Računanje povednih moči

Probabilistične metode,
metoda vektorskega prostora,
obratna pogostost dokumentov.

Povedne moči - uvod

- Osnovni namen postopkov avtomatskega indeksiranja je zgoščevanje zapisa ob nespremenjeni vsebini:
  - z blokiranjem zavržemo besede brez vsebine (20% - 40% zgostitev,
  - s krnjenjem nevtraliziramo vplive sintakse (pomembna dodatna zgostitev, ker se ohranijo samo unikatni krni).
- Zanemarimo dejstvo, da z blokiranjem in krnjenjem izgubimo del vsebine, skrite v medsebojnem odnosu besed.

Povedne moči - uvod

- Iz Luhnove krivulje sledi, da imajo različne besede različno vlogo pri opisovanju vsebine.
- Delež posamezne besede v zalogi vsebine dokumenta imenujemo njeno povedno moč.

Povedne moči - uvod

- Dokumenti v zbirki opisujejo različne vsebine.
- Beseda, ki je pomembna za vsebinski opis dokumenta, s svojim pojavljanjem dokument na nek način in do neke mere loči od ostalih dokumentov.
- Beseda, ki bi se v zbirki pojavljala naključno, dokumentom ne bi prispevala vsebine.
- Iz odmika od naključnega pojavljanja neke besede lahko sklepamo na njeno pomensko funkcijo.

Povedne moči - uvod

- Statistične metode računanja povednih moči besed ocenjujejo vlogo besed v dokumentih z dveh zornih kotov:
  - reprezentacija: sposobnost besed, da predstavljajo vsebino dokumenta,
  - diskriminacija: sposobnost besed, da dokument ločijo od ostalih v zbirki.

Probabilistične metode

- Probabilistične metode računanja povednih moči ocenjujejo verjetnost, da bo neka beseda vsebovana v relevantnem dokumentu.
- Za razumevanje načela potrebujemo miselni preskok: ko ocenjujemo sposobnost besede, da predstavlja vsebino dokumenta pravzaprav ocenjujemo relevantnost besede za dani dokument.
- Največjo povedno moč v dokumentu, ki govori o vsebini *a*, bo imela beseda, ki se z največjo verjetnostjo pojavlja v dokumentih o *a*, in z najmanjšo verjetnostjo v dokumentih, ki niso o *a*.

IDF

Sparck Jones, 1972:

- Že dolgo vemo, da je beseda, ki se v nekem dokumentu pogosto pojavlja, verjetno pomembna za vsebino dokumenta. Frekvenca besede je torej povezana z njeno povedno močjo.
- Primernost neke besede za opisovanje vsebine pa je odvisna tudi od števila dokumentov, v katerih se pojavlja.
- Za opisovanje vsebine v podatkovni zbirki je primerna beseda, ki se zelo pogosto pojavlja v majhnem številu dokumentov.

IDF

Odvisnost med pojavljanjem besede *b* v dokumentih in njeno povedno močjo opisuje *obratna pogostost dokumentov* (inverse document frequency, IDF).

$$IDF_b = \log \frac{N - n_b}{n_b}$$

N število dokumentov v zbirki,

nb število dokumentov z besedo b.

Normalizacija dolžine dokumentov

- Besede v dolgih dokumentih bodo v povprečju imele večje frekvence, kot v kratkih.
- Če ne bi upoštevali tega, bi bile povedne moči besed v dolgih dokumentih neupravičeno visoke in dolgi dokumenti bi imeli več možnosti da bodo poiskani, kot kratki.
- Da bodo frekvence besed dobro predstavljale vsebino dokumentov, moramo dolžino dokumentov normalizirati.
- Normalizacija: frekvenc besed popravimo, da bodo take, kot da bi izvirale iz enako dolgih dokumentov.

Probabilistične metode

Croft, Harper, 1983:

- Povedna moč besede *b* v dokumentu *d* je odvisna od njene frekvence v dokumentu in *IDF* v zbirki:

$$PM_{bd} = IDF_b \times f_{bd}$$

$$f_{bd} = K + (1 - K)\frac{freq_{bd}}{\max\_ freq_d}$$

freqbd = frekvenca besede *b* v dokumentu *d,*
max_freqd = frekvenca najpogostejše besede v dokumentu *d (normalizacija glede dolžine dokumentov),*
K = konstanta, namenjena prilagajanju dolžini dokumentov.

