

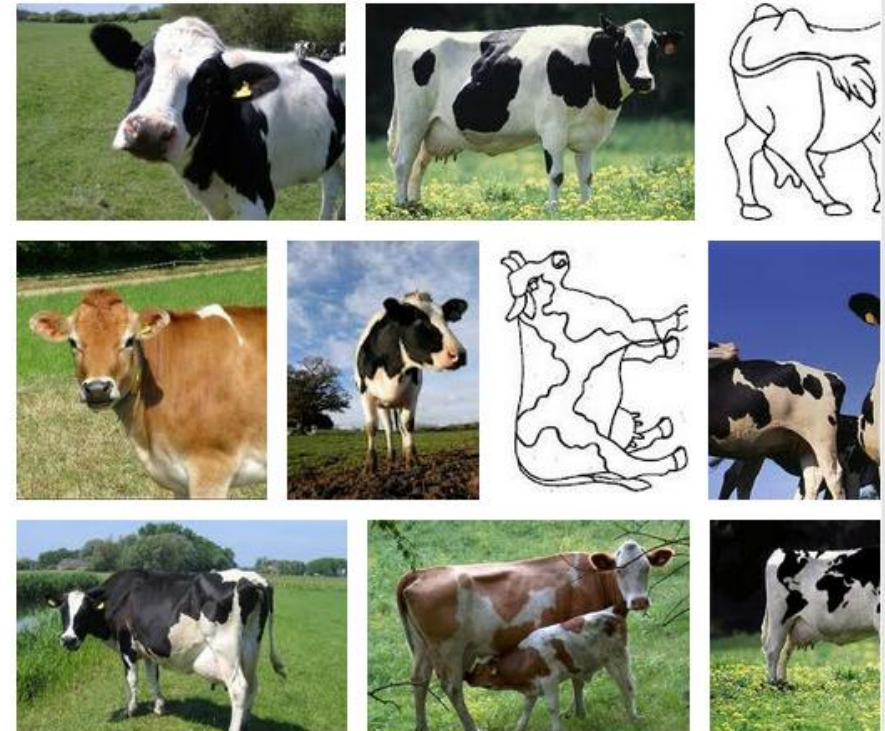


PRIDOBIVANJE INFORMACIJ: TEKST

Pridobivanje informacij

- *Information retrieval (IR)*
- Iskanje po nestrukturiranih gradivih (npr. dokumentih, zvoku, slikah ...)
 - navadno imamo opravka z **velikimi zbirkami** gradiv
- Ker so zbirke **velike**, potrebujemo **indekse**

Related searches: [happy cows](#) [cute cows](#) [funny cows](#) [cow face](#) [milk cow](#)

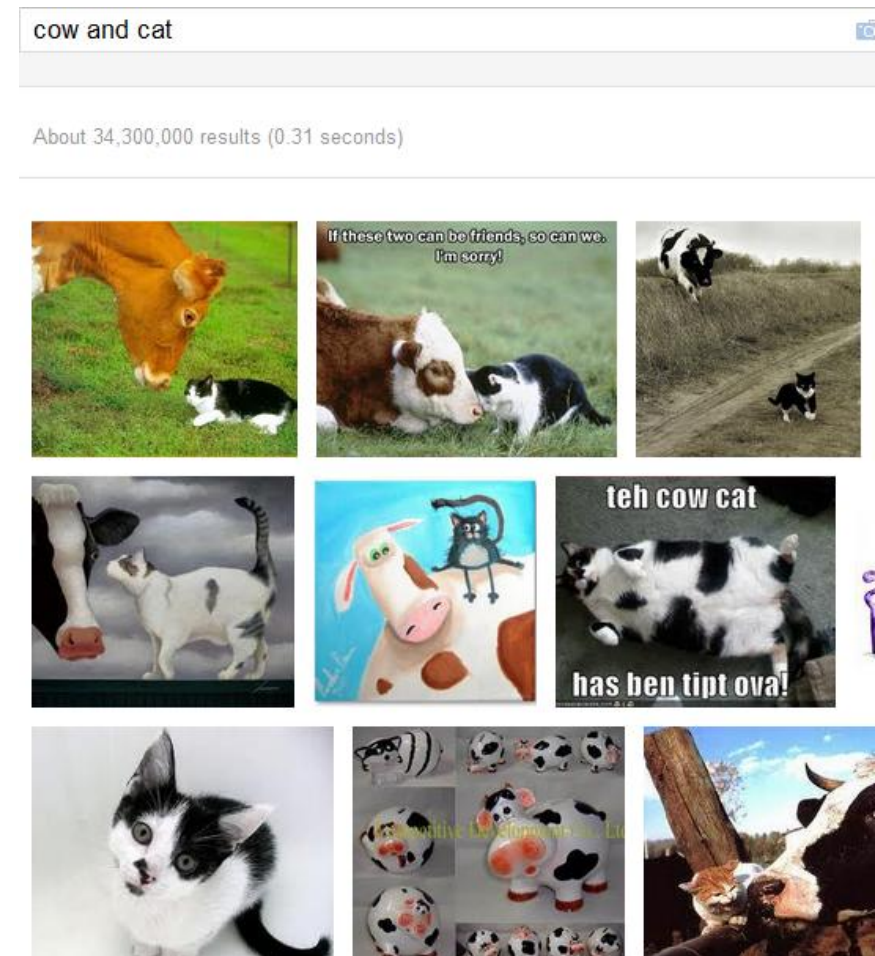




INDEKSIRANJE TEKSTA: OBRNJEN INDEKS

Iskanje z Boolovimi izrazi

- Preprost model za iskanje po gradivih
- Povpraševanja so Boolovi izrazi (AND, OR, NOT)
 - npr. "Cow AND Cat"
- Iskalnik vrne vse dokumente, ki ustrezajo izrazu
- Ali Google uporablja Boolove izraze?





- Imamo vsa zbrana dela
 - katera vsebujejo besede Brutus in Caesar, ne pa Calpurnia?
- Lahko po vrsti preiščemo vse besede vseh del in vrnemo rezultat
- Problem
 - počasi (za velike zbirke)
 - implementacija nekaterih operacij ni enostavna
 - npr. beseda Brutus naj bo blizu Caesar

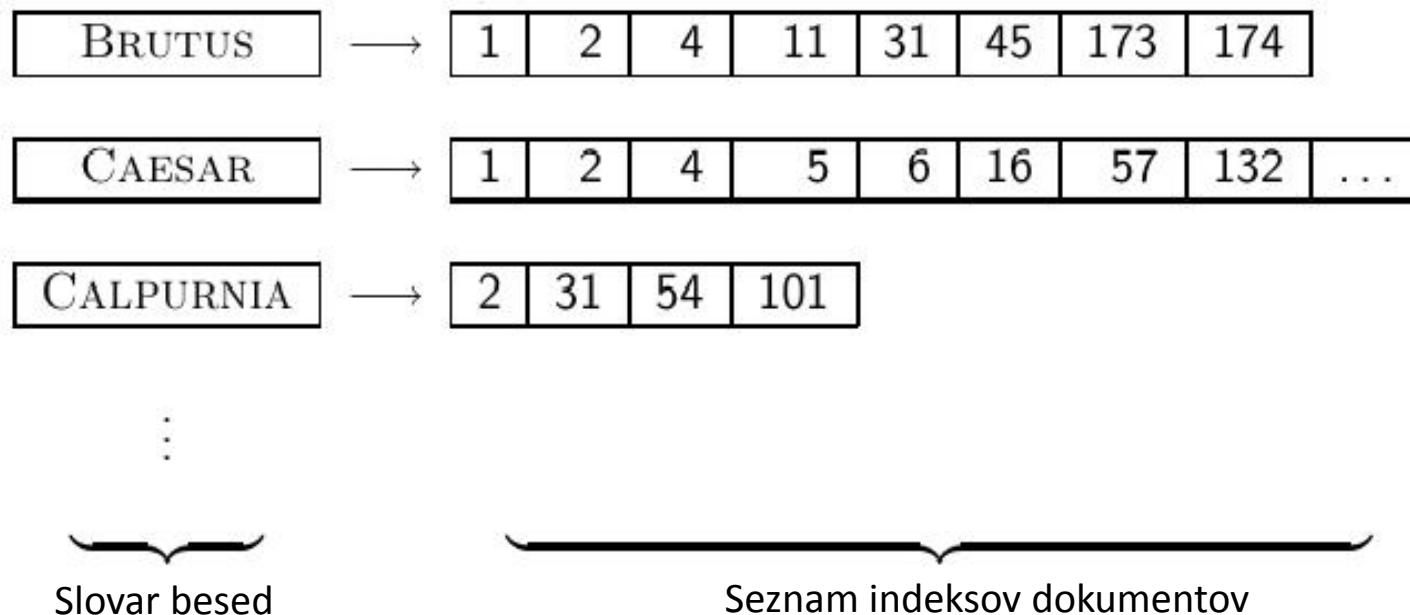


Primer: Shakespeare



Obrnjen indeks (*inverted index*)

- Dokument = seznam besed
- Obrnemo
 - za vsako besedo naredimo seznam dokumentov, ki besedo vsebujejo





Koraki pri gradnji

1. Zberemo dokumente za indeksiranje

Friends, Romans, countrymen. So let it be with Caesar ...

2. Iz dokumentov izločimo besede

Friends Romans countrymen So ...

