiranje vektorskih predstavitev:
  - modeli strojnega učenja,
  - vložitve besed (*word embeddings*) in
  - algoritmi za ekstrakcijo lastnosti (feature extraction algorithms).

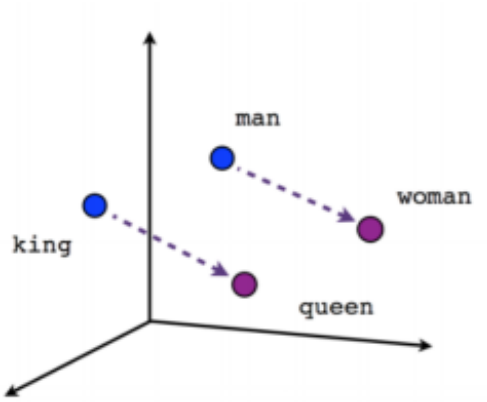
- Cilj: semantično podobni vhodni podatki se pretvorijo v vektorje, ki so si blizu v več-dimenzionalnem prostoru.

Vložitve besed

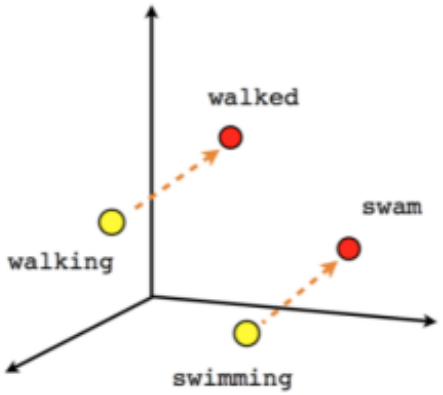


Vir slike

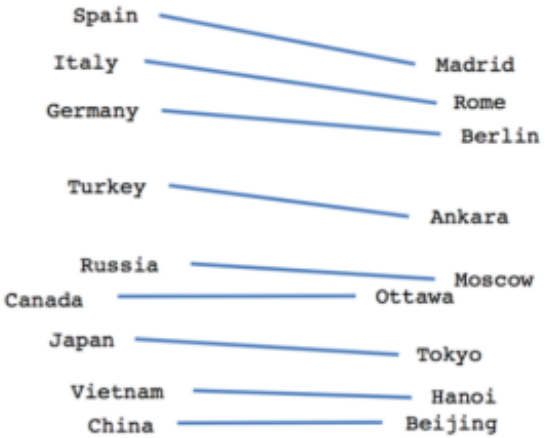
Lastnosti vložitev



