

Univerza v Ljubljani
Fakulteta za računalništvo
in informatiko



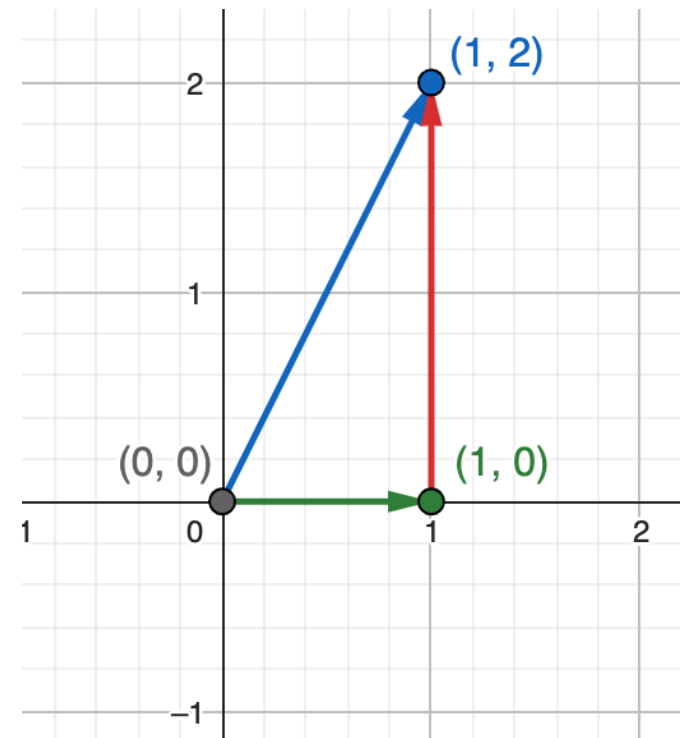
Predmet: Osnove podatkovnih baz

Modul:
Vektorske podatkovne baze

Gradivo:
v.2025

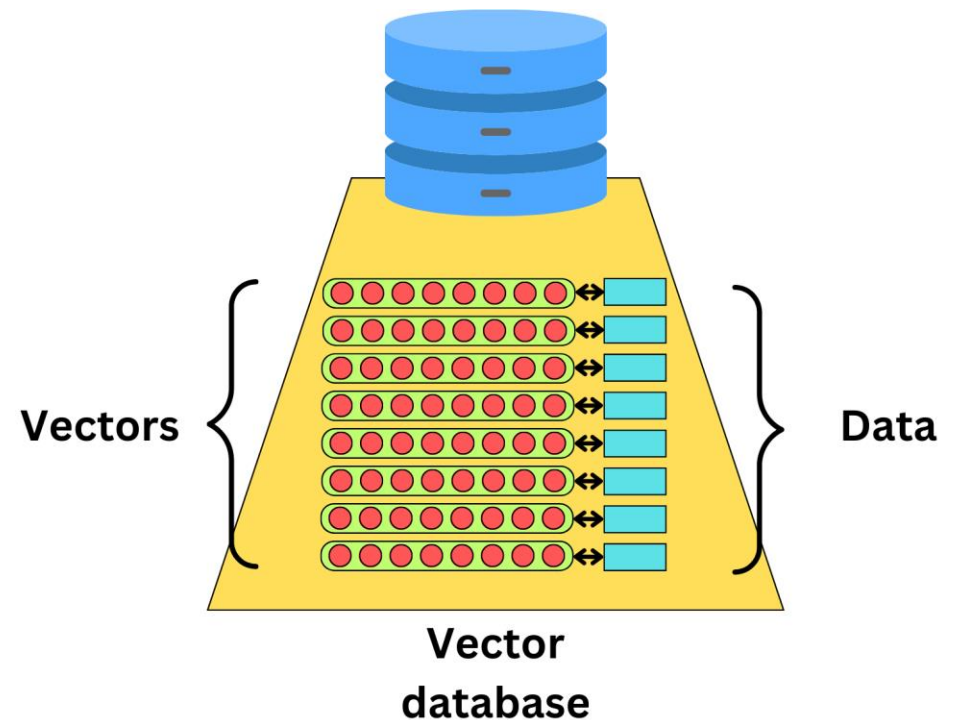
Vektorji

- 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).