3. Besede pretvorimo v simbole (*tokens*)

countryman so ... friend roman

4. Zgradimo obrnjen indeks





- Naredimo seznam vseh simbolov v dokumentih

Doc 1

I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

Doc 2

So let it be with
Caesar. The noble
Brutus hath told you
Caesar was ambitious

Gradnja indeksa

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2



- Seznam uredimo po abecedi



Gradnja indeksa

term	docID		term	docID
i	1		ambitious	2
did	1		be	2
enact	1		brutus	1
julius	1		brutus	2
caesar	1		capitol	1
i	1		caesar	1
was	1		caesar	2
killed	1		caesar	2
i'	1		did	1
the	1		enact	1
capitol	1		hath	1
brutus	1		i	1
killed	1		i	1
me	1	⇒	i'	1
so	2		it	2
let	2		julius	1
it	2		killed	1
be	2		killed	1
with	2		let	2
caesar	2		me	1
the	2		noble	2
noble	2		so	2
brutus	2		the	1
hath	2		the	2
told	2		told	2
you	2		you	2
caesar	2		was	1
was	2		was	2
ambitious	2		with	2



- Za vsak simbol naredimo seznam dokumentov
- Seznam naj bo urejen
- Zabeležimo tudi njegovo dolžino



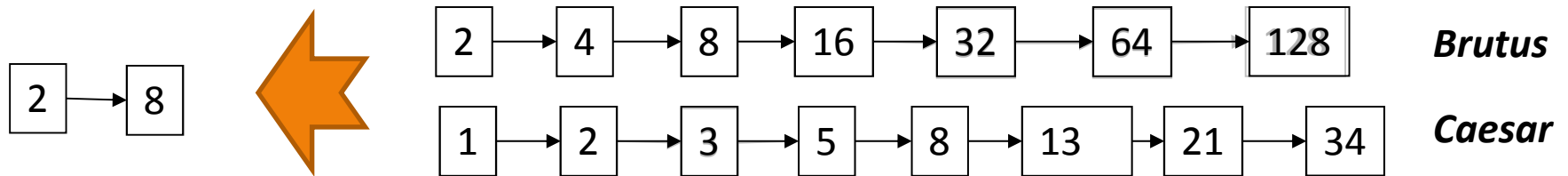
Gradnja indeksa

Term	docID	term	doc.	freq.	→	postings lists
ambitious	2	ambitious	1		→	2
be	2	be	1		→	2
brutus	1	brutus	2		→	1 → 2
brutus	2	capitol	1		→	1
capitol	1	caesar	2		→	1 → 2
caesar	1	did	1		→	1
caesar	2	enact	1		→	1
caesar	2	hath	1		→	2
did	1	i	1		→	1
enact	1	i'	1		→	1
hath	1	it	1		→	2
i	1	julius	1		→	1
i	1	killed	1		→	1
i'	1	let	1		→	2
it	2	me	1		→	1
julius	1	noble	1		→	2
killed	1	so	1		→	2
killed	1	the	2		→	1 → 2
let	2	told	1		→	2
me	1	you	1		→	2
noble	2	was	2		→	1 → 2
so	2	with	1		→	2
the	1					
the	2					
told	2					
you	2					
was	1					
was	2					
with	2					



Boolovi izrazi

- Za povpraševanje Brutus AND Caesar
 - najdemo seznam dokumentov za vsakega posebej
 - izračunamo presek

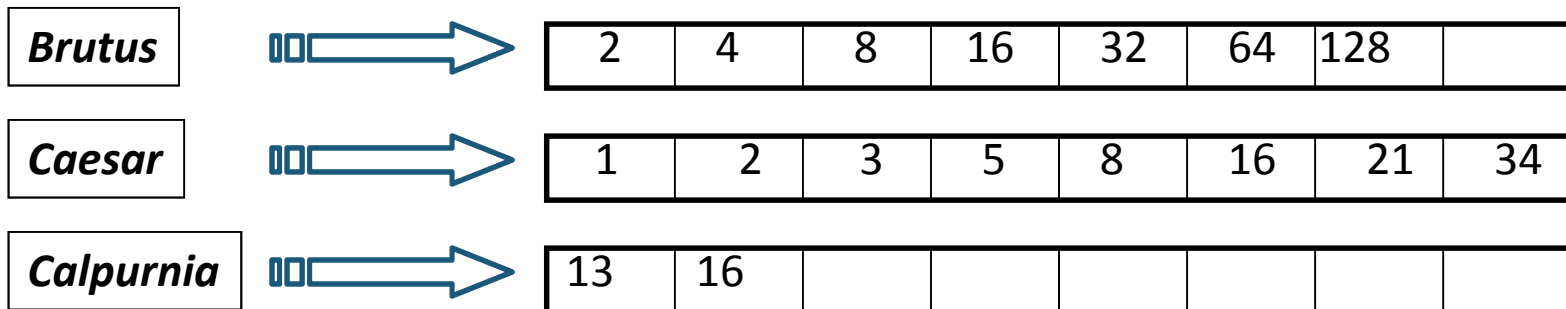


- linearen čas (vsota dolžine obeh), če so sezname dokumentov urejeni



Boolovi izrazi

- Kaj pa Brutus AND Calpurnia AND Caesar
 - procesiramo enako kot prej
- S katerim seznamom je najbolje začeti?
 - vzamemo najkrajšega najprej (najprej Calpurnia, potem Brutus, potem Caesar)



- V splošnem npr. (Madding OR Crowd) AND (Ignoble OR Strife)
 - ocenimo velikost vsakega OR bloka kot vsoto seznamov, nato vzamemo najkrajšega najprej

Ali Google uporablja Boolove izraze?

- Privzeta interpretacija povpraševanja $[w_1 w_2 \dots w_n]$ je
 - $w_1 \text{ AND } w_2 \text{ AND } \dots \text{ AND } w_n$
 - imamo tudi OR in NOT (-)
- Če rezultat ne vsebuje kake od teh besed:
 - besedo vsebuje povezava
 - stran vsebuje drugo verzijo besede (drugo črkovanje, sinonim...)
 - povpraševanje je dolgo
 - osnovni izraz vrne zelo malo zadetkov
- Google vrne **rangiran** spisek rezultatov – najpomembnejše oz. najboljše zadetke najprej
 - povpraševanje kot smo ga do sedaj omenjali, ne upošteva ranga

caessar and brutus and calpurnia and pipi

About 48,000 results (0.33 seconds)

Showing results for [caesar and brutus and calpurnia and pipi](#)

Search instead for [caessar and brutus and calpurnia and pipi](#)

[Julius Caesar by William Shakespeare. Search eText, Read Online ...](#)

[www.online-literature.com > William Shakespeare - Cached](#) 

He also uses contrasts between characters and relationships such as **Cassius** and **Brutus**, Octavius and Antony. Portia, **Brutus**, **Calpurnia**, and **Cesar** also paint ...

[SCENE II. A public place.](#)

[shakespeare.mit.edu/julius_caesar/julius_caesar.1.2.html - Cached](#) 

Enter **CAESAR**; ANTONY, for the course; **CALPURNIA**, PORTIA, DECIUS **BRUTUS**, CICERO, **BRUTUS**, **CASSIUS**, and **CASCA**; a great crowd following, among ...


[Rinse the Blood off My Toga](#)

[www.informalmusic.com/latinsoc/rinse.html - Cached](#) 

Brutus, senator and alleged friend of the deceased. **Calpurnia**, a (recent) widow ...

Brutus. I'm a Senator. I was **Caesar's** best friend. The name is **Brutus**. a small place with a few tables and a guy in the corner playing a crude, cool reed **pipe**. ...

[\[PDF\] Julius Caesar Study Guide.indd](#)

[www.ums.org/assets/pdf/studyguide/juliuscaesar-sg.pdf](#) 

File Format: PDF/Adobe Acrobat - [Quick View](#)

the prophetic dream of his wife **Calpurnia**, **Caesar** goes to the Capitol on the Ides aristocrats and even **Caesar's** friend **Brutus**, conspired to kill him. A **pipe** or long pole suspended horizontally above the stage, upon which ...



BESEDE -> SIMBOLI



- Želimo, da so besede v slovarju v nevtralni obliki
 - U.S.A -> USA
 - wanting -> want
 - windows -> window
 - ...
- Koraki
 - dokument razbijemo na besede
 - odstranimo pogoste besede
 - besede normaliziramo
 - besede pretvorimo v nevtralno obliko



besede -> simboli

window

About 2,360,000,000 results (0.25 seconds)

[Microsoft Windows](#)

[windows.microsoft.com/](#) - Cached

The official website for the Microsoft **Windows** operating system. Explore **Windows** info, get downloads, and find the latest PCs for **Windows**.

[Windows downloads](#) - [Windows 7](#) - [Windows XP](#) - [Microsoft Update](#)

[Microsoft Corporation: Software, Smartphones, Online, Games ...](#)

[www.microsoft.com/](#) - Cached

Shop for your new **Windows** Phone. ... Your search for great **Windows** Phone apps is over. **Windows** Phone apps to match your lifestyle. Whether you're a foodie, ...

[Window - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Window](#) - Cached

A half-glazed **window** of the 17th century from Scotland. A **window** is a transparent or translucent opening in a wall or door that allows the passage of light and, ...

[Microsoft Windows](#) - [Casement window](#) - [Sash window](#) - [Display window](#)

[Andersen Windows - Federal Energy Tax Credit - Energy Efficient ...](#)

[www.andersenwindows.com/](#) - Cached

Andersen offers a broad range of replacement **windows** and doors and new construction **windows** and doors. Federal energy tax credits are available for ...

[Pella Windows and Doors | Wood, Fiberglass, Vinyl Windows | Pella ...](#)

[www.pella.com/](#) - Cached

Discover Pella's energy efficient replacement **windows** and doors. Photo gallery, product builder, project ideas, expert consultations and more at Pella.com.



- Upoštevanje ločil, odstranjevanje ločil v besedah
 - U.S.A. -> USA
 - kaj vrne Google za C.A.T.?
- Ena ali dve besedi?
 - Hewlett-Packard
 - State-of-the-art
 - co-education
 - cheap San Francisco-Los Angeles fares
- Številke – želimo enoten zapis
 - 3/20/91
 - 20/3/91
 - Mar 20, 1991
- Kitajščina (ni presledkov)
- Združevanje besed
 - Lebensversicherungsgesellschaftsangestellter



Razbijanje na besede

莎拉波娃现在居住在美国东南部的佛罗里达。今年4月9日，莎拉波娃在美国第一大城市纽约度过了18岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Kitajščina nima presledkov

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるMOTTA INAIキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを800字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて10月20日までにお送りください。大賞受賞者には、50万円相当の旅行券とエコ製品2点の副賞が贈られます。

Japonščina nima presledkov + uporabljajo 4 vrste znakov (abeced) – povpraševanje je lahko v drugi kot dokumenti

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← → ← → ← START

Menjavanje smeri pisanja v arabščini



Pretvorba besed (normalizacija)

- Vse besede pretvorimo v male črke
- Odstranjevanje akcentov ipd.
 - Universität vs. Universitaet
 - résumé vs. resume
- Izpustimo vse “*stop words*”
 - pogoste besede, ki se skoraj vedno pojavljajo v dokumentih
 - a, an, and, ..., with, the ...
 - novejši trendi iskalnikov so, da izpuščajo zelo malo besed
 - sicer so povpraševanja kot “To be or not to be” problematična

The boy's cars are of different colors

→

boy's cars are different colors



Lematizacija

- Lematizacija
 - spregatve, množine itn. pretvorimo v **nevtrarno obliko**
 - am, are, is → be
 - car, cars, car's, cars' → car
 - ...
- Kompleksna – potrebujemo dober model za posamezni jezik
- Angleški, npr. [Wordnet Lemmatizer](#)
- Slovenščina, npr. [Obeliks, demo](#)

We went to see the prettiest flat
→

We go to see the pretty flat





- *stemming*
- Enostavnejša alternativa lematizaciji
 - **odrežemo konce besed**
 - še vedno odvisno od jezika
 - automate, automatic, automation → automat
- Za angleščino je nekaj znanih pristopov, npr. **Porterjev** algoritem
 - nabor pravil, npr.:
 - briši “ement” če je ostanek daljši od enega znaka
 - replacement → replac
 - cement → cement
 - demo
- Drugi jeziki (vključno s slovenščino so še bolj zapleteni)
 - finski glagol ima lahko 12000 oblik
 - demo za slovenščino

Korenjenje

Pravilo	Primer
SSSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

boy's cars are different colors

→

boy ' car are differ color



Primer treh angleških korenjenj:

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to pictur of express that is more biolog transpar and access to interpret

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret





- “Stop words”? – niti ne
- Normalizacija
- Razbitje na simbole
- Pretvorba v majhne črke
- Korenjenje
- Upoštevanje ne-latinskih abeced
- Akcenti
- Sestavljene besede
- Števila



Google?

window

About 2,360,000,000 results (0.25 seconds)

[Microsoft Windows](#)

[windows.microsoft.com/](#) - Cached

The official website for the Microsoft **Windows** operating system. Explore **Windows** info, get downloads, and find the latest PCs for **Windows**.