Male-Female



Verb tense

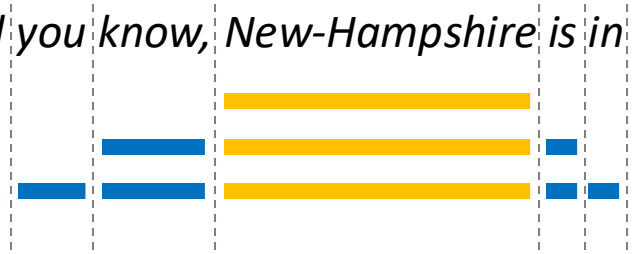


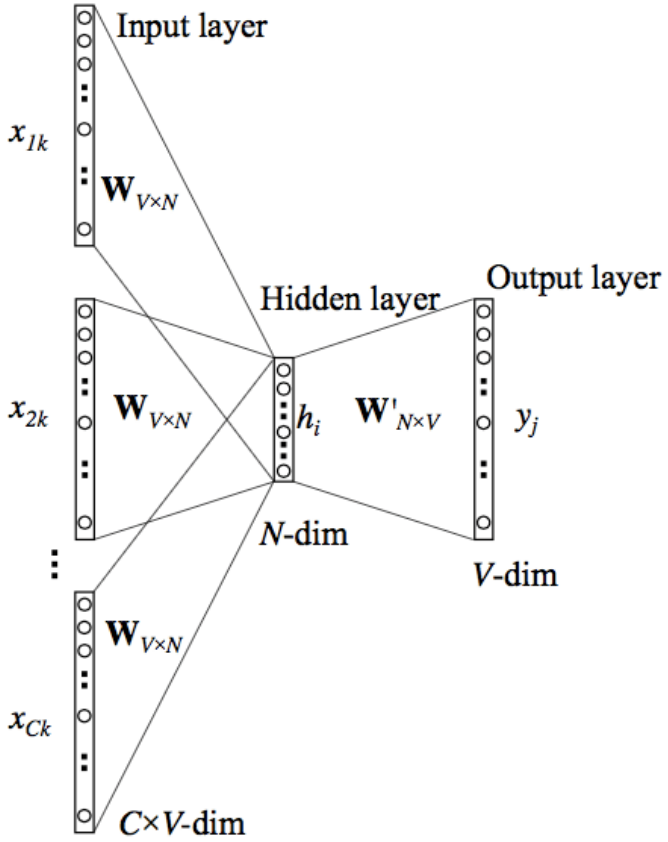
Country-Capital

Vir slike

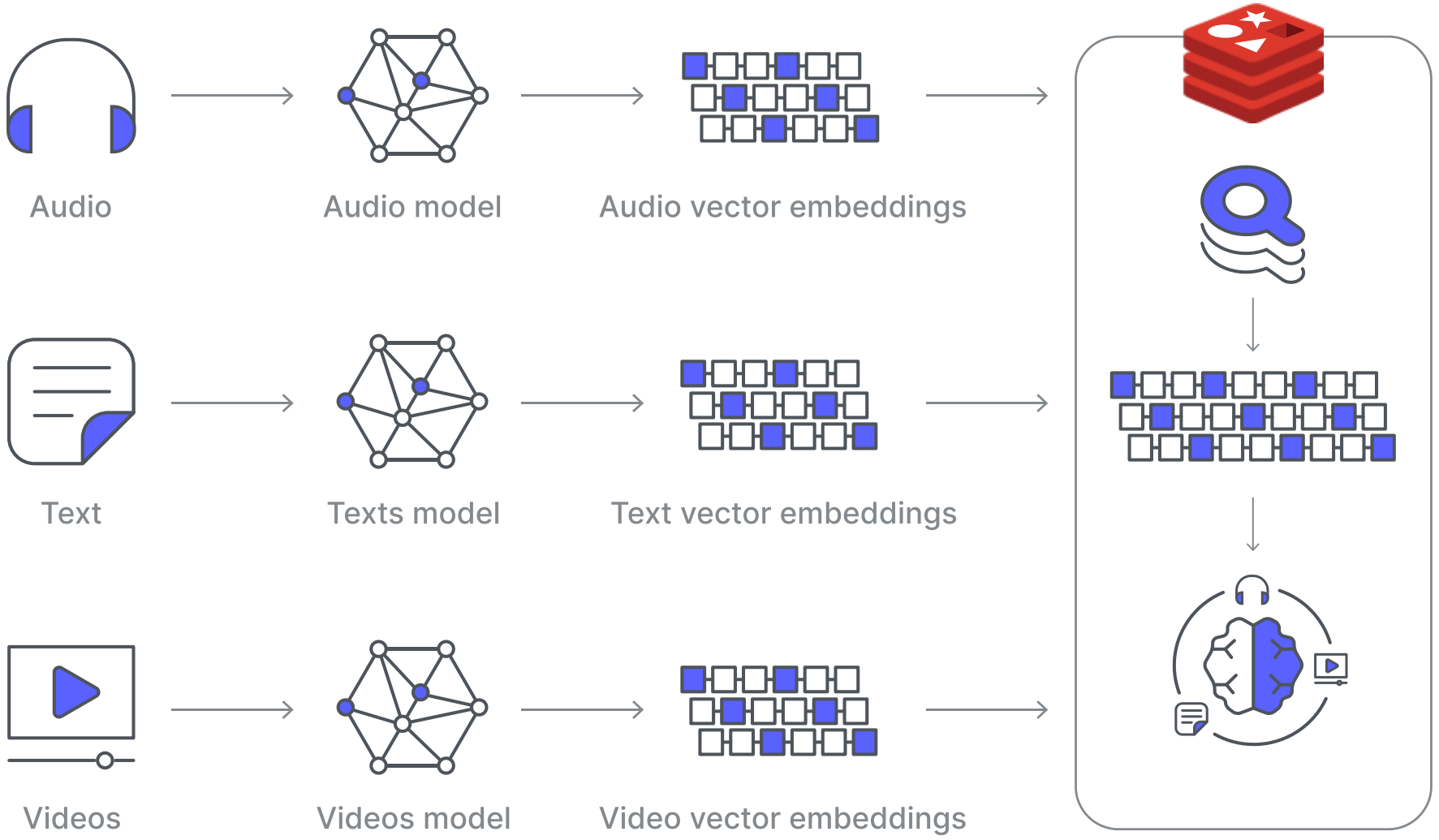
Pridobivanje vektorskih vložitev

Did you know, New-Hampshire is in New-England.





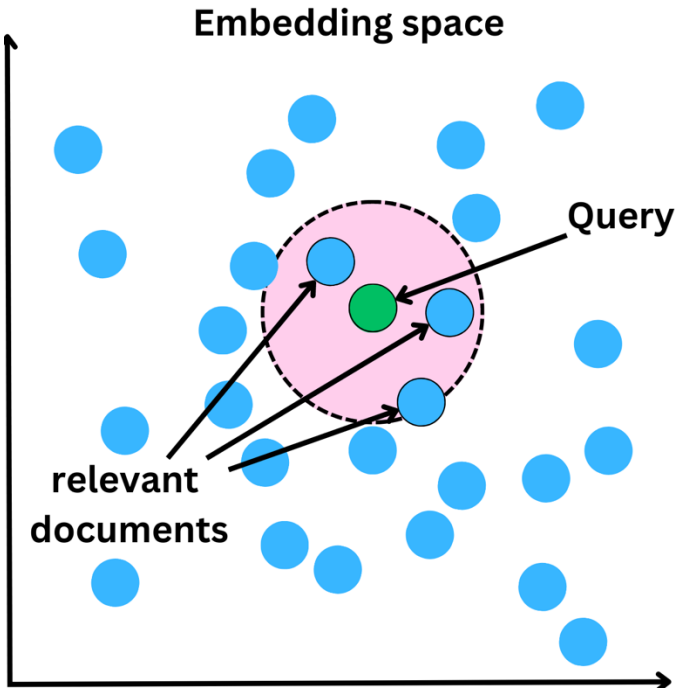
Vir slike


Audio
Audio model
Audio vector embeddings
Text
Texts model
Text vector embeddings
Videos
Videos model
Video vector embeddings