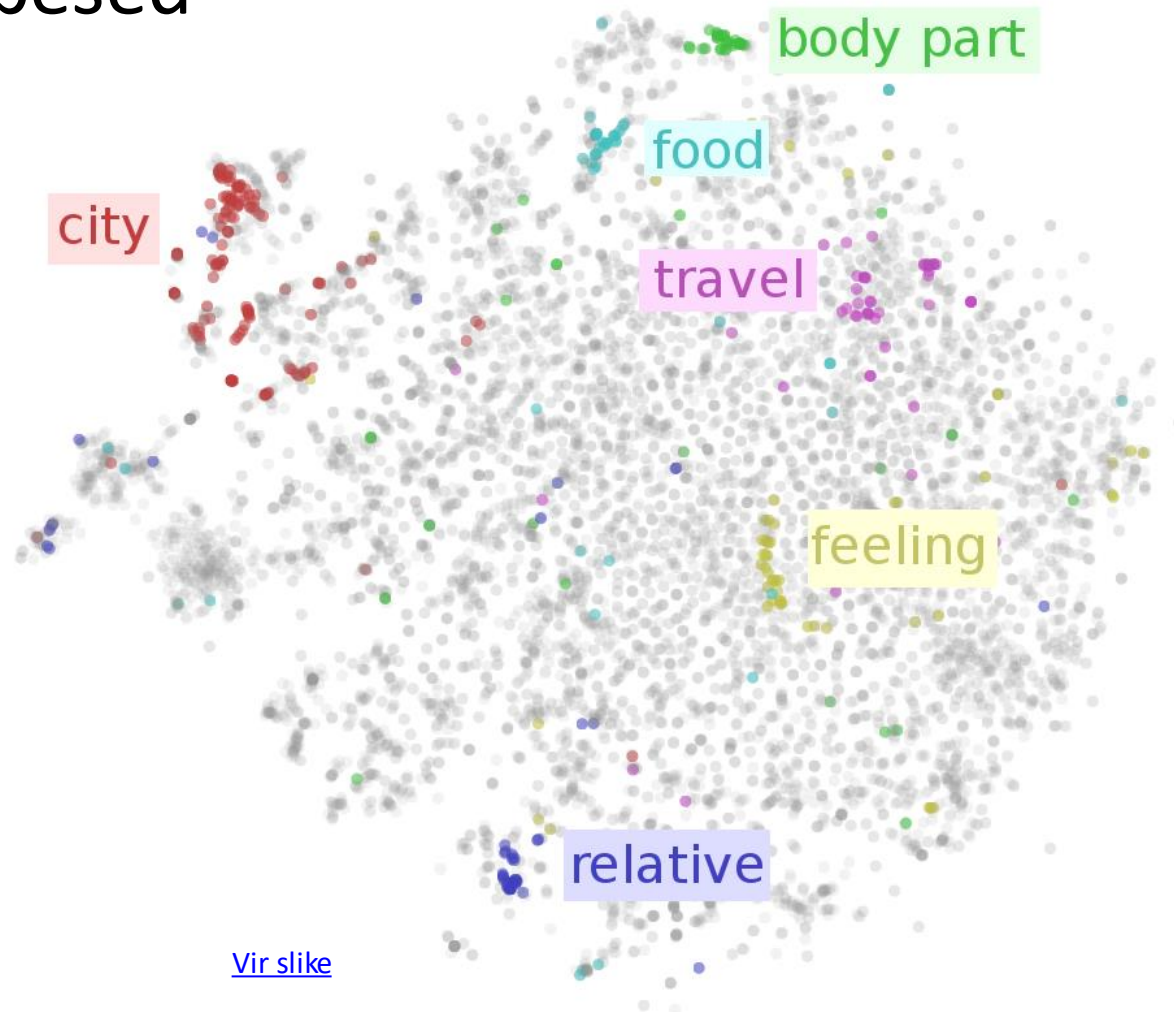


[Vir slike](#)

Vektorske podatkovne baze - VPB

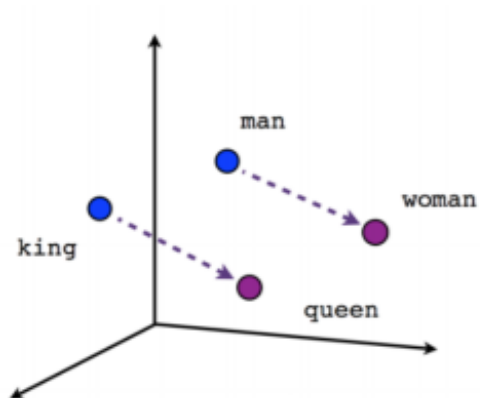
- VPB se uporabljajo za hranjenje kompleksnih vektorjev, ki nastanejo s strojno obdelavo podatkov in iz podatkov izluščijo številne (tudi skrite) lastnosti.
- Različne metode za generiranje vektorskih predstavitev:
 - modeli strojnega učenja,
 - vložitve besed (*word embeddings*) in
 - algoritmi za ekstrakcijo lastnosti (feature extraction algorithms).
- Cilj: semantično podobni vhodni podatki se pretvorijo v vektorje, ki so si blizu v več-dimenzionalnem prostoru.

Vložitve besed

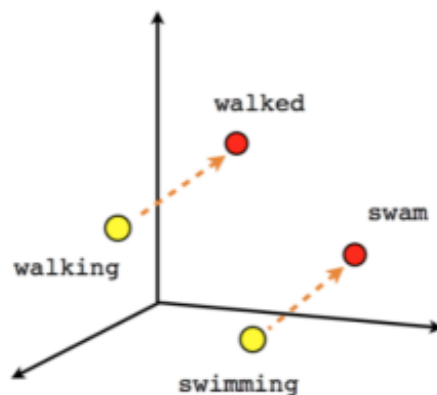


[Vir slike](#)