[Windows downloads](#) - [Windows 7](#) - [Windows XP](#) - [Microsoft Update](#)

[Microsoft Corporation: Software, Smartphones, Online, Games ...](#)

[www.microsoft.com/](#) - Cached

Shop for your new **Windows** Phone. ... Your search for great **Windows** Phone apps is over. **Windows** Phone apps to match your lifestyle. Whether you're a foodie, ...

[Window - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Window](#) - Cached

A half-glazed **window** of the 17th century from Scotland. A **window** is a transparent or translucent opening in a wall or door that allows the passage of light and, ...

[Microsoft Windows](#) - [Casement window](#) - [Sash window](#) - [Display window](#)

[Andersen Windows - Federal Energy Tax Credit - Energy Efficient ...](#)

[www.andersenwindows.com/](#) - Cached

Andersen offers a broad range of replacement **windows** and doors and new construction **windows** and doors. Federal energy tax credits are available for ...

[Pella Windows and Doors | Wood, Fiberglass, Vinyl Windows | Pella ...](#)

[www.pella.com/](#) - Cached

Discover Pella's energy efficient replacement **windows** and doors. Photo gallery, product builder, project ideas, expert consultations and more at Pella.com.

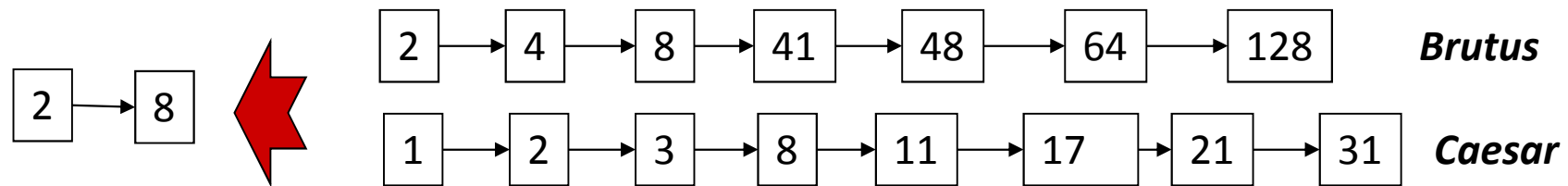


RAZŠIRITVE OBRNJENEGA INDEKSA

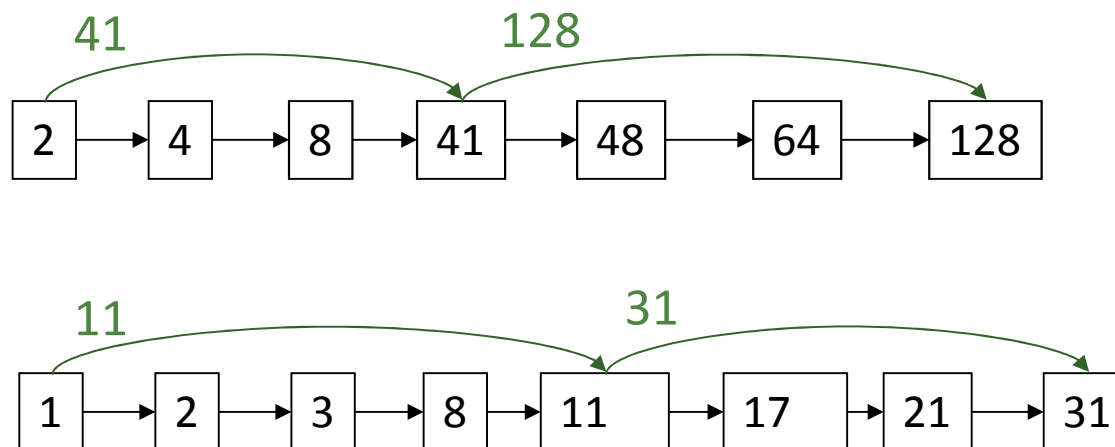


Pohitritev iskanja

- Osnovno iskanje za AND je linearno v času $O(n+m)$



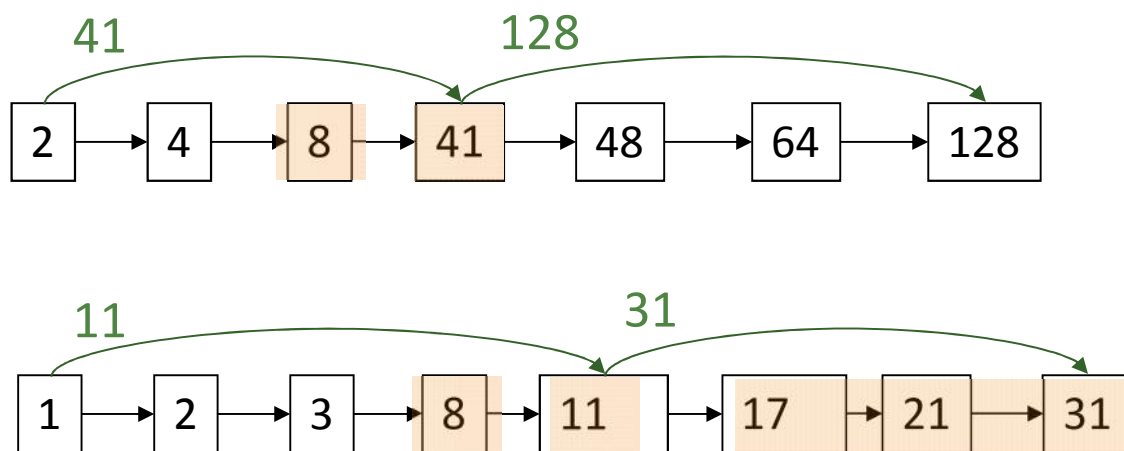
- Uvedemo dodatne kazalce – skoke
 - hitreje preskočimo dele seznama, ki ne vsebuje zadetkov





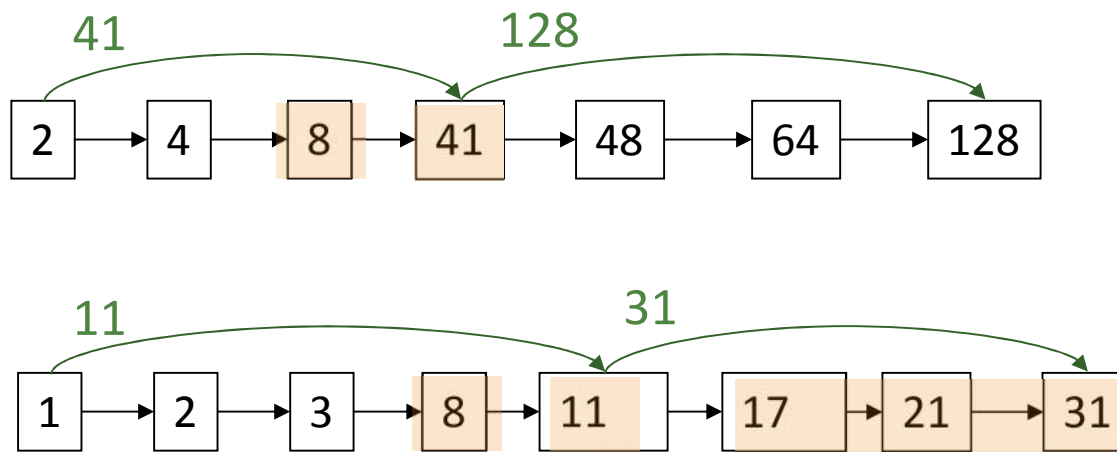
Pohitritev iskanja

- Če smo v obeh seznamih na 8. dokumentu, lahko pri pomiku na naslednjega v drugem preskočimo precej elementov



Pohitritev iskanja

- Kam postavimo skoke?
 - več jih je, manj preskočimo, vendar večkrat
 - manj jih je, več preskočimo, vendar manjkrat
- Primer enostavnega pravila:
 - za dolžino seznama L postavimo \sqrt{L} skokov





- Kako iščemo zaporedja več besed, npr. “Stanford University”?
 - cca. 10% povpraševanj je take oblike
- Lahko bi v indeksu hranili še vse **pare** besed (*biword index*)
 - daljša zaporedja pretvorimo v pare, npr.
 - “stanford university palo alto” pretvorimo v “stanford university” AND “university palo” AND “palo alto”
 - vendar bi morali še dodatno filtrirati rezultate, da bi res vrnili prave zadetke – počasi, ker je potrebno pregledati vse dokumente
- Problem
 - indeks se zelo poveča
 - pri daljših zaporedjih besed počasno, če želimo vrniti prave rezultate

Iskanje fraz

“Stanford University”

About 33,200,000 results (0.28 seconds)

[Stanford University](#)

[www.stanford.edu/](#) - Cached

Stanford University is one of the world's leading research and teaching institutions. It is located in Palo Alto, California.

[Admission](#)

Introduction. Introduction. Stanford students stand out for their ...

[Stanford Engineering](#)

Stanford Engineering pushes the frontiers of modern science and ...

[School of Medicine](#)

Offers programs leading to an MD, MS or PhD degree.

[Photo - Stanford Campus](#)

About Stanford, Admission, Academics, Research, Life on ...

[Stanford Graduate School of ...](#)

Stanford Graduate School of Business offers full-time MBA ...

[Jobs](#)

At Stanford, you can make all the difference in the world. Go to ...

Search stanford.edu

[Stanford University - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Stanford_University](#) - Cached

The Leland Stanford Junior University, commonly referred to as **Stanford University** or Stanford, is a private research university on an 8180-acre (3310 ha) ...