Vir slike

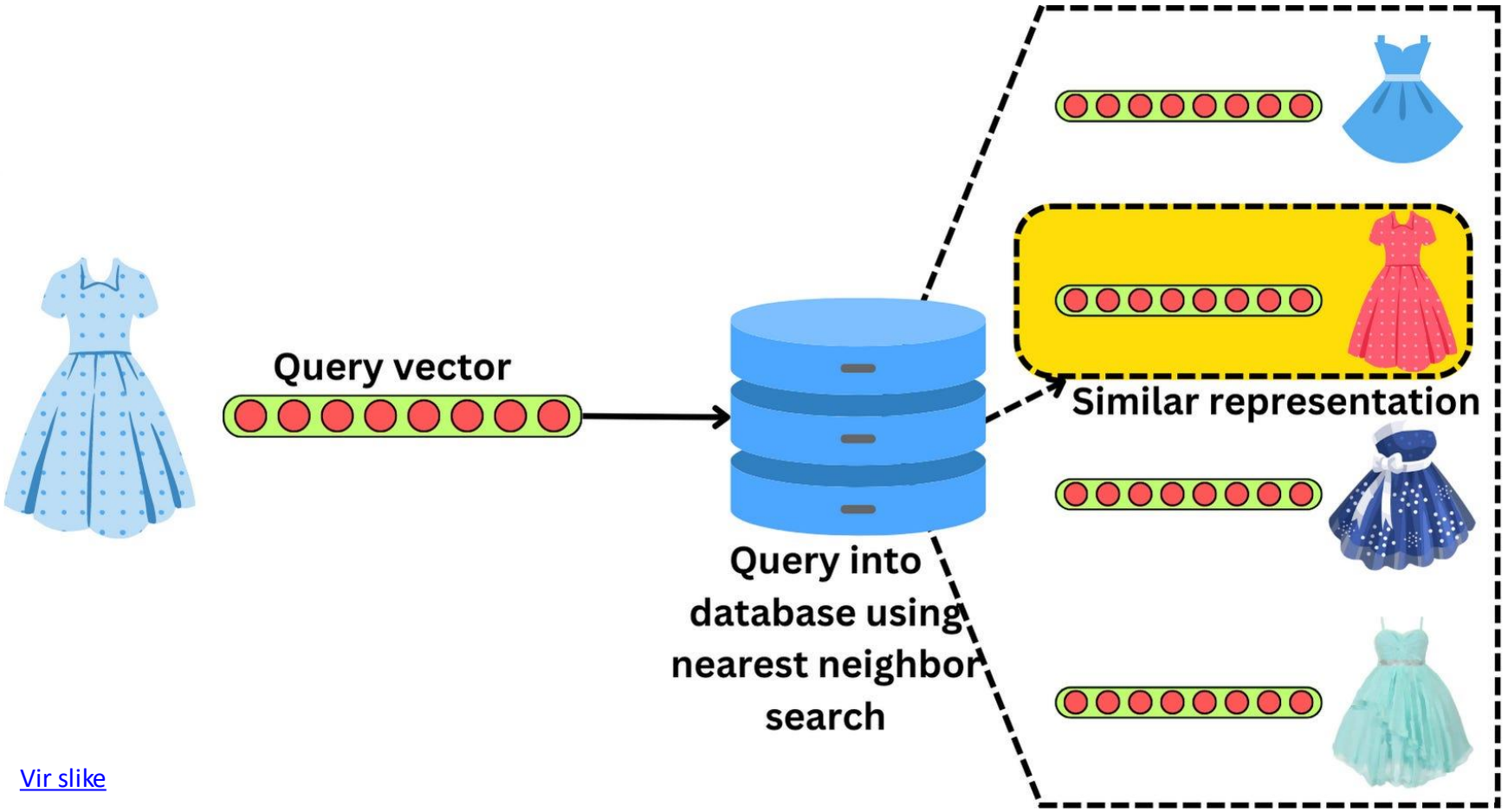
Prednosti VPB

- Učinkovito iskanje podobnih podatkov (bližina v vektorskem prostoru).
- Iskanje na podlagi semantične relevantnosti namesto zanašanja na natančno ujemanje.
- VPB so optimizirane za analitične poizvedbe; gručenje, kategorizacija podatkov, …