Probabilistične metode

Croft, Harper, 1983:

- Sorodnost iskalne zahteve q in dokumenta d je enaka vsoti povednih moči skupnih besed:

$$sorodnost(q,d) = \sum_{i=1}^{n} (C + IDF_i) \times f_{id}$$

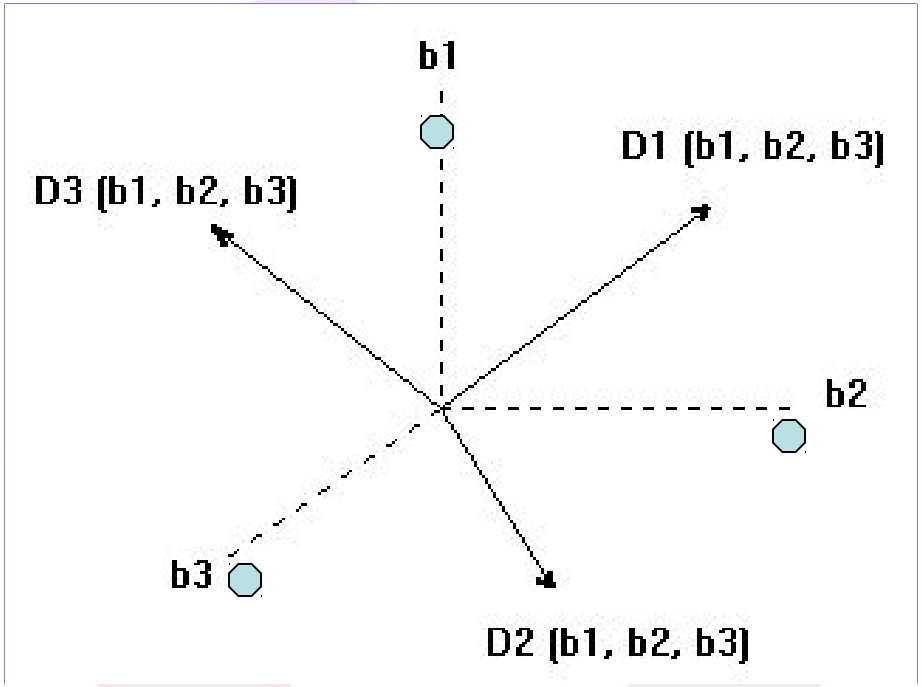
n = število besed, skupnih iskalni zahtevi *q* in dokumentu *d,*
C = konstanta, namenjena prilagajanju načinu indeksiranja.

Model vektorskega prostora

Salton, 1975

- Predstavljajmo si množico različnih besed v zbirki dokumentov tridimenzionalno, v prostoru.
- Vsaka beseda je točka s pozicijo v tem prostoru.
- Dokumente si predstavljajmo kot vektorje, sestavljene iz besed v tem prostoru.
- Izhodišča vektorjev naj bodo v središču prostora in vektorji naj bodo usmerjeni navzven.
- Smer vektora je odvisna od pozicij besed, ki ga sestavljajo.

Model vektorskega prostora



Poenostavljen prostor besed b1, b2, b3 in dokumentov D1, D2, D3.

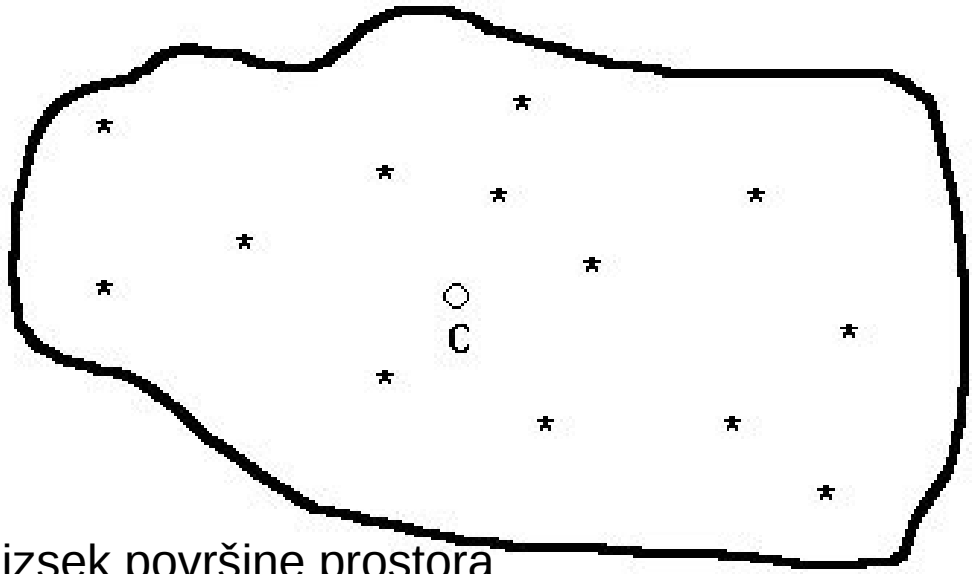
Tri različne besede oblikujejo tri-dimenzionalni prostor.

- Prisotnost besed v dveh vektorjih je izražena z kotom med vektorjema.
- Več ko imata vektorja različnih besed, večji je kot med njima.