[Stanford Football - Fiesta Bowl - Stanford University Official Athletic ...](#)

[www.gostanford.com/](#) - Cached

The Stanford Cardinal Official Athletic Site, partner of CBS College Sports Networks, Inc. The most comprehensive coverage of Stanford Athletics on the web.

[News for “Stanford University”](#)



Iskanje fraz

- **Pozicijski indeks**
 - v obrnjen indeks shranimo še položaj besed v dokumentu
- V katerem dokumentu je “ $to_1 be_2 or_3 not_4 to_5 be_6$ ”
- Pri iskanju moramo upoštevati tudi položaje besed in razdalje med položaji
 - počasneje
- Pozicijski indeks lahko uporabimo tudi za bolj splošna iskanja, npr. besed, ki so si blizu
 - najdi vse pojavitve, ko sta si besedi *employment in place* največ 4 besede narazen

to, 993427:

1: ⟨7, 18, 33, 72, 86, 231⟩;

2: ⟨1, 17, 74, 222, 255⟩;

4: ⟨8, 16, 190, 429, 433⟩;

5: ⟨363, 367⟩;

be, 178239:

1: ⟨17, 25⟩;

4: ⟨17, 191, 291, 430, 434⟩;

5: ⟨14, 19, 101⟩; . . .



Iskanje fraz

- Velikost pozicijskega indeksa je precej večja kot običajnega obrnjenega indeksa
 - velikost je odvisna od povprečne dolžine dokumentov
 - v povprečju 2-4 krat večji od običajnega
 - 35-50% velikosti originalnih dokumentov
- Navadno se pri iskanju fraz kombinira *biword* in pozicijski indeks
 - *biword* za pogoste fraze (npr. Michael Jackson ali The Who)
 - iskanje je precej hitrejše
 - pozicijski za ostale neindeksirane fraze





- Lahko preverimo kaj je počasneje:
 - iskanje z Boolovim izrazom
 - iskanje fraze

Google?

"the who"

About 54,900,000 results (0.24 seconds)

[The Who's Official Website - News from Roger Daltrey, Pete ...](#)
[www.thewho.com/](#) - Cached

ENTER HERE >> - Welcome to thewho.com, the official home of **The Who** online.

[The Who - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/The_Who](#) - Cached

The Who are an English rock band formed in 1964 by Roger Daltrey (lead vocals, harmonica and guitar), Pete Townshend (guitar, keyboards and vocals), John ...

[Members](#) - [The Who discography](#) - [Keith Moon](#) - [Roger Daltrey](#)

[The Who – Free listening, videos, concerts, stats, & pictures at Last.fm](#)



[www.last.fm/music/The+Who](#)

16 Dec 2010

Watch videos & listen free to **The Who**: My ...

[The Who - My Generation](#)



[www.youtube.com/watch?v=594WLz3JI](#)

30 Apr 2008 - 3 min - polydorclassics

Album: **The Who** Sings My Generation (1965)

[The Who - Pinball Wizard](#)



[www.youtube.com/watch?v=aOUqRZkR8dE](#)

20 Jun 2008 - 3 min - eaglerocktv

Album: **Tommy** (1969)





PRIBLIŽNA POVPRAŠEVANJA



Približna povpraševanja

- Imamo obrnjen indeks
 - slovar simbolov
 - seznam dokumentov (opcijsko še lokacije pojavitve simbolov)
- Kako najdemo pojavitve vseh besed, ki se začnejo na ka ipd.:
 - ka*, *ka, ka*ki ...
- In če se zmotimo v črki
 - pinbal namesto pinball?

BRUTUS	→	1	2	4	11	31	45	173	174
CAESAR	→	1	2	4	5	6	16	57	132 ...
CALPURNIA	→	2	31	54	101				
⋮									
dictionary		postings							

pinbal

About 53,000,000 results (0.18 seconds)

Showing results for [pinball](#)

Search instead for [pinbal](#)

[Mr Men Pinball - Action Games at Miniclip.com - Play Free Online ...](#)

[www.miniclip.com/games/mr-men-pinball/en/](#) - Cached

... AOL Instant Messenger; Google Talk; Skype. arrows Down to launch the ball, left and right to flip. Activate multiple targets for bonus points in Mr Bump **Pinball!** ...

[pinball](#)

[starskyandhutchmovie.warnerbros.com/pinball/](#) - Cached

[Pinball - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Pinball](#) - Cached

Pinball is a type of arcade game, usually coin-operated, where a player attempts to score points by manipulating one or more metal balls on a playfield inside a ...

[Teagames.com - Play Free Flash Games - Pinball Play](#)

[www.teagames.com/games/pinball/play.php](#) - Cached

Old-school gaming with a retro **pinball** simulation! Reckon you're a **pinball** wizard?

[Waterpark Pinball - Free Online Pinball | Candystand.com](#)



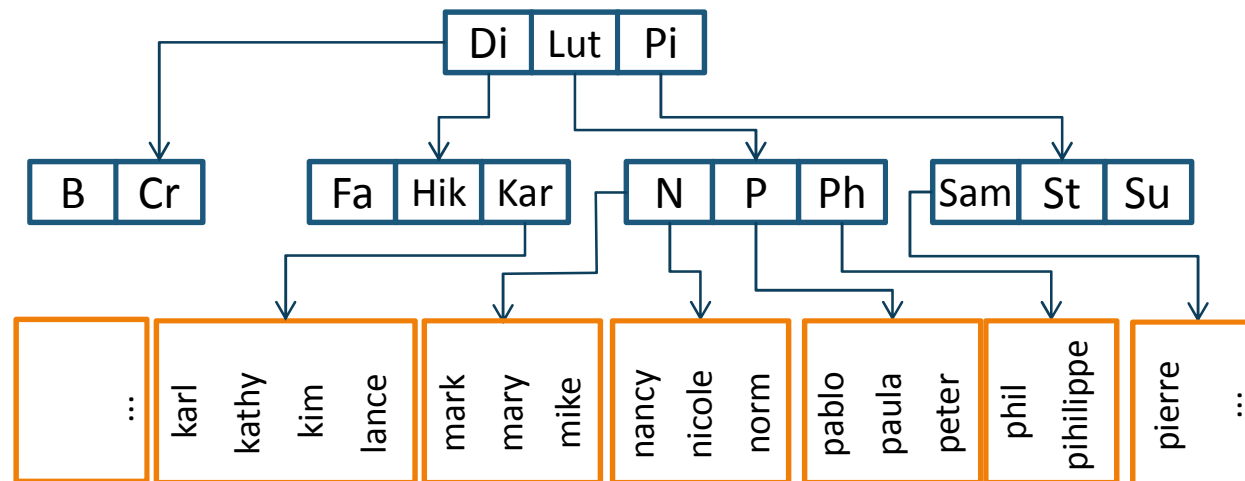
[www.candystand.com/play/waterpark-pinball](#)

21 Jul 2006

Challenge your friends in this action packed **pinball** game. Water Park **Pinball** will keep you playing for ...

[More videos for pinball »](#)

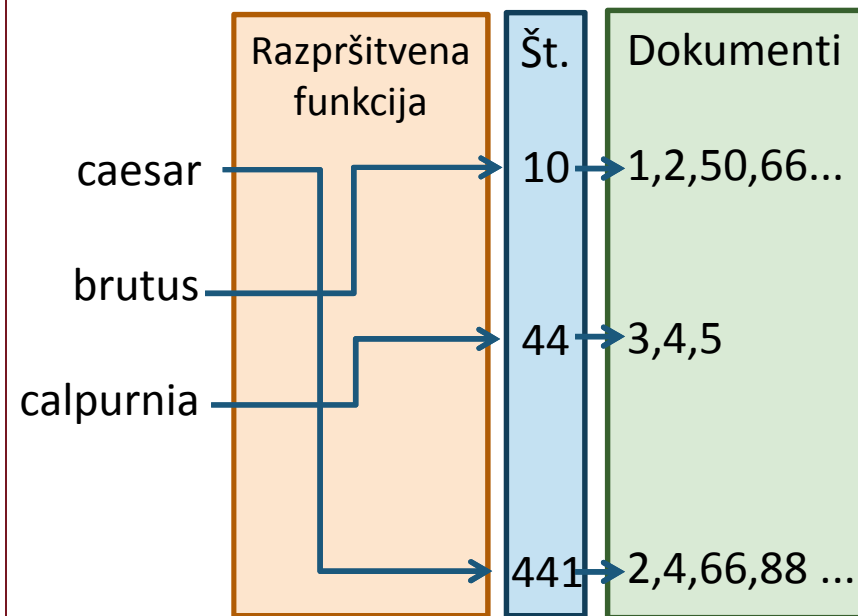
- Vsebuje simbole
- Navadno hranjen v
 - razpršeni tabeli
 - drevesu





- *Hash table*
- Vsak simbol se preslika v (različno) število
- Hitro $O(1)$ iskanje
- Minusi
 - težko poiščemo približne variante, npr cesar/caesar
 - težja implementacija iskanj ka*
 - ko se tabela širi, jo je občasno potrebno na novo preračunati