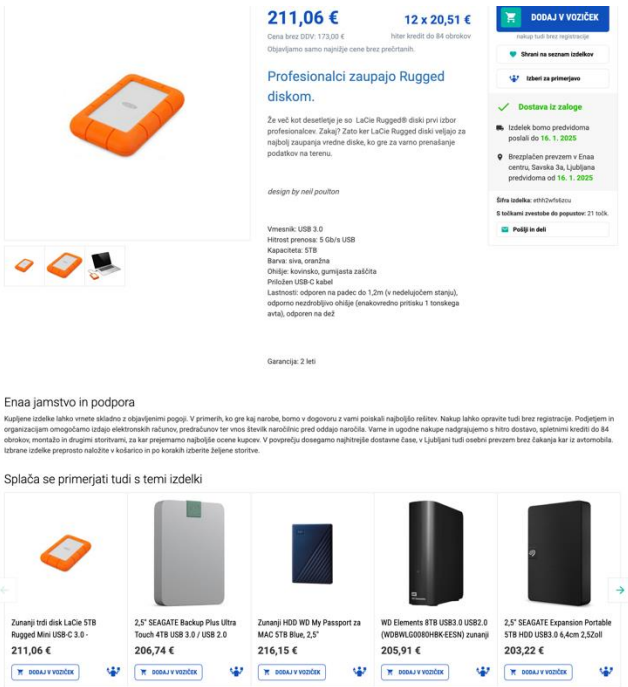
Vir slike

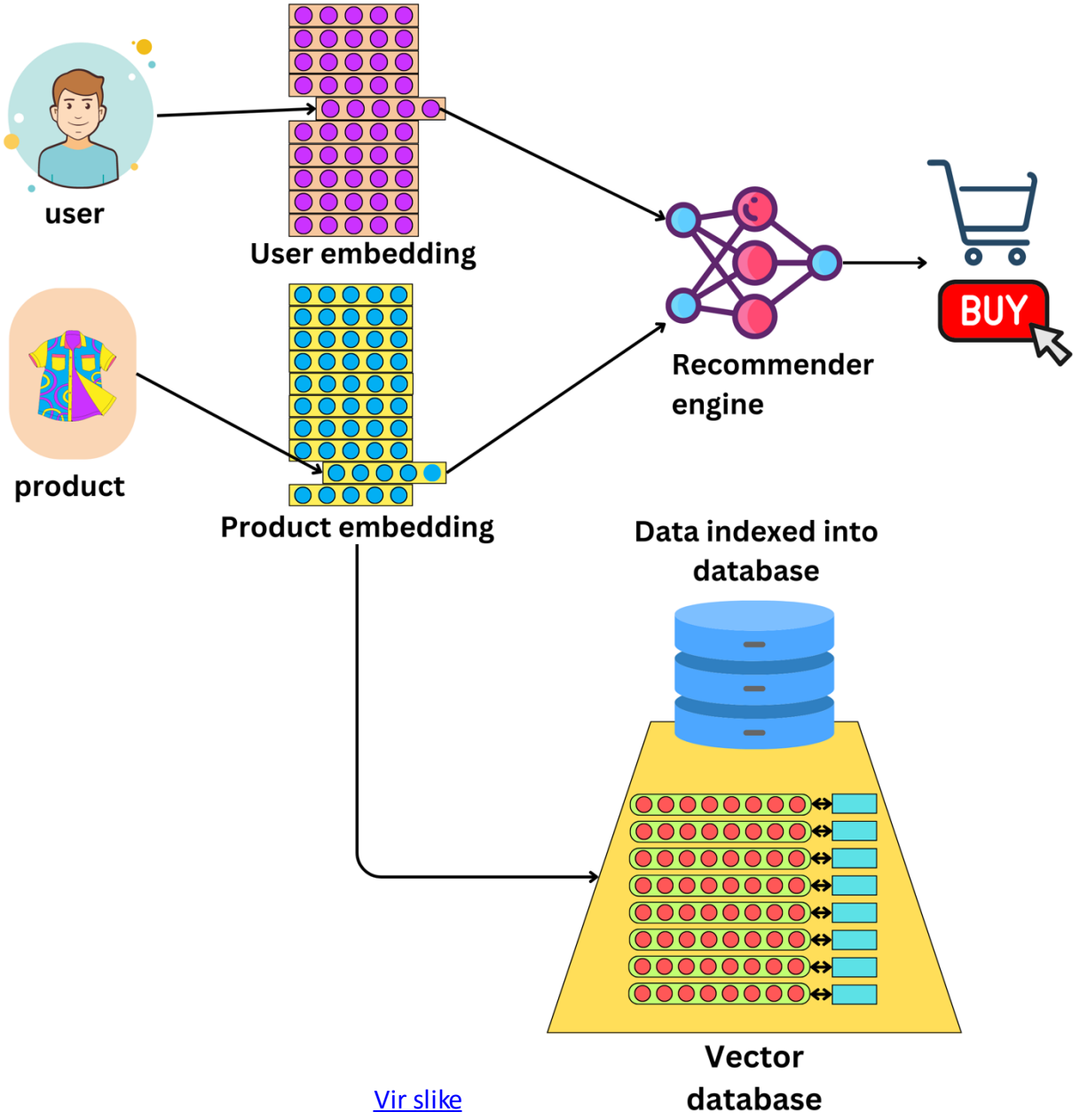
Iskanje po podobnosti



Vir slike

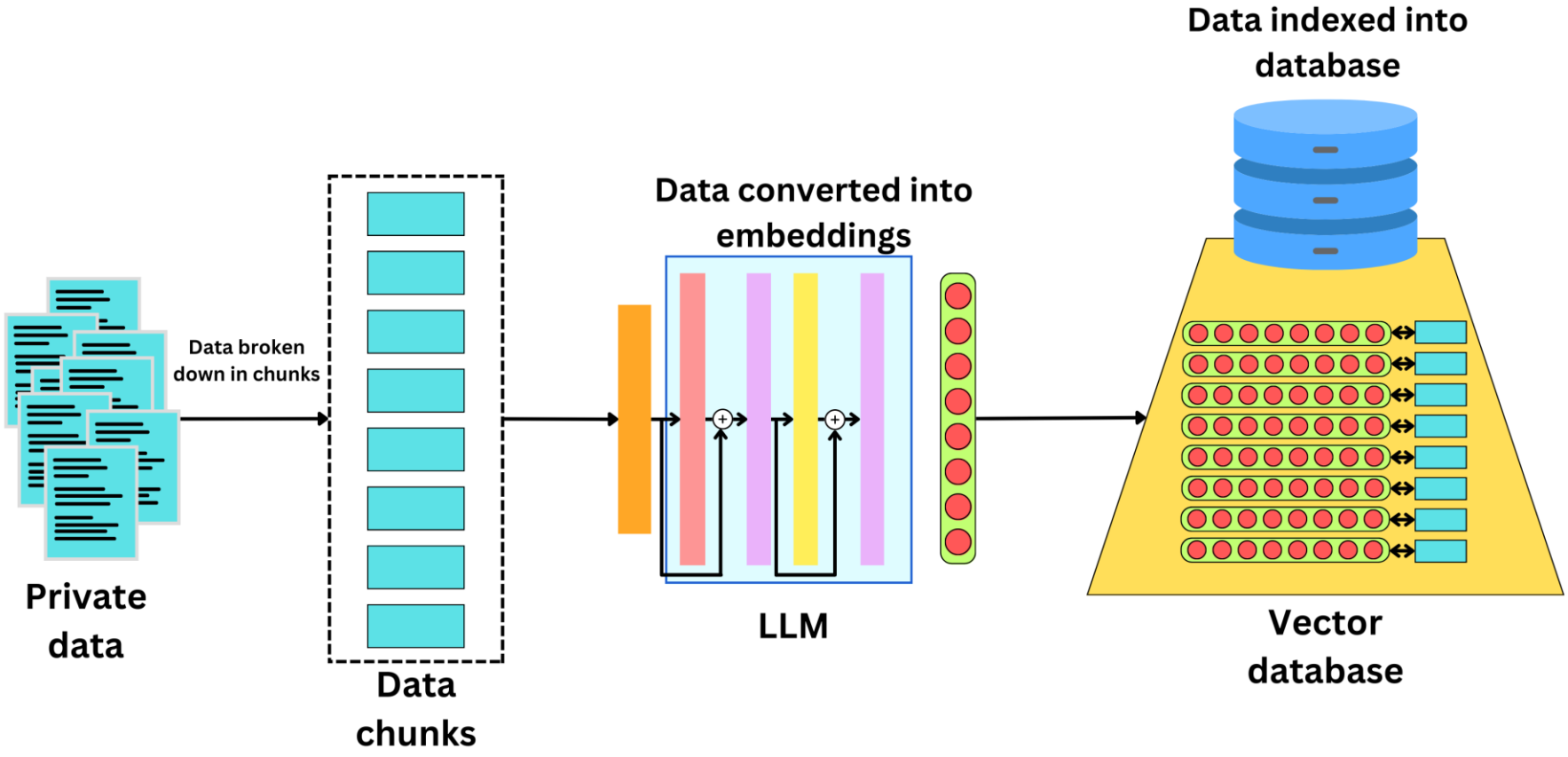
Priporočilni sistemi, e-poslovanje





Vir slike

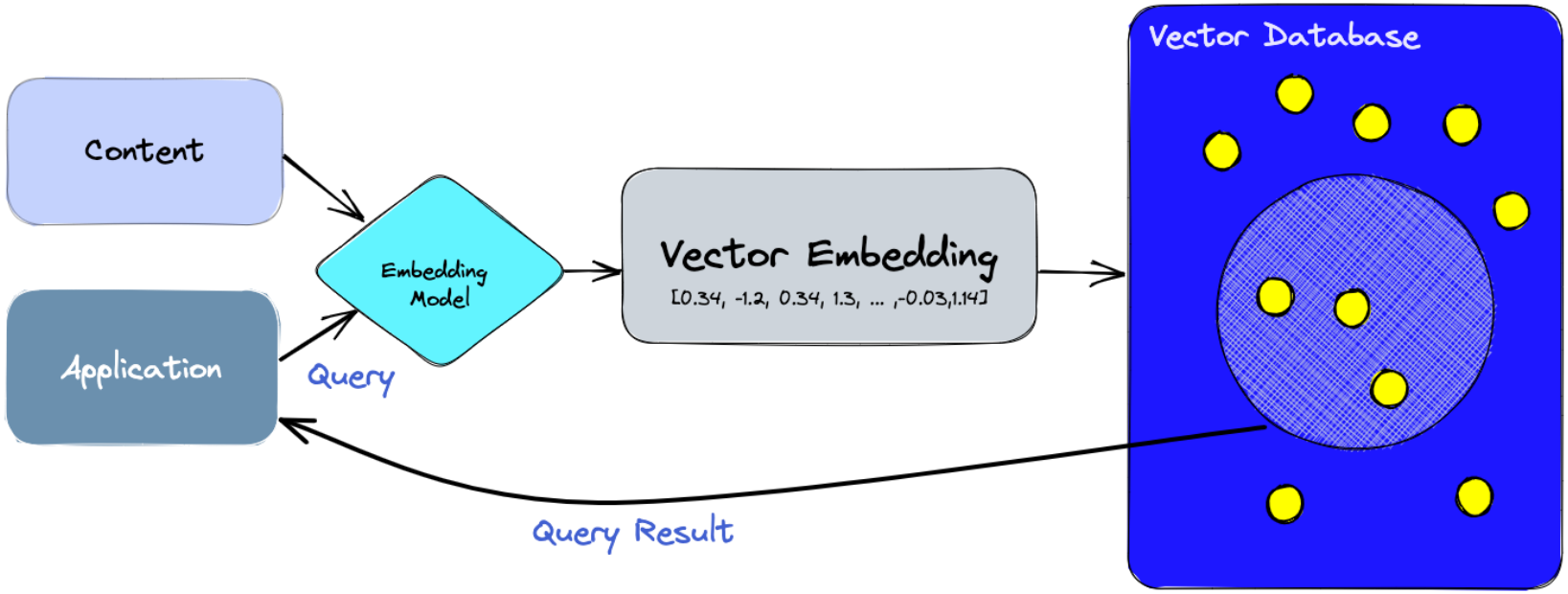
LLM in RAG



Vir slike

Delovanje VPB

- Polnjenje:
  - vektorje generiramo z uporabo ene izmed možnih funkcij za izračun vektorjev
  - izračunane vektorje zapišemo v VPB z referencami na izvirne podatke
- Poizvedovanje:
  - Poizvedbo pretvorimo v vektor (z isto funkcijo).
  - Z uporabo algoritmov v VPB poiščemo najbližje vektorje glede na izbrane metrike.


Content
Application
Embedding Model
Query
Vector Embedding
[0.34, -1.2, 0.34, 1.3, ... ,-0.03,1.14]
Vector Database
Query Result