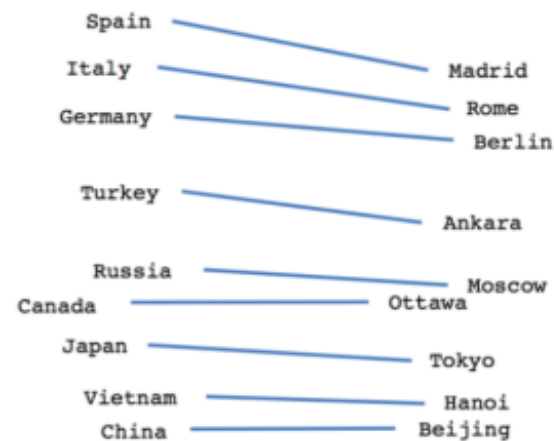
Lastnosti vložitev



Male-Female



Verb tense

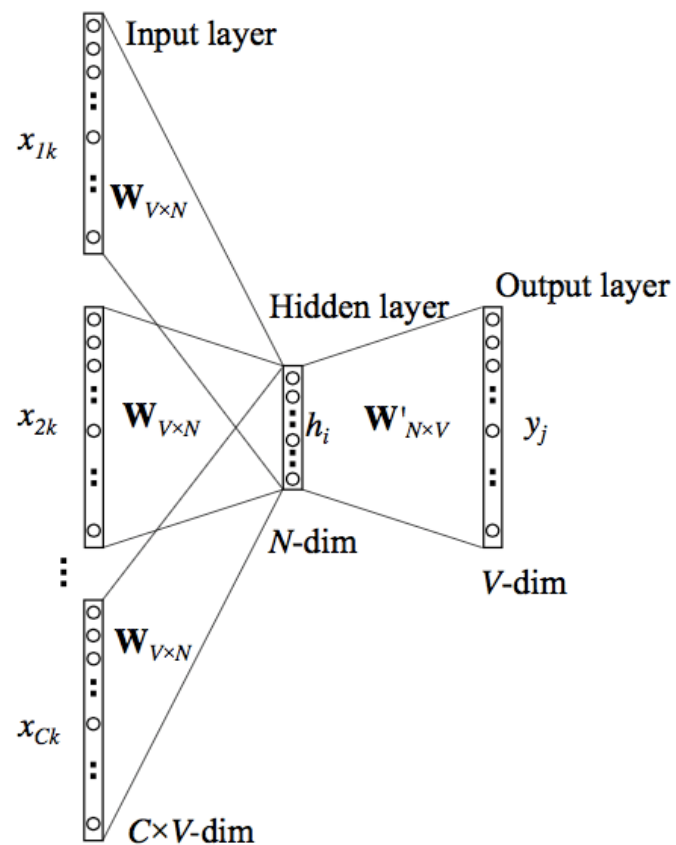
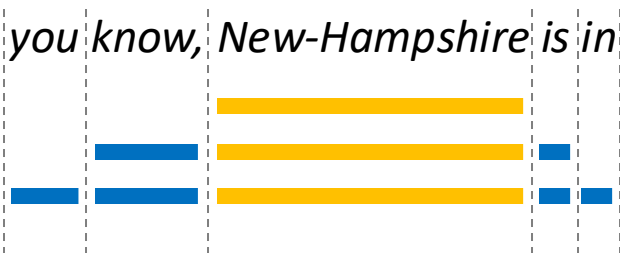


Country-Capital

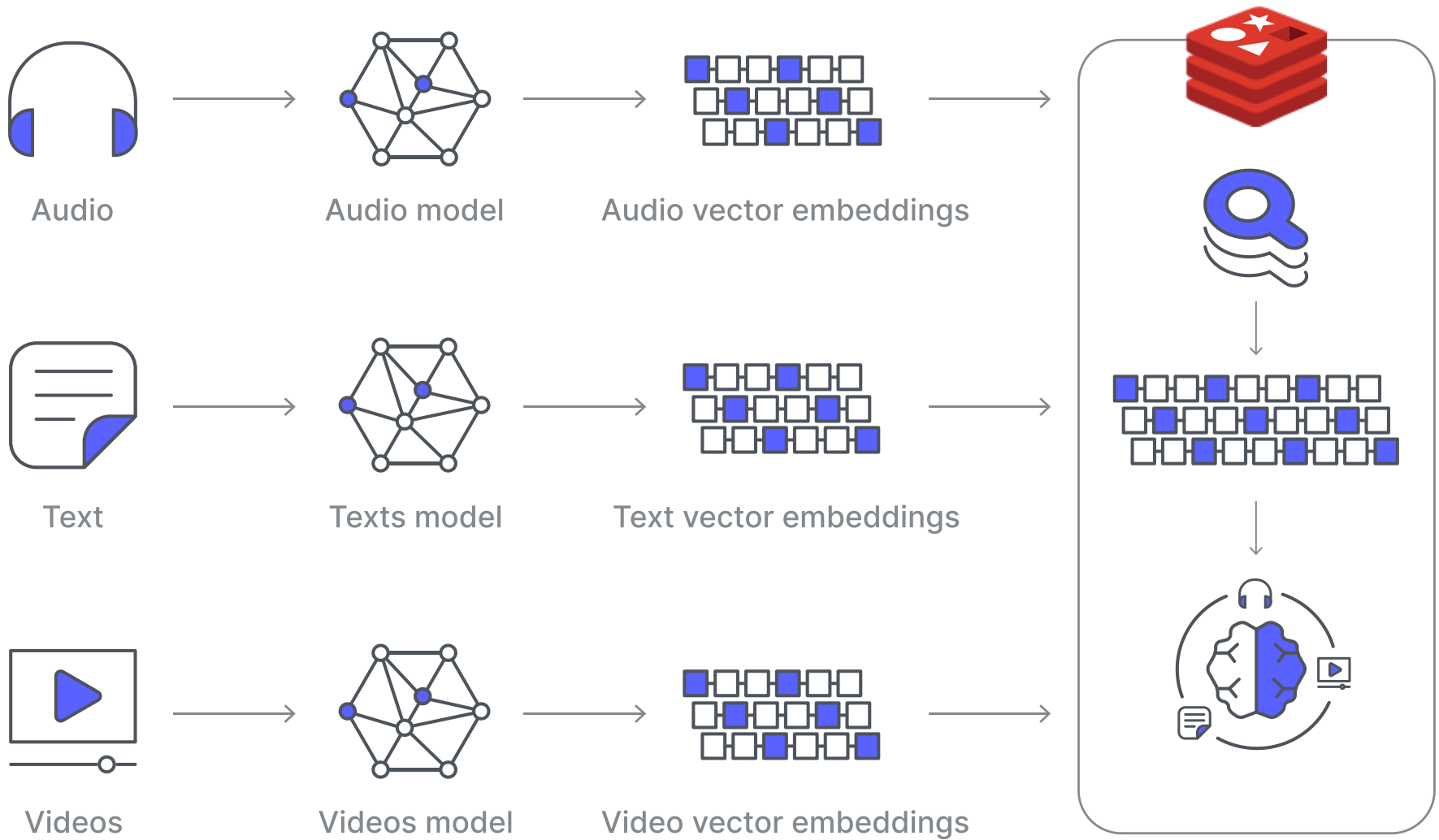
[Vir slike](#)