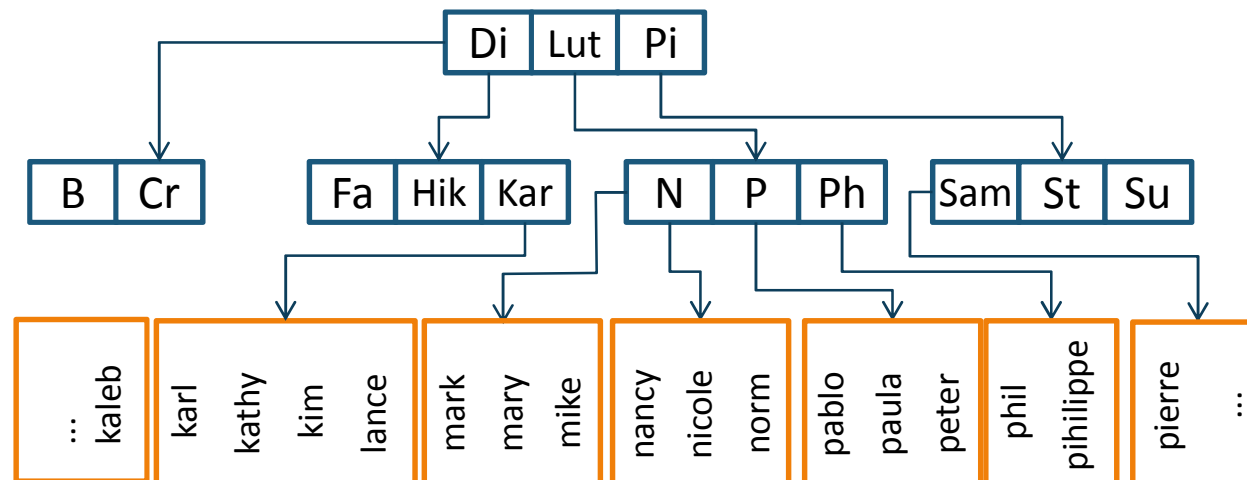
Razpršena tabela





B drevesa

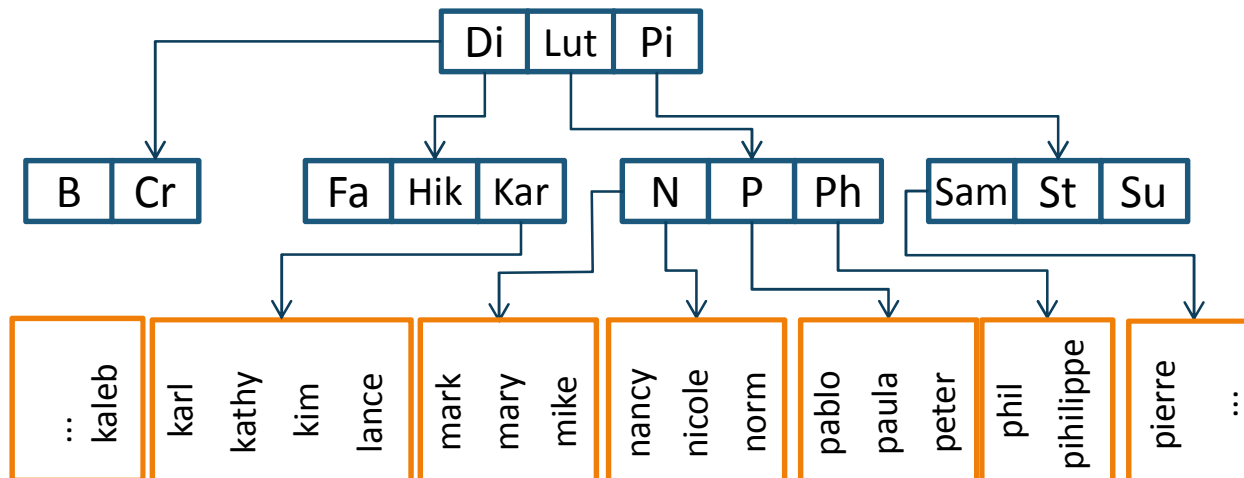
- Urejeno drevo
 - vsako vozlišče ima lahko med več naslednikov (v intervalu $[a,b]$)
 - je uravnoteženo – vsi listi so na enaki globini
- Čas povpraševanja je $O(\log(N))$
- Enostavna implementacija povpraševanj tipa ka^*





B drevesa

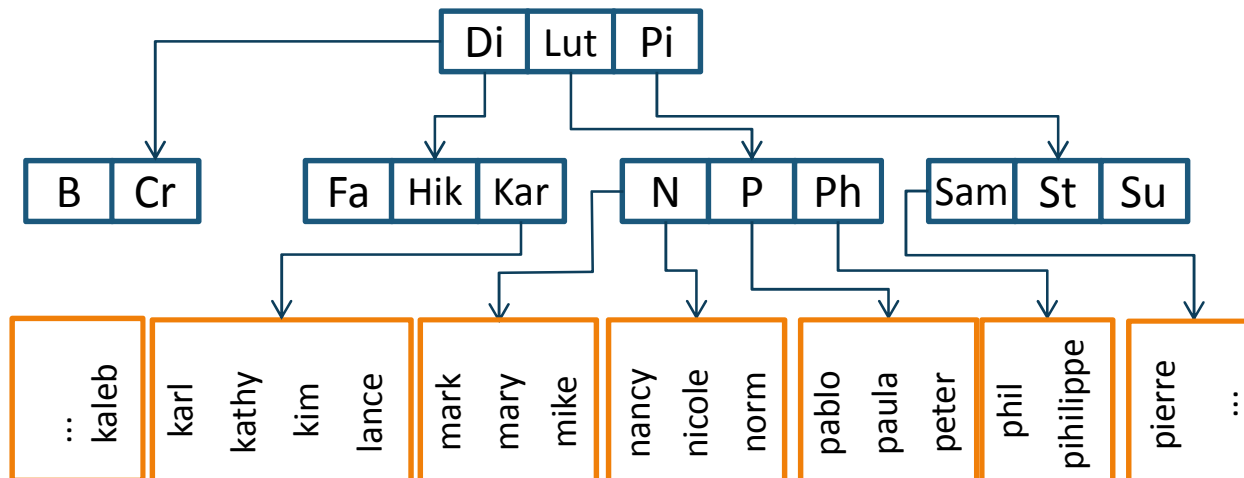
- Kako najdemo ka^*
 - torej vse simbole, ki se začnejo na ka
- V drevesu enostavno, ker je urejeno
 - poiščemo vse simbole w : $ka \leq w < kb$
- Kaj pa $*ka$?
 - vzdržujemo še eno drevo s simboli v obratnem vrstnem redu





B drevesa

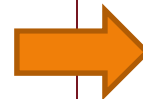
- Kaj pa ka*ki?
 - obravnavamo kot ka* AND *ki, torej naredimo presek vseh simbolov, ki jih najdemo v slovarju
 - časovno zahtevno





Permuterm indeks

- Permuterm indeks omogoča hitrejše približno povpraševanje tipa ka^*ki
- Ideja: besedo indeksiramo v vseh permutacijah
- Povpraševanje obrnemo tako, da je $*$ na koncu:
- Problem je, da se velikost indeksa precej poveča
 - obstajajo tudi druge tehnike, npr. bigrami itn.



npr. kabuki indeksiramo kot

- kabuki\$
abuki\$
buki\$
uki\$
ki\$
i\$
\$
- \$ je posebna oznaka

ka^*ki povprašamo kot kika^*$

- $*ki$ povprašamo kot ki*$
- ki^* povprašamo kot $$ki^*$
- $*ki^*$ povprašamo kot ki^*

Približno povpraševanje in Boolovi izrazi

- Če kombiniramo približna povpraševanja z Boolovimi izrazi, lahko pri iskanju dokumentov dobimo zelo dolge sezname lokacij
 - npr. *gen* universit**: pomeni
 - geneva university OR geneva université OR genève university OR genève université OR general universities OR . . .
 - iskanje postane dolgotrajno
 - uporabniki so leni, kar pomeni da bodo * veliko uporabljali
- Ima google to možnost?
 - * lahko nastopa le namesto cele besede

en*li*h

About 3,100,000,000 results (0.53 seconds)

[David H. Li - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/David_H_Li](#) - Cached

... name; the family name is Li. David H. Li is an author on Chinese history and chess. ... Retrieved from "http://en.wikipedia.org/w/index.php?title=David_H. ...

[David Daokui Li - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/David_Daokui_Li](#) - Cached

28, issue 4, pages 716-738; Gordon, Roger H.; Bai, Chong-En; Li, David D. Efficiency losses from tax distortions vs. government control. European Economic ...

[Wen-Hsiung Li - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Wen-Hsiung_Li](#) - Cached

Wen-Hsiung Li (Traditional Chinese: 李文雄, 1942-) is a Taiwanese American scientist working in the fields of molecular ... Li, W.-H. (2006). ... Retrieved from "http://en.wikipedia.org/w/index.php?title=Wen-Hsiung_Li&oldid=450179082" ...

[LONG ISLAND CITY Hotels - Holiday Inn Hotels & Resorts L.I. CITY ...](#)

[www.holidayinn.com/hotels/us/en/long-island.../hoteldetail](#) - Cached

Official Site for L.I. CITY-MANHATTAN VIEW hotel in LONG ISLAND CITY. Best Price Guarantee or your first night is free! Book early & save, plus earn rewards ...

[PDF] [E EN NL LI IG GH HT TE EN NM ME EN NT ~](#)

[www.kundaliniawakeningsystems1.com/.../chrism_on_-_enlightenme...](#)

File Format: PDF/Adobe Acrobat

E EN NL LI IG GH HT TE EN NM ME EN NT ~. l l s s t t t h h h i i s s w w w h h h a a t t y y y o o u u s s s e e e e e k k k ? ? W. W W h h a a t t d d d o o e e s s i i t t ...

[Huiying Li Ph.D - People - CNSI](#)

[www.cnsi.ucla.edu/institution/personnel?personnel_id...](#) - Cached

Ochoa MT, Teles R, Haas BE, Zaghi D, Li H, Sarno EN, Rea TH, Modlin RL, Lee DJ, A role for interleukin-5 in promoting increased immunoglobulin M at the site ...



- Kaj, če povpraševanje vsebuje napake?
 - npr. pinbal namesto pinball
- Napako želimo odpraviti in uporabniku ponuditi možne variante popravkov
- Popravljamo lahko
 - posamezne **nepravilne besede** npr. pinbal
 - besede , ki so sicer pravilne ampak ne v tem **kontekstu**
 - an asteroid that fell **form** the sky



Popravljanje napak

pinbal

About 53,000,000 results (0.18 seconds)

Showing results for [pinball](#)

Search instead for pinbal

[Mr Men Pinball - Action Games at Miniclip.com - Play Free Online ...](#)

[www.miniclip.com/games/mr-men-pinball/en/](#) - Cached

... AOL Instant Messenger; Google Talk; Skype. arrows Down to launch the ball, left and right to flip. Activate multiple targets for bonus points in Mr Bump **Pinball**! ...

[pinball](#)

[starskyandhutchmovie.warnerbros.com/pinball/](#) - Cached

[Pinball - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Pinball](#) - Cached

Pinball is a type of arcade game, usually coin-operated, where a player attempts to score points by manipulating one or more metal balls on a playfield inside a ...

[Teagames.com - Play Free Flash Games - Pinball Play](#)

[www.teagames.com/games/pinball/play.php](#) - Cached

Old-school gaming with a retro **pinball** simulation! Reckon you're a **pinball** wizard?

[Waterpark Pinball - Free Online Pinball | Candystand.com](#)



[www.candystand.com/play/waterpark-pinball](#)

21 Jul 2006

Challenge your friends in this action packed **pinball** game. Water Park **Pinball** will keep you playing for ...

[More videos for pinball »](#)

Popravljanje posameznih nepravilnih besed

- Predpostavke:
 - imamo **seznam** besed
 - pravih besed – slovar
 - vseh besed v indeksu (niso vse pravilne), ter podatek o pogostosti uporabe
 - znamo izračunati **razdaljo** med besedami
- Tako lahko popravimo povpraševanje, s tem da izberemo najbližjo pravo besedo

pinbal

About 53,000,000 results (0.18 seconds)


Showing results for [pinball](#)
Search instead for pinbal

[Mr Men Pinball - Action Games at Miniclip.com - Play Free Online ...](#)
[www.miniclip.com/games/mr-men-pinball/en/](#) - Cached
... AOL Instant Messenger; Google Talk; Skype. arrows Down to launch the ball, left and right to flip. Activate multiple targets for bonus points in Mr Bump **Pinball!** ...