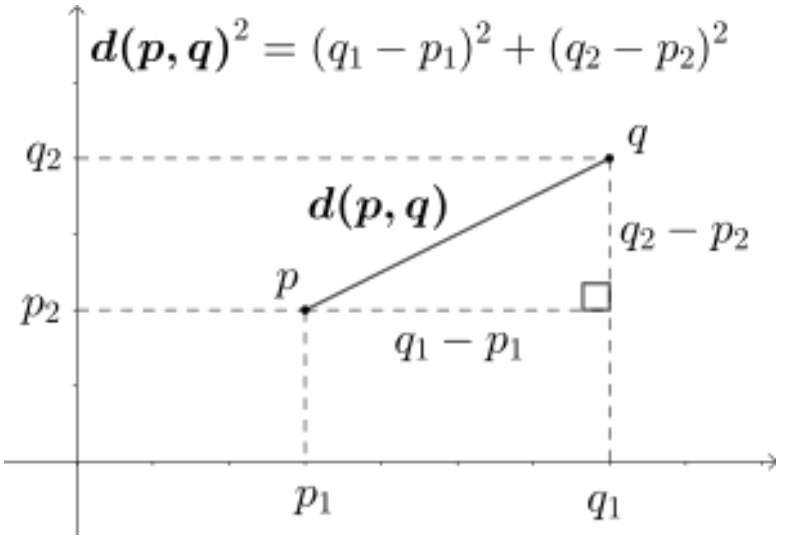
Vir slike

Metrike razdalje

- Matematične funkcije, ki določajo "razdaljo" med dvema vektorjema v vektorskem prostoru.
- Različne metrike razdalje zajamejo različne vidike podobnosti (izbira metrike ključna za specifične aplikacije).
- Najbolj priljubljene metrike razdalje so:
  - Evklidska razdalja
  - Manhattanska razdalja
  - Jaccardova podobnost
  - Skalarni produkt
  - Kosinusna razdalja

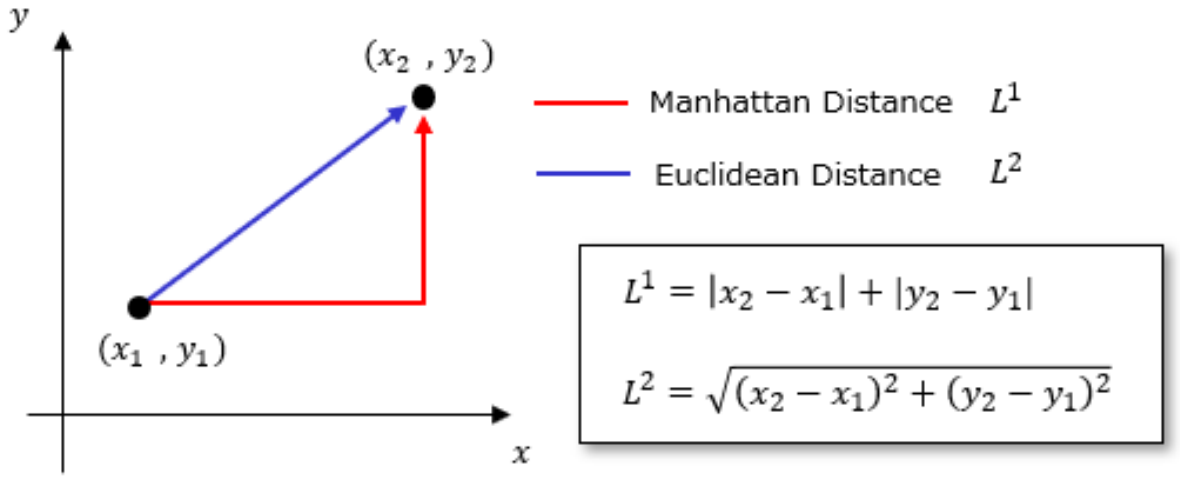
Evklidska razdalja

- Najpogostejša metrika razdalje, imenovana tudi L2 norma.
- Meri razdaljo med dvema točkama v vektorskem prostoru.
- Zelo občutljiva na velikost vektorjev.



Manhattanska razdalja

- L1 norma ali taksimetrična geometrija.
- Sešteje absolutne razlike med istoležnimi koordinatami vektorjev.
- Uporabimo, kadar želimo poudariti, kako se razlikuje vsaka značilnost, ne le kako različne so v celoti.
- Pogosto se uporablja za iskanje slik in finančno analizo.



$$L^1 = |x_2 - x_1| + |y_2 - y_1|$$

$$L^2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

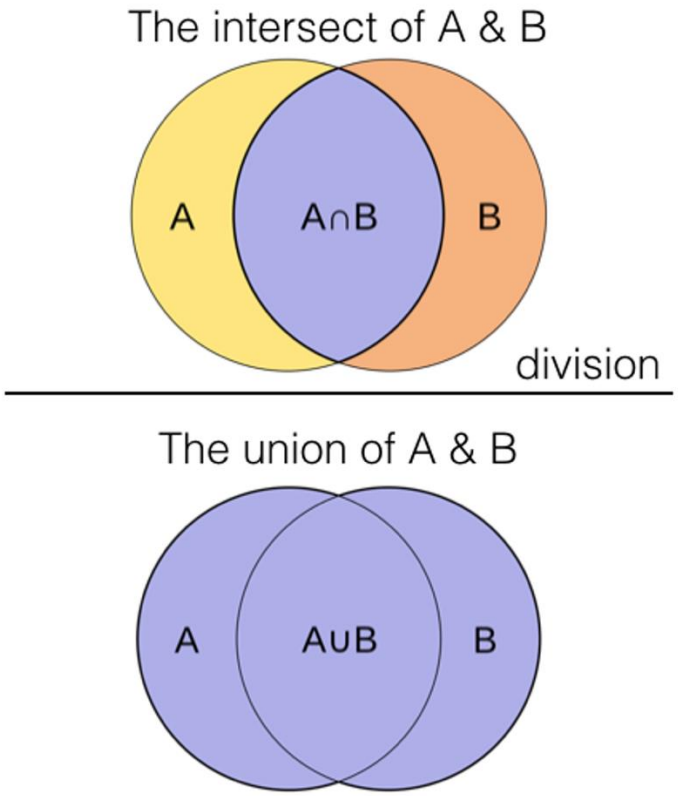
Vir

Jaccardova podobnost

- Jaccardova podobnost je podobnost med dvema množicama:

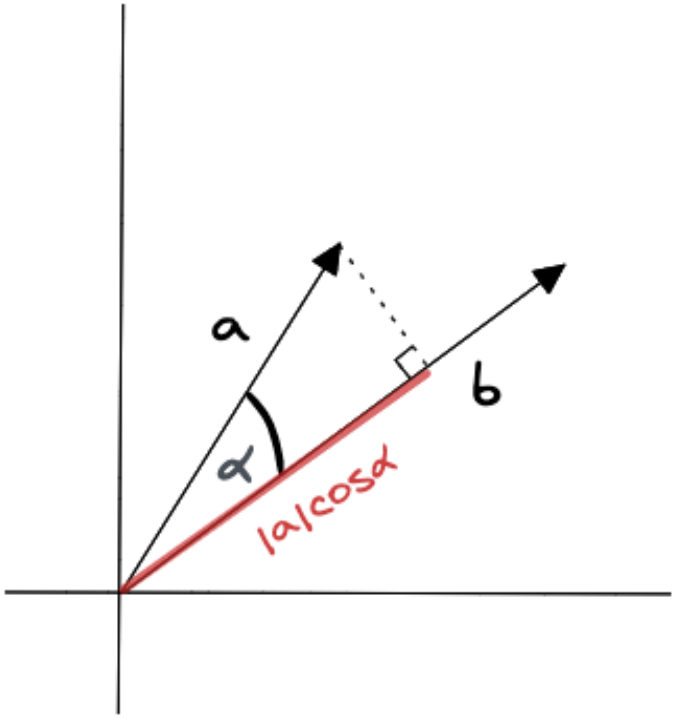
$$J(A,B) = \frac{A \cap B}{A \cup B}$$

- Primeri uporabe:
  - Gručenje in klasifikacija
  - Informacijsko poizvedovanje, npr. izračun podobnosti med dvema besediloma glede na besedišče, ki ga uporabljata.



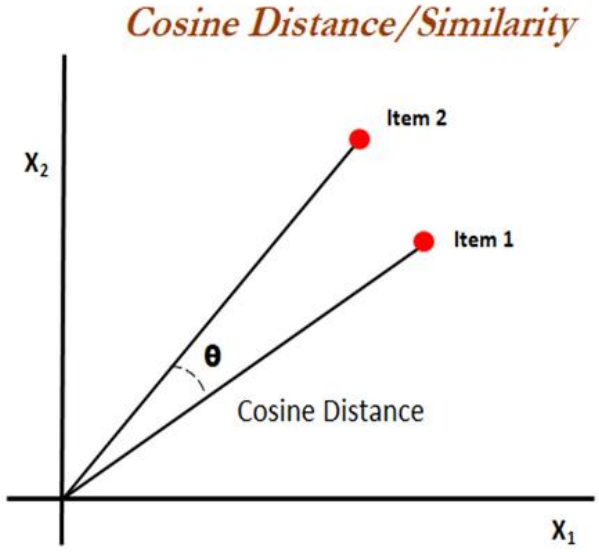
Skalarni produkt

- Meri produkt velikosti dveh vektorjev in kosinus kota med njima.
- Obseg je od -∞ do ∞, kjer pozitivna vrednost predstavlja vektorje, ki kažejo v isto smer, 0 predstavlja ortogonalne vektorje, negativna vrednost pa predstavlja vektorje, ki kažejo v nasprotnih smereh.
- Dolžina vektorjev ključna.



Kosinusna razdalja

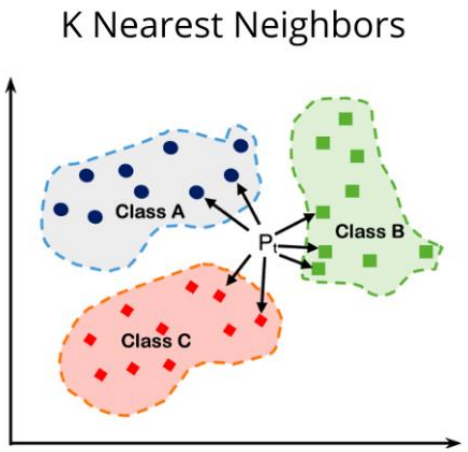
- Meri kot med dvema vektorjema – bolj kot sta vektorja narazen, manj podobne podatke predstavljata.
- Uporabna za iskanje podobnosti med besedili.
- Dolžina vektorjev nepomembna.



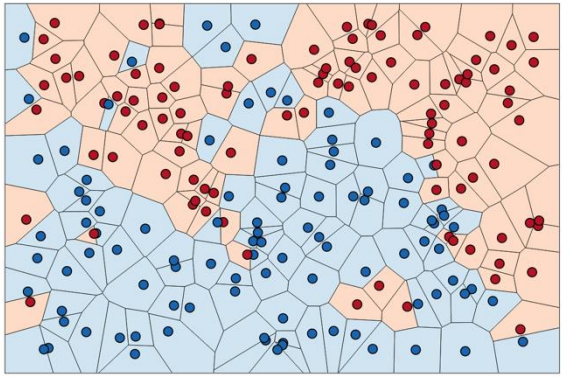
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Iskanje podobnosti

- VPB imajo implementirane optimizacijske algoritme za hitro iskanje, npr.:
  - Iskanje podobnih vektorjev
  - KNN – k najbližjih sosedov
  - ANN – približno najbližji sosed