Model vektorskega prostora

- Kot med vektorjema dveh identičnih dokumentov bi bil 0 stopinj.
- Prostor n različnih besed je n-dimenzionalen in vsebuje toliko vektorjev, kolikor je dokumentov v zbirki.

Model vektorskega prostora

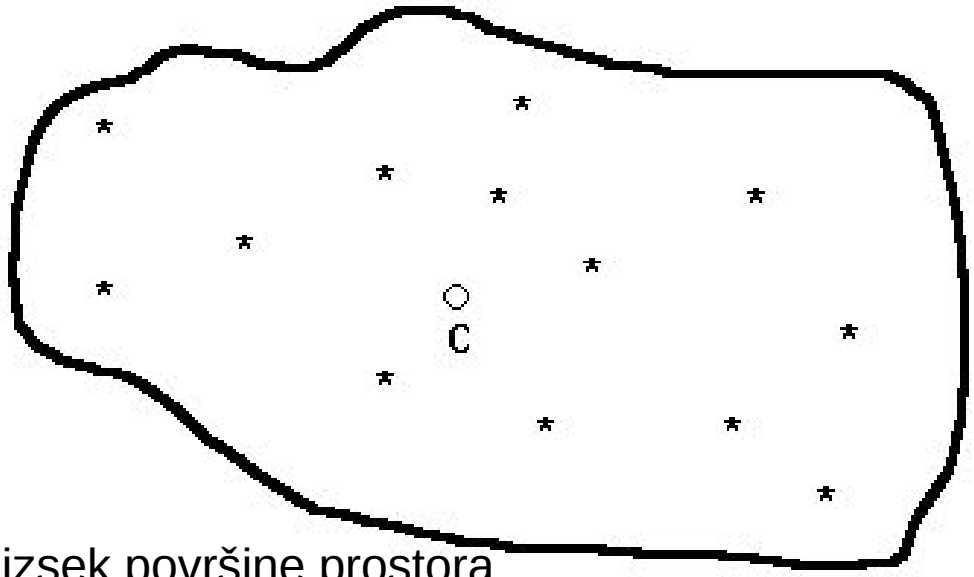


izsek površine prostora

Prostor besed je omejen s številom besed in ima zato svoj končen volumen in površino. Površino razgrnemo v dve dimenziji.

- Zvezdice predstavljajo konice vektorjev (dokumentov).
- C (centralni dokument ali centroid) je navidezni dokument, ki predstavlja povprečje vseh stvarnih dokumentov.

Model vektorskega prostora



izsek površine prostora

Kot med vektorjema ponazarja njuno različnost (različnost besed v dokumentih).

- Različnost dveh dokumentov se na zgornji sliki površine izraža kot razdalja med konicama njunih vektorjev.
- Stopnjo posebnosti dokumenta lahko prikažemo kot razdaljo konice njegovega vektorja od centroida.

Model vektorskega prostora

- V modelu vektorskega prostora ocenjujemo diskriminacijsko sposobnost besed.
- Diskriminacijska sposobnost besed - lastnost besed, da s svojo vsebino razločujejo (diskriminirajo) dokumente.
- Beseda z veliko diskriminacijsko sposobnostjo konice vektorjev razprši.
- Večja ko je razpršenost, večja je diskriminacijska sposobnost besede.

Model vektorskega prostora

- Diskriminacijsko sposobnost besed lahko ocenjujemo tako, da merimo razpršenost, ki jo povzročajo.

Model vektorskega prostora

$$Q = \sum_{i=1}^{n} S(C, D_i)$$

- Mero *razpršenosti* dokumentov *Q* predstavlja vsota *sorodnosti S* središčnega dokumenta *C* z vsakim posameznim dokumentom *D*.
- *i* predstavlja posamezno od *n* besed.

Model vektorskega prostora

- Sorodnost dokumenta in centroida je obratno sorazmerna s kotom med njunima vektorjema.
- Lahko jo izračunamo z *enačbo za kozinus kota*.

Model vektorskega prostora

$$DV_b = Q_b - Q$$

- Diskriminacijska vrednost *DV* besede *b* je mera sprememb, ki jo povzroči uvajanje te besede v prostor.
- Če izračunamo mero razpršenosti prostora brez in z prisotnostjo besede *b*, je razlika razpršenosti v obeh primerih ravno diskriminacijska vrednost besede *b*.

Model vektorskega prostora

- Diskriminacijska vrednost besede velja za vsak dokument v zbirki.
- Upošteva samo povprečno pogostost pojavljanja te besede v zbirki dokumentov.

Model vektorskega prostora

- Zanima nas količina informacije v besedi (njena povedna moč) v konkretnem dokumentu.
- Povedno moč *PV* besede *b* v dokumentu *D* dobimo tako, da diskriminacijsko vrednost *DV* te besede pomnožimo z njeno frekvenco *f* v dokumentu *D*.

$$PV_{bD} = DV_b \times f_{bD}$$