[pinball](#)
[starskyandhutchmovie.warnerbros.com/pinball/](#) - Cached

[Pinball - Wikipedia, the free encyclopedia](#)
[en.wikipedia.org/wiki/Pinball](#) - Cached
Pinball is a type of arcade game, usually coin-operated, where a player attempts to score points by manipulating one or more metal balls on a playfield inside a ...

[Teagames.com - Play Free Flash Games - Pinball Play](#)
[www.teagames.com/games/pinball/play.php](#) - Cached
Old-school gaming with a retro **pinball** simulation! Reckon you're a **pinball** wizard?

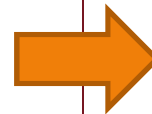
[Waterpark Pinball - Free Online Pinball | Candystand.com](#)
 [www.candystand.com/play/waterpark-pinball](#)
21 Jul 2006
Challenge your friends in this action packed **pinball** game. Water Park **Pinball** will keep you playing for ...

[More videos for pinball »](#)



Računanje razdalj med besedami

- Popravna razdalja (*edit distance*)
 - koliko (najmanj) operacij potrebujemo da niz s_1 pretvorimo v s_2
- Levenshteinova razdalja
 - tri operacije:
 - vstavljanje črke
 - brisanje črke
 - spreminjanje črke
- Damerau-Levenshtein razdalja
 - še zamenjavo sosednjih črk



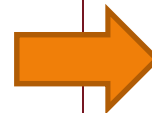
Levenshtein:

dog-do: 1

cat-cart: 1

cat-cut: 1

cat-act: 2



Damerau-Levenshtein:

cat-act: 1



Levenshteinova razdalja

- Gradimo matriko vseh možnih pretvorb iz ene besede v drugo
- V celicah matrike (i,j) seštevamo cene operacij potrebnih za transformacijo niza (1..j) v drugega (1..i)
 - če sta znaka (i,j) enaka:
 - $m[i,j] = \min(m[i-1,j]+1, m[i,j-1]+1, m[i-1,j-1])$
 - če znaka nista enaka
 - $m[i,j] = \min(m[i-1,j]+1, m[i,j-1]+1, m[i-1,j-1]+1)$

Razdalja med
fast in *cat* je 2

		f	a	s	t
	0	1	2	3	4
c	1	1	2	3	4
a	2	2	1	2	3
t	3	3	2	2	2

vstavljanje – cena == 1

brisanje – cena == 1

zamenjava – cena == 1

kopiranje – cena == 0

Demo

Levenshteinova razdalja

LEVENSHTEINDISTANCE(s_1, s_2)

```
1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7      do if  $s_1[i] = s_2[j]$ 
8          then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9          else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 
```

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

Levenshteinova razdalja

- Algoritem je primer **dinamičnega programiranja**, kjer v matriki akumuliramo ceno vseh možnih poti, da najdemo rešitev
- Lahko bi vpeljali še **uteži** za posamezne operacije
 - npr. menjava med m in n je bolj verjetna (ima manjšo ceno) kot med m in q
- Ali uvedli **dodatne operacije**, kot npr. zamenjava sosednjih črk
 - dobimo algoritem Damerau-Levenshtein
- Slednjega [uporablja Google](#) za popravljanje povpraševanj
 - potrebno je še smiselno indeksiranje, da lahko hitro izračunamo razdaljo med povpraševanjem in vsemi simboli v indeksu

		f	a	s	t
	0	1	2	3	4
c	1	1	2	3	4
a	2	2	1	2	3
t	3	3	2	2	2

Operacije za spremembo lahko preberemo s sledenjem v matriki od zadaj naprej, gledamo kateri minimum je bil izbran.

Primer: cat -> fast

kopiraj *t*

dodaj *s*

kopiraj *a*

spremeni *c* v *f*



Soundex

- Najde **fonetično** razdaljo med besedami
 - kako so si podobne ob (angleški) izgovorjavi
 - npr. chebyshev / tchebyscheff
- Definiran kot nabor pravil, ki preslika niz v štiriznakovno kodo
 - indeksiramo to kodo
- Za splošno iskanje ni preveč smislen
 - je pa za posebne aplikacije, kjer je važno, da najdemo čimveč pravih zadetkov (poleg njih lahko tudi veliko nepravih)
 - npr. iskanje po policijskih arhivih ipd.

Demo

HERMAN:

obdržimo H

ERMAN → *ORMON*

ORMON → *06505*

06505 → *06505*

06505 → *655*

Rezultat: H655

Za *HERMANN* bomo dobili enako kodo



RANGIRANJE DOKUMENTOV



- Dokumentov, ki ustrezajo iskalnim pogojem, je lahko veliko
- Rangiranje prikaže **najpomembnejše** dokumente najprej



Rangiranje dokumentov

large caterpillar species

About 4,250,000 results (0.24 seconds)

[Stinging Caterpillars: A Guide to Recognition of Species Found on ...](#)
[www.ag.auburn.edu/enpl/bulletins/caterpillar/caterpillar.htm](#) - Cached

Among the members of this family are some of the largest and most striking and fearsome-looking of our native **caterpillars**. Some **species** spin **large**, thick ...

[What's this caterpillar?!](#)

[www.texasento.net/rearing.htm](#) - Cached

Here are information links to a few **species** of **caterpillars** that have periodic outbreaks: Juniper Budworm in ... Is your **caterpillar** particularly **large**, bordering on ...

[IPM1019 Caterpillars in Your Yard and Garden | Page 9](#)

[extension.missouri.edu > ... > Integrated pest management](#) - Cached

These **large caterpillar species** are usually not considered pests. Although a single individual can consume relatively large amounts of foliage, their numbers ...

[Saturniid Moths | University of Kentucky Entomology](#)

[www.ca.uky.edu/entomology/entfacts/ef008.asp](#) - Cached

10 Jan 2010 – One or two of the **larger types** of **caterpillars** can cause severe defoliation. Hand picking these **caterpillars** is sufficient control, but wear gloves if ...

[Identifying Australian Caterpillars](#)

[lepidoptera.butterflyhouse.com.au/faqs/ident.html](#) - Cached

The Identification of **Caterpillars** of Australia. Don Herbison-Evans ... SPECIAL CATERPILLAR SPECIES ... Big red rump: Day-Flying Moths AGARISTINAE ...

[Forest Tent Caterpillars in Minnesota](#)

[www.extension.umn.edu/distribution/horticulture/dg7563.html](#) - Cached

It periodically infests aspen and many other tree **species** over **large** areas of northern Minnesota. The **caterpillars** are commonly, but mistakenly, called ...

Rangiranje dokumentov

- Kako lahko izračunamo rang?
- Primer: Jaccardov koeficient za množici A in B:
- Primer:
 - Povpraševanje: “ides of March”
 - Dokument “Caesar died in March”
 - $JACCARD(q, d) = 1/6$
- Problemi:
 - ne upošteva kolikokrat se simboli pojavijo v dokumentu

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



- *Term frequency*
- $tf_{t,d}$ - število pojavitev simbola t v dokumentu d
- Samo število pojavitev ni dobra mera
 - 1000 pojavitev ne pomeni da je dokument 1000x bolj pomemben
- Uvedemo neko funkcijo števila pojavitev, npr. logaritmično
- Za par povpraševanje, dokument seštejemo po vseh simbolih:

$$tfScore(q, d) = \sum_{t \in q \cap d} 1 + \log_{10} tf_{t,d}$$

Pogostost simbolov

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d} & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$tf_{t,d} \rightarrow w_{t,d}$:

$0 \rightarrow 0, 1 \rightarrow 1, 2 \rightarrow 1.3, 10 \rightarrow 2, 1000 \rightarrow 4 \dots$

Inverzna pogostost v dokumentih

- *Inverse document frequency*
- Besede, ki se pogosto pojavljajo v dokumentih niso zelo pomembne
 - npr. *good*
 - želimo, da imajo nizko utež
- Besede, ki so redke, so bolj pomembne
 - npr. *arachnocentric*
 - želimo visoko utež
- *Document frequency* df_t
 - število dokumentov, v katerih se simbol t pojavi
- *Inverse document frequency*
 - izračunamo kot je podano, kjer je N število vseh dokumentov
- Če imamo povpraševanje iz ene besede, *idf* ni relevantna mera
- Če iz več besed, npr. *arachnocentric person*, je utež na *arachnocentric* večja kot na *person*

$$idf_t = \log_{10} (N/df_t)$$

term	df_t	idf_t
calpurnia	1	6
animal	100	4
sunday	1.000	3
fly	10.000	2
under	100.000	1
the	1.000.000	0

$N=1.000.000$



tf-idf

- *term-frequency – inverse-document frequency*
- Produkt *tf* in *idf* mer
- Najboljša mera za uteževanje
- Utež se povečuje s številom pojavitev simbola v dokumentu in z redkostjo simbola

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \log_{10} \frac{N}{df_t}$$





tf-idf

Povpraševanje: *animals fly on sunday*
 $N=1.000.000$

term	df_t
animal	100
sunday	1.000
fly	10.000
on	1.000.000

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \log_{10} \frac{N}{df_t}$$

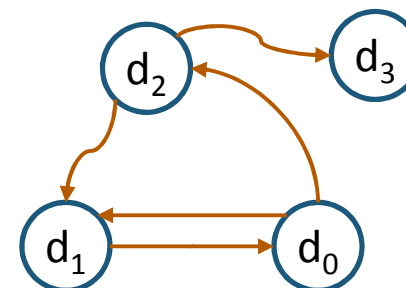
term	$tf_{t,1}$	$tf_{t,2}$	$tf_{t,3}$	$tf_{t,4}$
animal	1	1	3	0
fly	1	3	1	3
on	1	6	0	10
sunday	1	3	3	0

term	tf-idf	tf-idf	tf-idf	tf-idf
animal	4	4	6	0
fly	2	3	2	3
on	0	0	0	0
sunday	3	4.5	4.5	0
skupaj	9	11.5	12.5	3