Vir slike



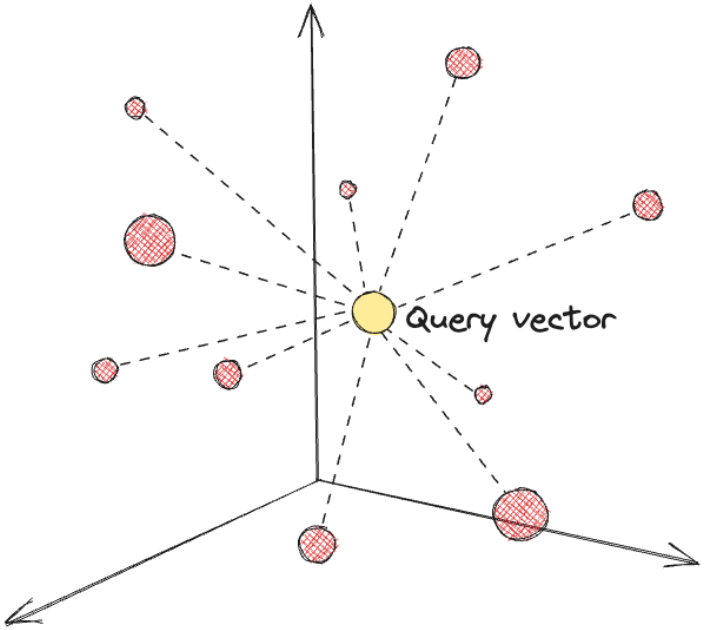
Vir slike

Indeksiranje

- Omogočajo hitro zoženje iskalnega prostora in pospešitev pridobivanja vektorjev.
  - Ravni indeks (*Flat index*)
  - Lokalno občutljivo zgoščevanje (*Locally Sensitive Hahsing - LSH*)
  - Približni najbližji sosed Oh Yeah (*ANNOY - Aproximate Nearest Neighbor Oh Yeah*)
  - …

Ravni indeks

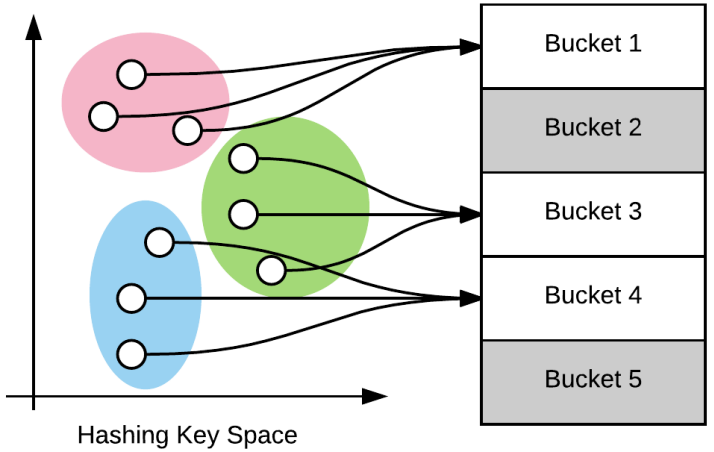
- Ravno indeksiranje (*Flat indexing*) – vektor hranimo v originalni obliki.
- Omogočajo pridobivanje natančnih rezultatov (natančno ujemanje)
- Hiter, natančen, majhna poraba prostora
- Primeren za majhne baze
  <10 000 vektorjev.
- Slabost: linearna časovna komp.



Vir slike

Lokalno občutljivo zgoščevanje (LSH)

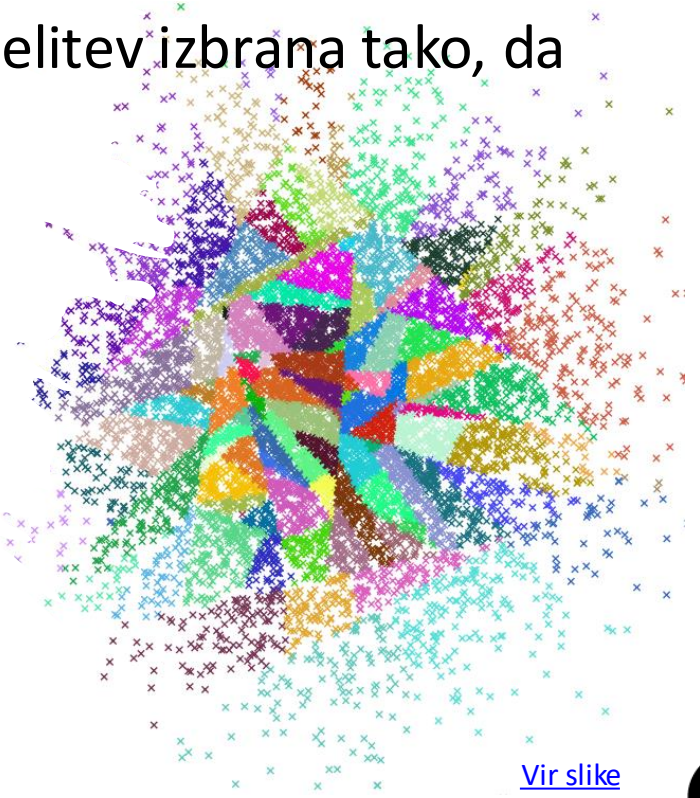
- Namenjen iskanju po visoko dimenzijskem prostoru.
- LSH podobne vektorje z visoko verjetnostjo vstavi v iste skupine (lokalno občutljive hash funkcije).
- Iskanje podobnosti zgolj znotraj istega ali sosednjih skupin



- Omogoča iskanje približno najbližjega soseda (ANN).

ANNOY – Appr. Nearest Neighbors Oh Yeah

- Knjižnica, zgrajena pri Sppotify-u
- Zgradi več dreves, tako da podatkovni nabor rekurzivno razdeli vzdolž naključno izbranih osi. Vsaka razdelitev izbrana tako, da čim bolj uravnoteži drevo.
- Ko so drevesa zgrajena, jih ANNOY indeksira za učinkovito poizvedovanje.
- Iskanje najbližjih sosedov temelji na pregledovanju dreves namesto vsakega posameznega vektorja.



Vir slike

Popularne vektorske baze