Pridobivanje vektorskih vložitev

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

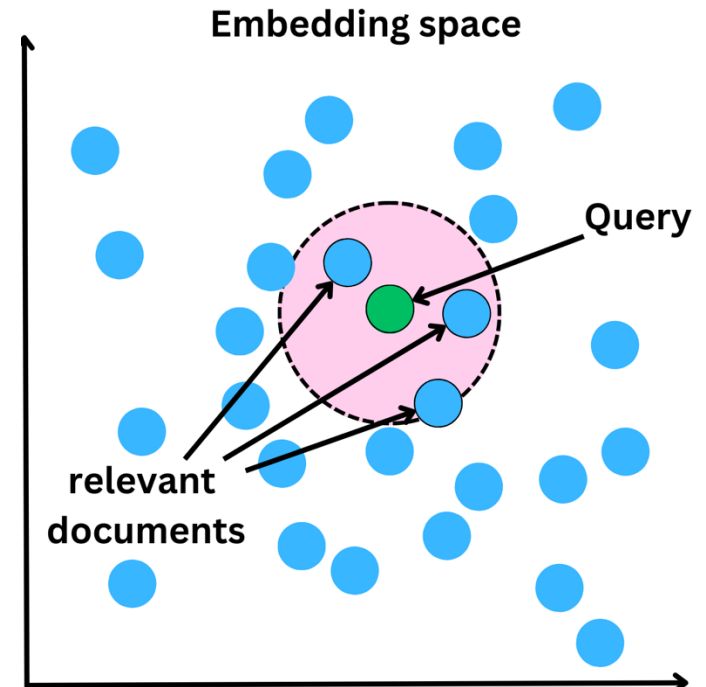


[Vir slike](#)



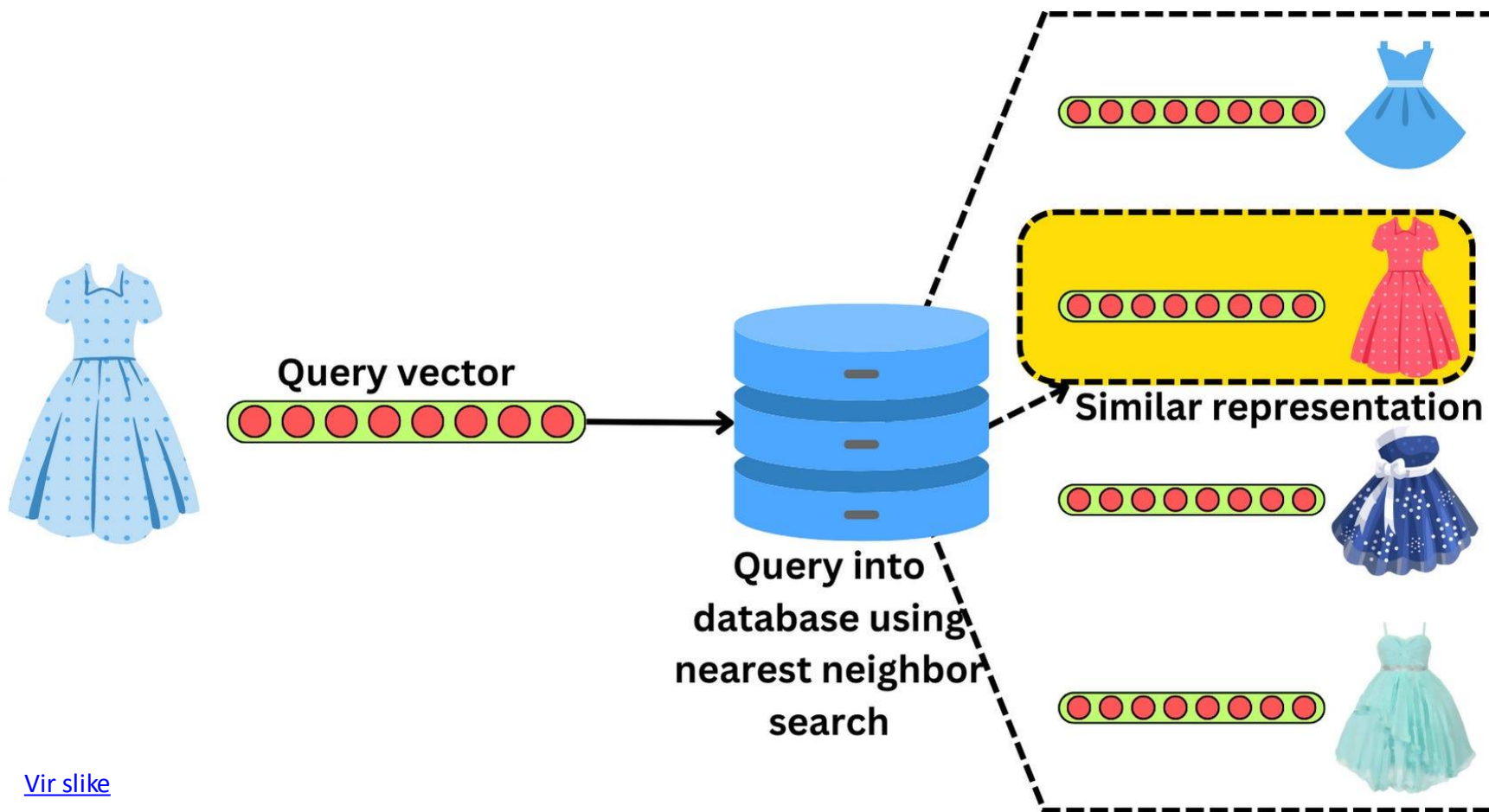
Prednosti VPB

- 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, ...