- Rangiranje ala Google:
 - strani, na katere kaže **več povezav**, so bolj **obiskane**
 - so bolj pomembne
 - povezave iz bolj pomembnih strani naj bodo bolj pomembne
 - in manj kot je povezav iz strani, bolj so pomembne
 - podobno kot citati v znanstveni literaturi
 - več ko ima članek citatov, bolj je pomemben
- PageRank lahko računamo z modeliranjem **naključnega sprehajalca** po spletu
 - verjetnost, da sprehajalec izbere katerokoli od k povezav iz neke strani naprej je enaka $(1/k)$
 - če stran nima povezav naprej, izbere katerokoli stran (če je skupaj N strani, z verjetnostjo $1/N$)
 - vedno lahko z neko verjetnostjo d skoči na katerikoli drugo stran (npr. neposredno vtipka URL v brskalnik)

PageRank



Enostaven primer, število strani $N=4$

	d_0	d_1	d_2	d_3
d_0	0	0.5	0.5	0
d_1	1	0	0	0
d_2	0	0.5	0	0.5
d_3	0.25	0.25	0.25	0.25

Verjetnosti prehodov med stranmi (matrika M)

	d_0	d_1	d_2	d_3
d_0	0.05	0.45	0.45	0.05
d_1	0.85	0.05	0.05	0.05
d_2	0.05	0.45	0.05	0.45
d_3	0.25	0.25	0.25	0.25

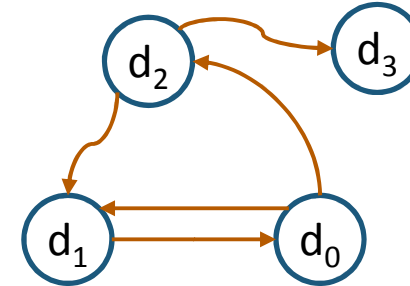
Popravljenе verjetnosti prehodov, če je verjetnost naključnega skoka $d=0.2$
 $M=d/N+(1-d)M$



PageRank

- Računanje PageRank

- obiskovalec naključno izbere neko stran, npr. z verjetnostmi $t=[0.2 \ 0.4 \ 0.1 \ 0.3]$
- po eni izbiri povezave ali sprehodu na naključen URL, bodo verjetnosti strani:
$$t'=t*M$$
- po dveh izbirah povezav ali sprehodu na naključen URL:
$$t'=t*M*M$$
- po n izbirah povezav ali sprehodih na naključen URL
$$t'=t*M^n$$
- vrednost t' se z večanjem n ustali na stabilni vrednosti (konvergira)
 - to je PageRank – bolj pomembne strani imajo višjo vrednost



Enostaven primer, število strani $N=4$

$$M = \begin{bmatrix} 0.05 & 0.45 & 0.45 & 0.05 \\ 0.85 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.45 & 0.05 & 0.45 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \quad t = \begin{bmatrix} 0.2 & 0.4 & 0.1 & 0.3 \end{bmatrix}$$

po 10 izbirah povezav

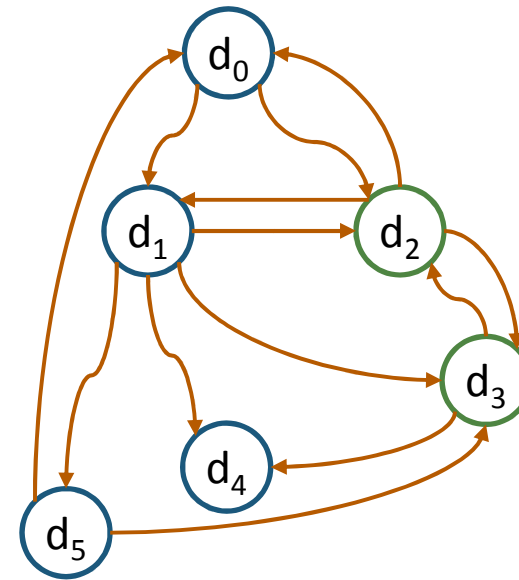
t'	0.32	0.3	0.21	0.17
----	------	-----	------	------



PageRank

- Splet = povezan graf
 - povezave kažejo iz dokumentov na druge
- Če
 - na stran A kažejo strani $B_1 \dots B_n$
 - $C(B)$ je število povezav, ki gredo iz strani B
 - d je neka konstanta, npr. 0.85 – pomeni verjetnost, da bo oseba nadaljevala s klikanjem po povezavah
 - N je število vseh strani
 - je PageRank $PR(A)$

$$PR(A) = \frac{(1-d)}{N} + d \sum_{i=1}^n \frac{PR(B_i)}{C(B_i)}$$



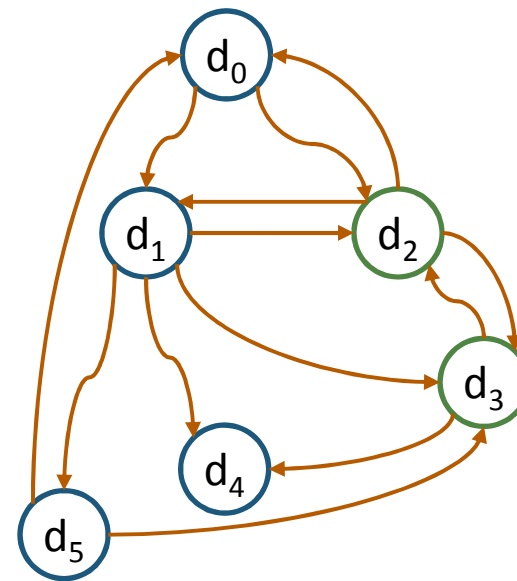
Primer :

$$PR(d_0) = (1-c) + c\left(\frac{PR(d_2)}{3} + \frac{PR(d_5)}{2}\right)$$



PageRank

- PageRank dejansko modelira naključnega sprehajalca, ki se sprehaja po spletu in
 - začne na naključni strani B
 - vsakič naključno izbere eno izmed povezav iz strani B z verjetnostjo $d/C(B)$
 - ali gre na katerokoli drugo naključno stran z verjetnostjo $(1-d)/N$
 - če stran nima povezav naprej, izbere naključno stran z verjetnostjo $1/N$





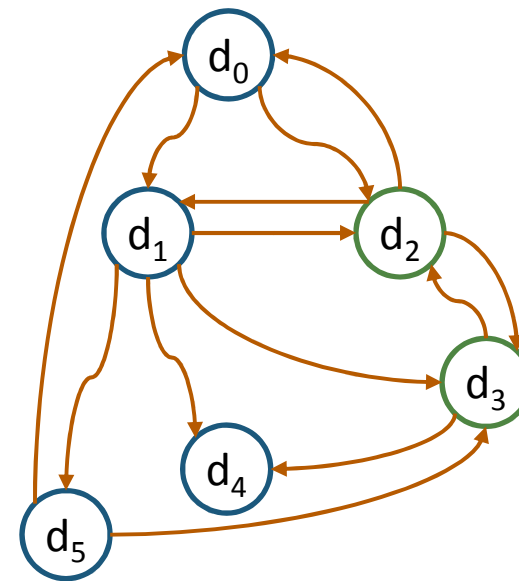
PageRank

- Verjetnosti prehodov med povezavami

	d ₀	d ₁	d ₂	d ₃	d ₄	d ₅
d ₀	0	0.5	0.5	0	0	0
d ₁	0	0	0.25	0.25	0.25	0.25
d ₂	0.33	0.33	0	0.33	0	0
d ₃	0	0	0.5	0	0.5	0
d ₄	0	0	0	0	0	0
d ₅	0.5	0	0	0.5	0	0

- Verjetnosti popravljene z verjetnostjo naključnega skoka
 $d=1-0.85$: $M=d*M+(1-d)/N$

	d ₀	d ₁	d ₂	d ₃	d ₄	d ₅
d ₀	0.025	0.45	0.45	0.025	0.025	0.025
d ₁	0.025	0.025	0.2375	0.2375	0.2375	0.2375
d ₂	0.3083	0.3083	0.025	0.3083	0.025	0.025
d ₃	0.025	0.025	0.45	0.025	0.45	0.025
d ₄	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
d ₅	0.45	0.025	0.025	0.45	0.025	0.025



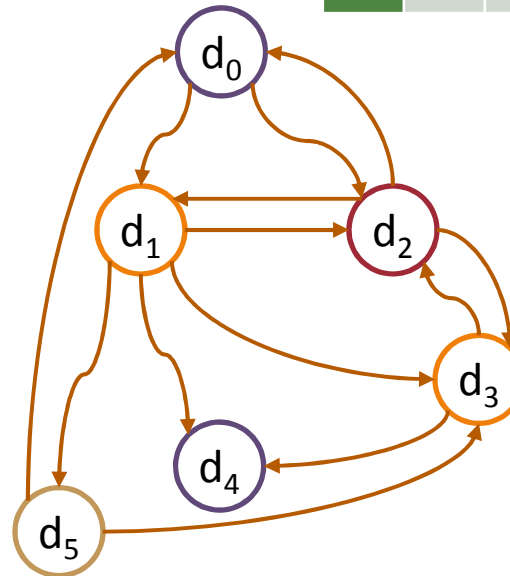


■ Primer računanja PageRanka:

- naključno izberemo verjetnosti, da sprehajalec izbere neko stran
 - $t = \text{rand}(1, N)$; $t = t / \text{norm}(t, 2)$;
- prehod na naslednjo stran je $t * M$
 - na naslednjo $(t * M) * M$
 - itn...
- množenje ponavljamo do konvergence
 - dobimo stabilno verjetnost, da bo sprehajalec končal na neki strani
- dobimo PageRank

PageRank

	d_0	d_1	d_2	d_3	d_4	d_5
rank	0.357	0.422	0.546	0.447	0.394	0.205



	d_0	d_1	d_2	d_3	d_4	d_5
d_0	0.025	0.45	0.45	0.025	0.025	0.025
d_1	0.025	0.025	0.2375	0.2375	0.2375	0.2375
d_2	0.3083	0.3083	0.025	0.3083	0.025	0.025
d_3	0.025	0.025	0.45	0.025	0.45	0.025
d_4	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
d_5	0.45	0.025	0.025	0.45	0.025	0.025



- Nam da pomembnost strani, ki jo lahko uporabimo za rangiranje
- Ni vsemogočen
- Zato ga Google kombinira z ostalimi kriteriji - najdene besede v naslovu, povezavah, velikost pisave ...
- Problemi, zlorabe
 - Link Spam – postavljanje strani, ki vsebujejo povezave na neko drugo stran za zvišanje PageRank-a
 - Google bombs – zviševanje ranga s spreminjanjem teksta povezav
 - npr. "miserable failure" je prikazalo stran G.W. Busha
 - l. 2007 je Google popravil iskalnik, da je preprečil tovrstne primere
 - se še vedno pojavljajo



PageRank

miserable failure

About 1,340,000 results (0.18 seconds)

[Google bomb - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Google_bomb](#) - Cached

However, by January 2007 Google had made changes to search results to counter popular Google bombs, such as "**miserable failure**", which now lists pages ...