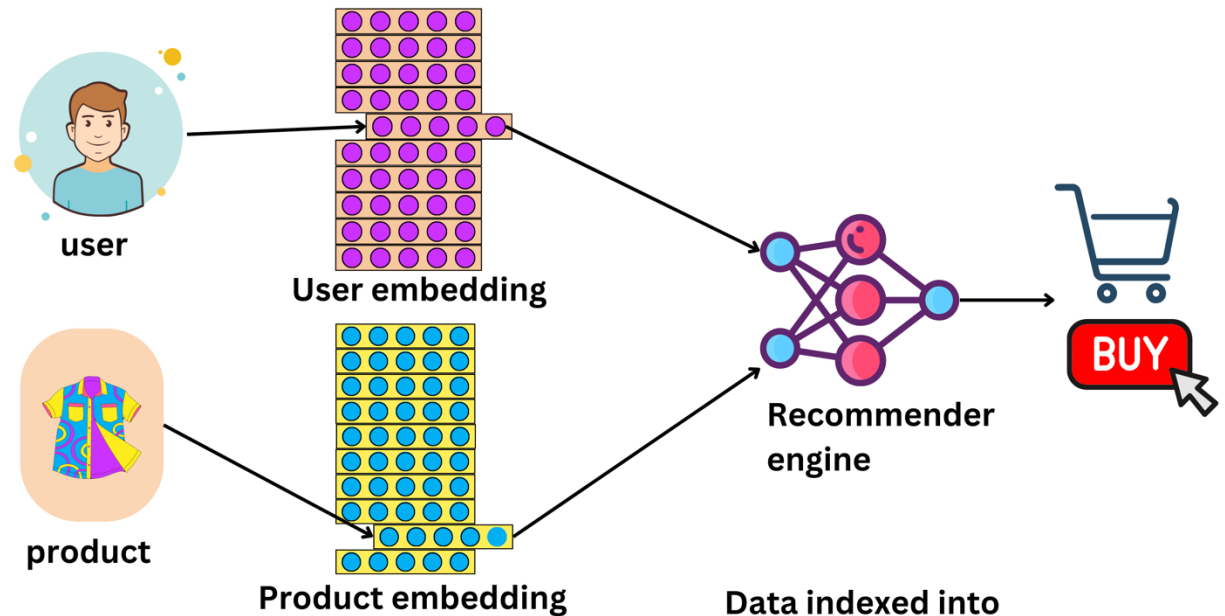
[Vir slike](#)

Iskanje po podobnosti



[Vir slike](#)

Priporočilni sistemi, e-poslovanje



211,06 € 12 x 20,51 €

Čena brez DDV: 173,00 € Hiter kredit do 84 obrokov

Objavljamo samo najnižje cene brez predatih.

Profesionalci zaupajo Rugged diskom.

Za več kot desetletje je so LaCie Rugged® diski prvi izbor profesionalcev. Zakaj? Zato ker LaCie Rugged diski veljajo za najbolj zaupanja vredne diske, ki gre za varno prenašanje podatkov na terenu.

design by neil poulton

Vmesnik: USB 3.0
Hitrost prenosa: 5 Gb/s USB
Kapaciteta: 5TB
Barva: siva, srebrna
Ohišje: kovinski, gumijasta zaščita
Priložen USB-C kabel
Lastnosti: odporen na padce do 1,2m (v nedeljujem stanju), odporno neopredeljeno ohišje (enakovredno pritisku 1 tonskega avta), odporen na del

Garancija: 2 leti

DODAJ V VOZIČEK

Shraniti na seznam izdelkov

Shraniti za primerjavo

Dostava iz zaloge

Izdelki bomo predvidoma poslali do 16. 1. 2025

Brezplačen prevoz v Enaa centru, Savača 3a, Ljubljana predvidoma od 16. 1. 2025

Bra izdelka: eničevskost

1 izdelki v skladu do prejetosti 21 točk

Popolni in deli

Enaa jamstvo in podpora

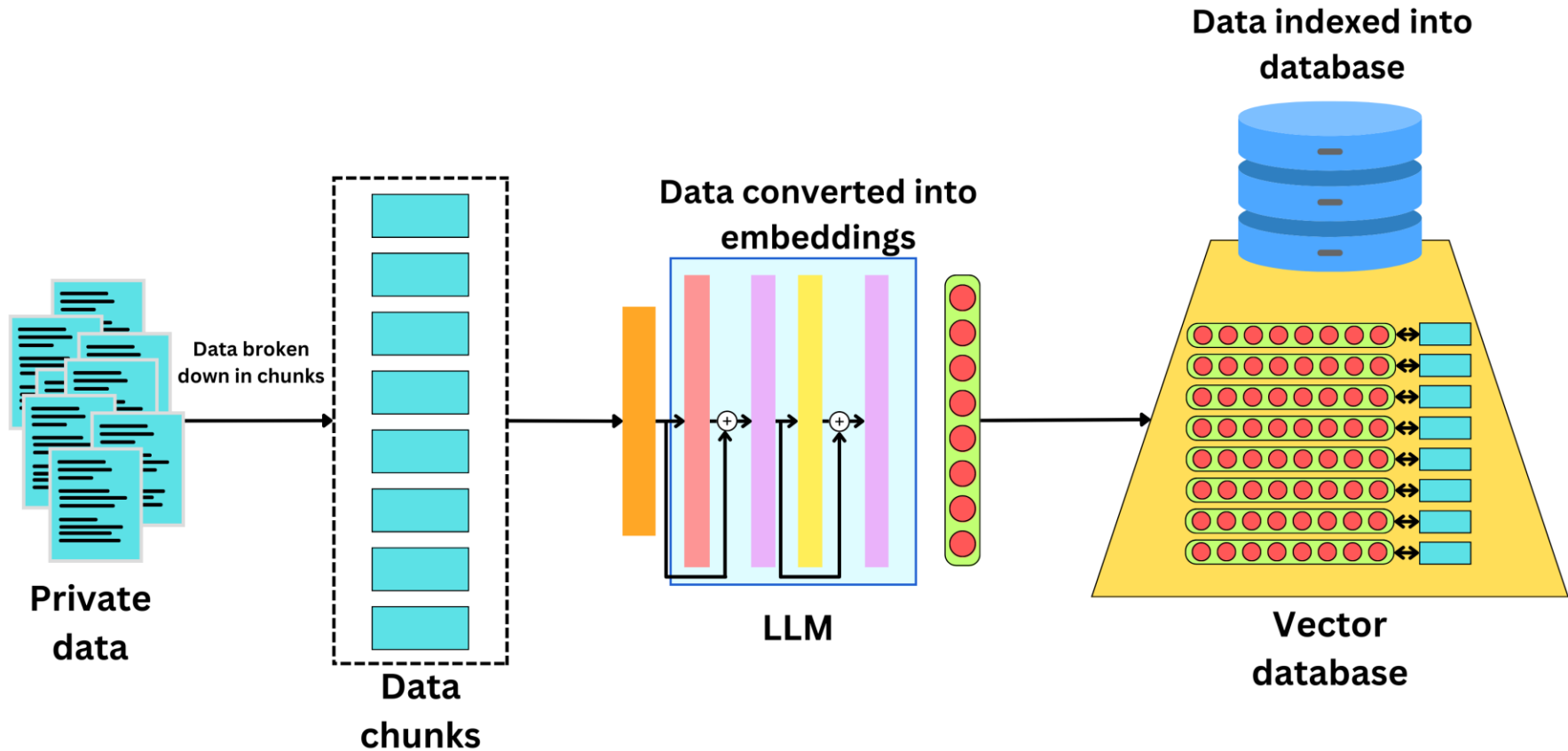
Kupljene izdelke lahko vrnete skladno z objavljenimi pogoji. V primerih, ko gre kaj narobe, bomo v dogovoru z vami poskušali najhitrejši rešitev. Nakup lahko opravite tudi brez registracije. Podjetjem in organizacijam omogočamo ugodne elektronske račune, predračune ter vnos števil naročil pred oddajo naročila. Varno in ugodno nakup nadgrajujemo s hitro dostavo, spletnimi krediti do 84 obrokov, montazo in drugimi storitvami, za kar prejmemo najboljše cene kupcev. V povprečju dosegamo najhitrejšo dostavo časa, v Ljubljani tudi osebni prevoz brez čakanja kar iz avtomobila. Izbire izdelke prejemate najhitreje v koderito in po lokalnih obsevnih željah storitev.

Splača se primerjati tudi s temi izdelki

Zunanji trdi disk LaCie 5TB Rugged Mini USB-C 3.0 - 211,06 €	2,5" SEAGATE Backup Plus Ultra Touch 4TB USB 3.0 / USB 2.0 - 206,74 €	Zunanji HDD WD My Passport za MAC 5TB Blue, 2,5" - 216,15 €	WD Elements 8TB USB 3.0 / USB 2.0 (WDBL6000HKBKESN) zunanji - 205,91 €	2,5" SEAGATE Expansion Portable 5TB HDD USB 3.0 6,4cm 2,52oll - 203,22 €
DODAJ V VOZIČEK	DODAJ V VOZIČEK	DODAJ V VOZIČEK	DODAJ V VOZIČEK	DODAJ V VOZIČEK

[Vir slike](#)

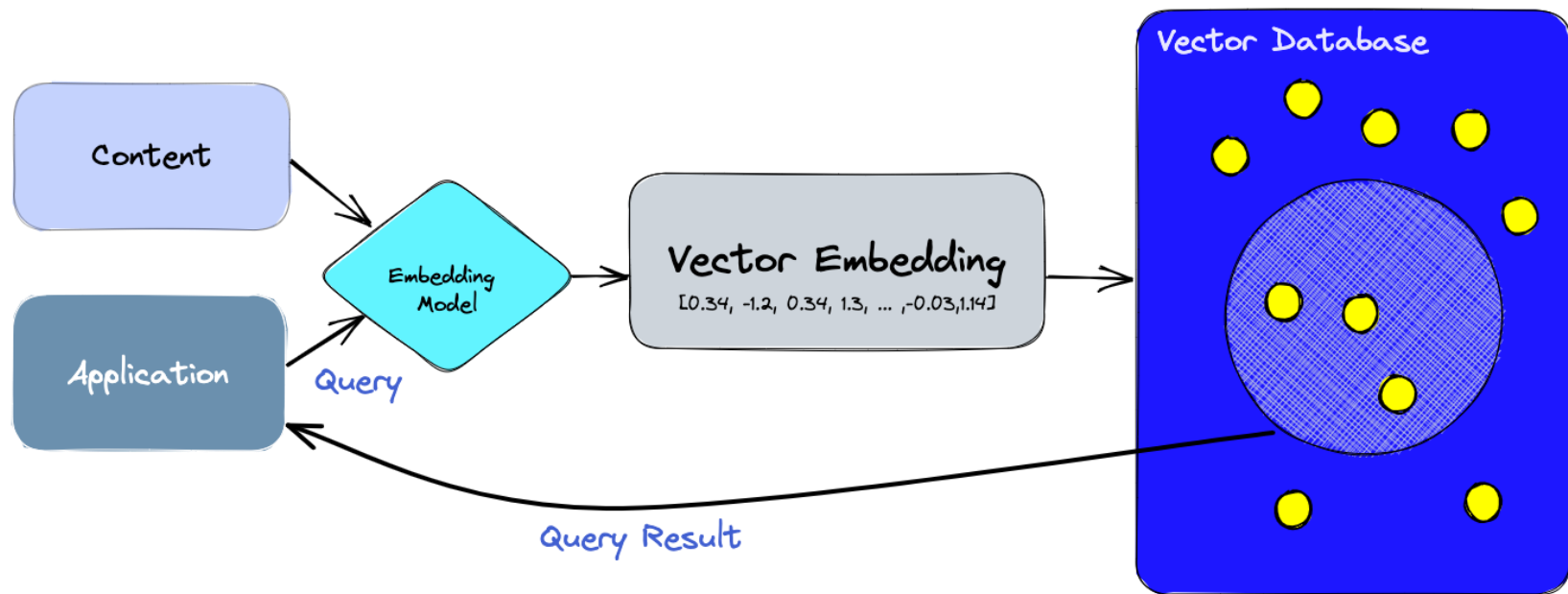
LLM in RAG





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.



[Vir slike](#)

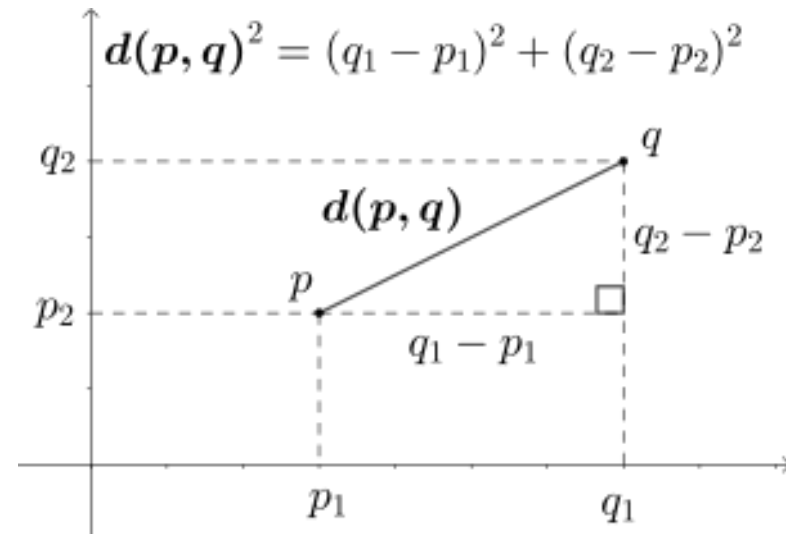


Metrike razdalje

- 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

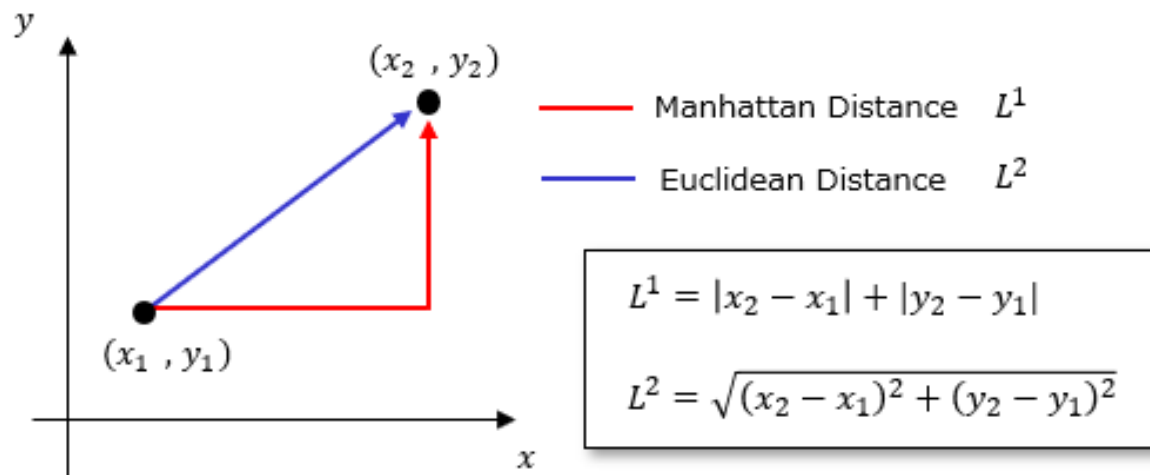
Evklidska razdalja

- 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.



[Vir](#)

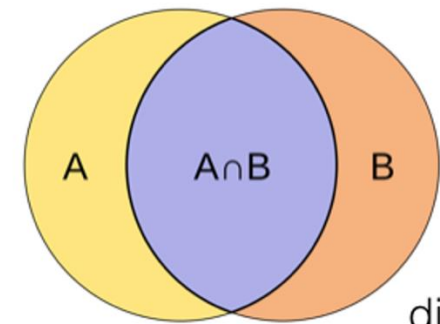
Jaccardova podobnost

- Jaccardova podobnost je podobnost med dvema množicama:

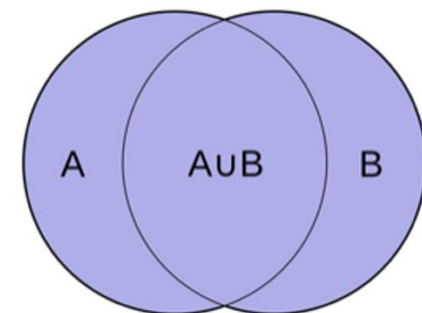
$$J(A, B) = \frac{A \cap B}{A \cup B}$$

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

The intersect of A & B

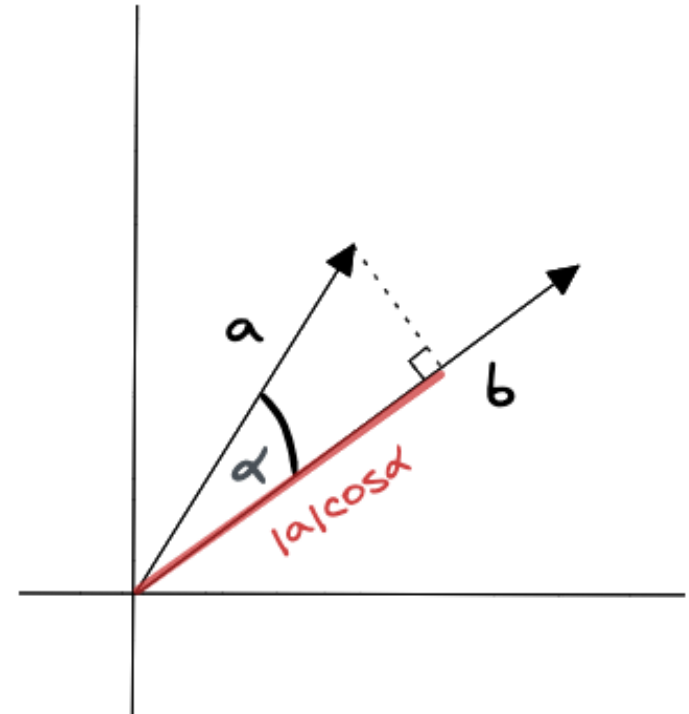


The union of A & B



Skalarni produkt

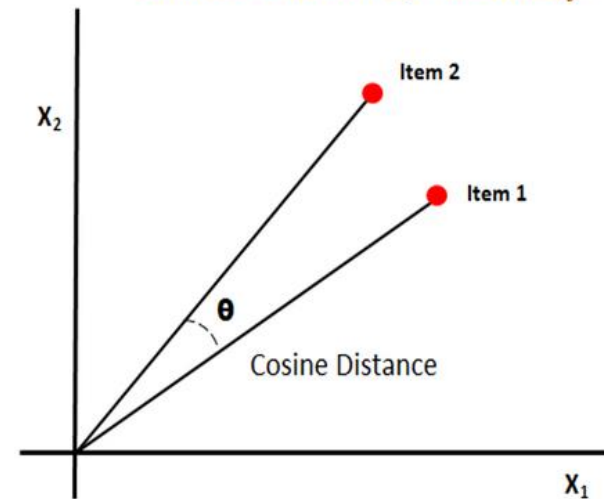
- Meri produkt velikosti dveh vektorjev in kosinus kota med njima.
- Obseg je od $-\infty$ do ∞ , kjer pozitivna vrednost predstavlja vektorje, ki kažejo v isto smer, 0 predstavlja ortogonalne vektorje, negativna vrednost pa predstavlja vektorje, ki kažejo v nasprotnih smereh.
- Dolžina vektorjev ključna.



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.

Cosine Distance/Similarity

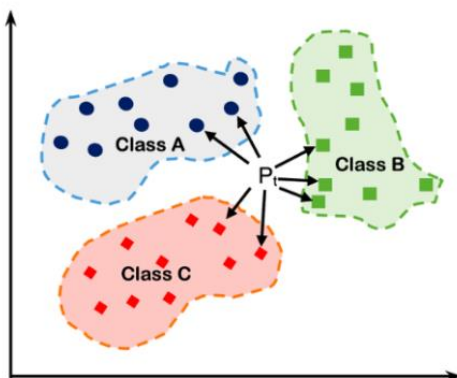


$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

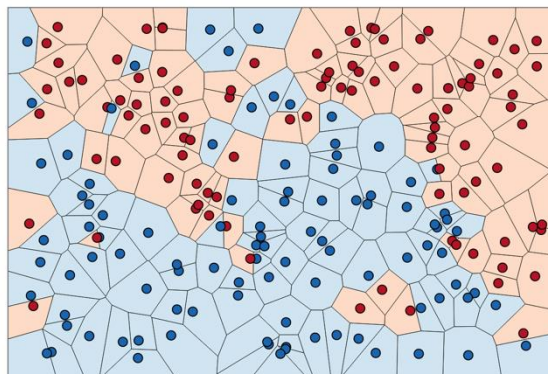
Iskanje podobnosti

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

K Nearest Neighbors



[Vir slike](#)



[Vir slike](#)

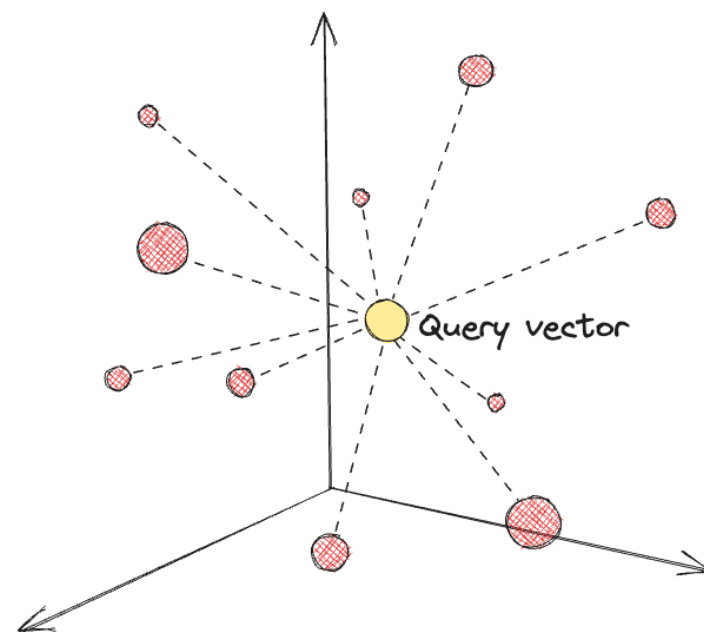


Indeksiranje

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

Ravni indeks

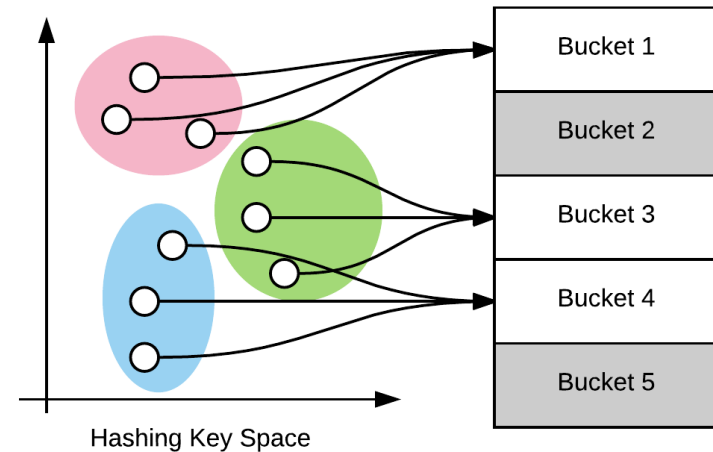
- Ravno indeksiranje (*Flat indexing*) – vektor hranimo v originalni obliki.
- Omogočajo pridobivanje natančnih rezultatov (natančno ujemanje)
- Hiter, natančen, majhna poraba prostora
- Primeren za majhne baze
<10 000 vektorjev.
- Slabost: linearna časovna komp.



[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 Spotify-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