[History](#) - [Uses as tactical media](#) - [Alternative meanings](#) - [Google bowling](#)

[Political Google bombs in the 2004 U.S. Presidential election ...](#)

[en.wikipedia.org/.../Political_Google_bombs_in_the_2004_U...](#) - Cached

Screenshot of the Google search results for "**Miserable Failure**" in March 2007 ... Two of the first were the "**miserable failure**" Google bomb linked to George W. ...

[Google Kills Bush's Miserable Failure Search & Other Google Bombs](#)

[searchengineland.com/google-kills-bushs-miserable-...](#) - Cached



by Danny Sullivan · in 162,486 Google+ circles · [More by Danny Sullivan](#)

25 Jan 2007 – Google has finally defused the "Google Bomb" that has returned US President George W. Bush at the top of its results in a search on **miserable** ...

[snopes.com: Miserable Failure](#)

[www.snopes.com/politics/bush/google.asp](#)

13 Aug 2007 – Why is the phrase '**miserable failure**' tied to President Bush's biography in Google?

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)

[news.bbc.co.uk/2/hi/3298443.stm](#) - Cached

7 Dec 2003 – Web users manipulate a popular search engine so an unflattering description leads to the president's page.

[Miserable Failure - Funny Bush Picture](#)

[politicalhumor.about.com/library/.../blbushmiserablefailure.ht...](#) - Cached

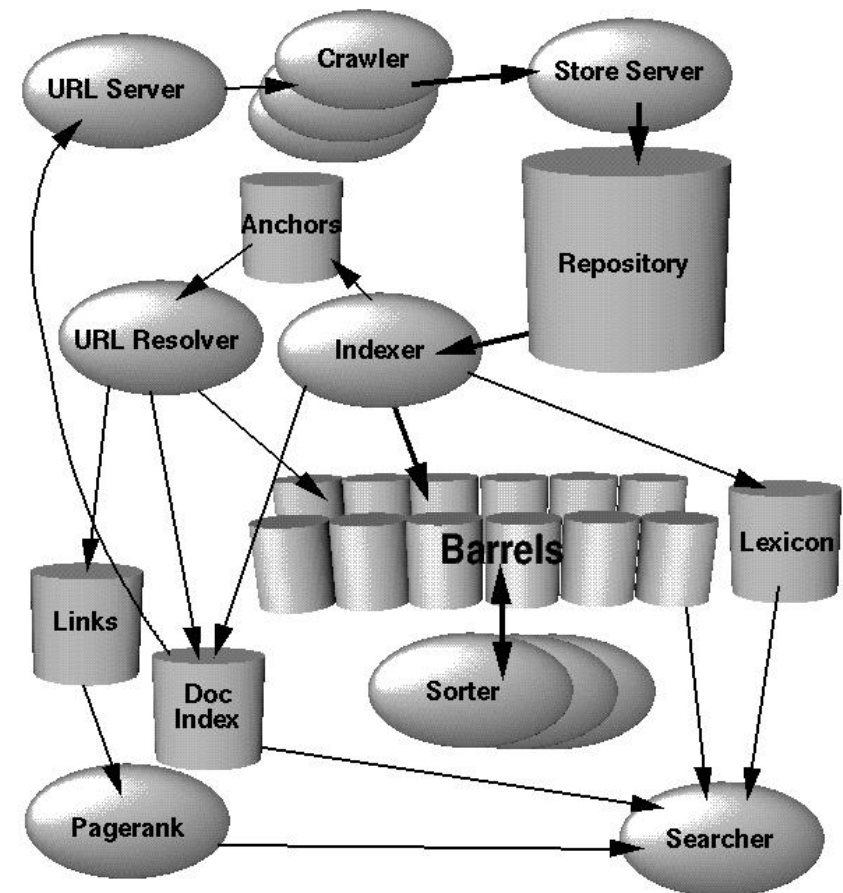
George W. Bush stands on the USS Lincoln under a banner that says **Miserable Failure**.



GOOGLE INTERNALS

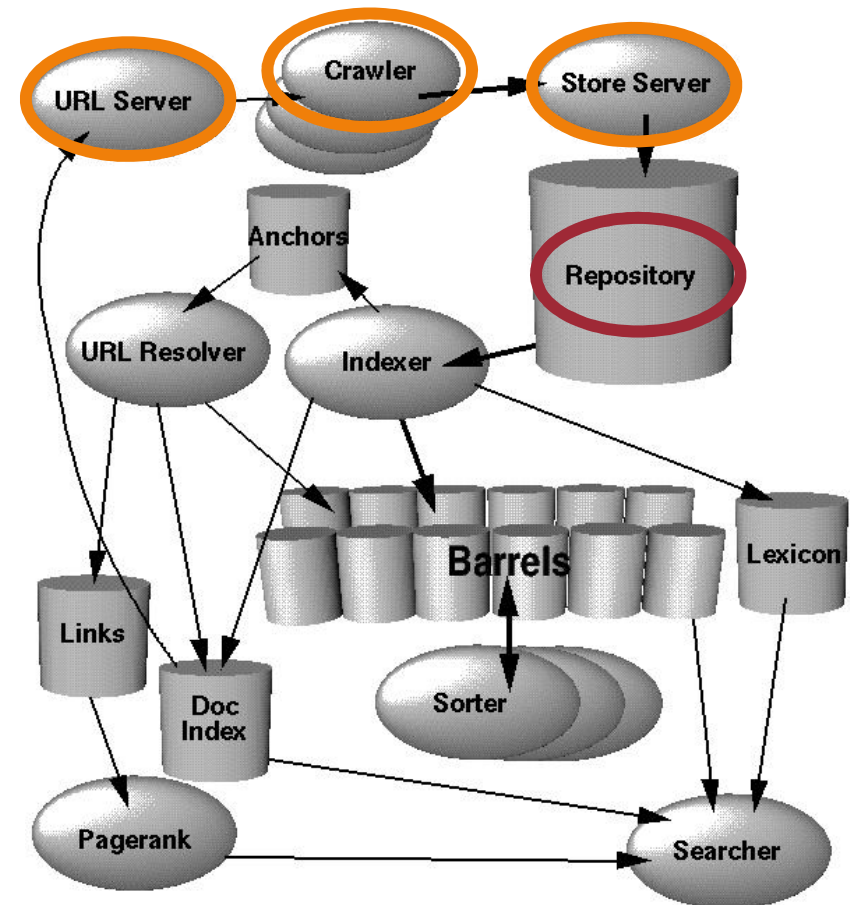
Kako dejansko dela Google?

- Brin, Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine



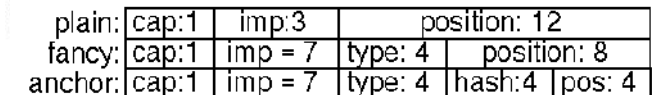
Kako dejansko dela Google?

- **URL Server** sestavlja sezname URLjev
 - na podlagi povezav v obstoječih dokumentih oz. novih povezav
- Sezname da (porazdeljenemu) web **crawlerju**, ki prenese dokumente s spleta
- **Store Server** prenešene dokumente stisne in jih shrani v **repozitorij**
 - shranjene strani vsebujejo celoten HTML
 - stisnjene z bzip-om
 - poleg strani se hrani še docID, dolžina in URL, torej:
docID, length, URL, page



1000000

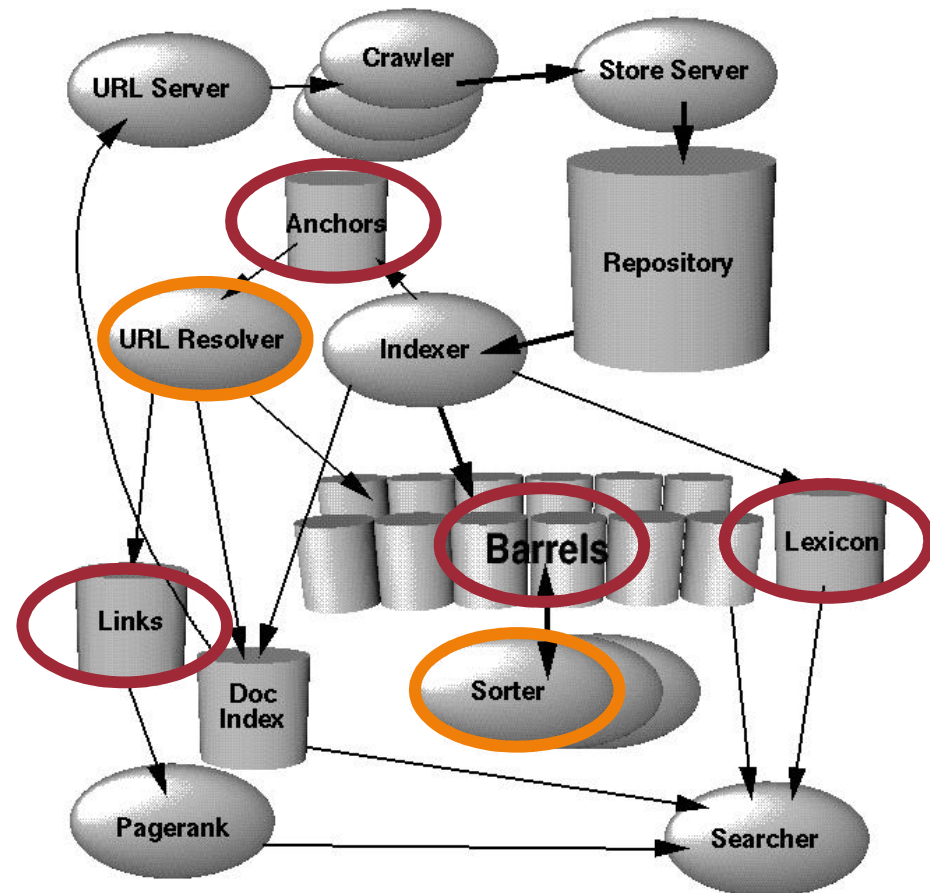
-



docid	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	null wordid		
docid	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	wordid: 24	nhits: 8	hit hit hit hit
	null wordid		

Kako dejansko dela Google?

- **URL Resolver** prebere povezave v **anchors**
 - jih pretvori v IDje dokumentov
 - doda pare povezanih docID-jev v Links
 - se uporabijo za rangiranje dokumentov
- **Sorter** prebere *forward index* zapisan v *barrels*
 - ta je urejen po docID
 - preuredi ga z urejanjem po wordID-jih, da dobi **obrnjen index**
 - zaradi lažjega rangiranja sta dva obrnjena indeksa,
 - kratek za besede v naslovih in povezavah
 - eden za ostale besede
 - v slovar besed doda povezave na besede v obrnjenem indeksu



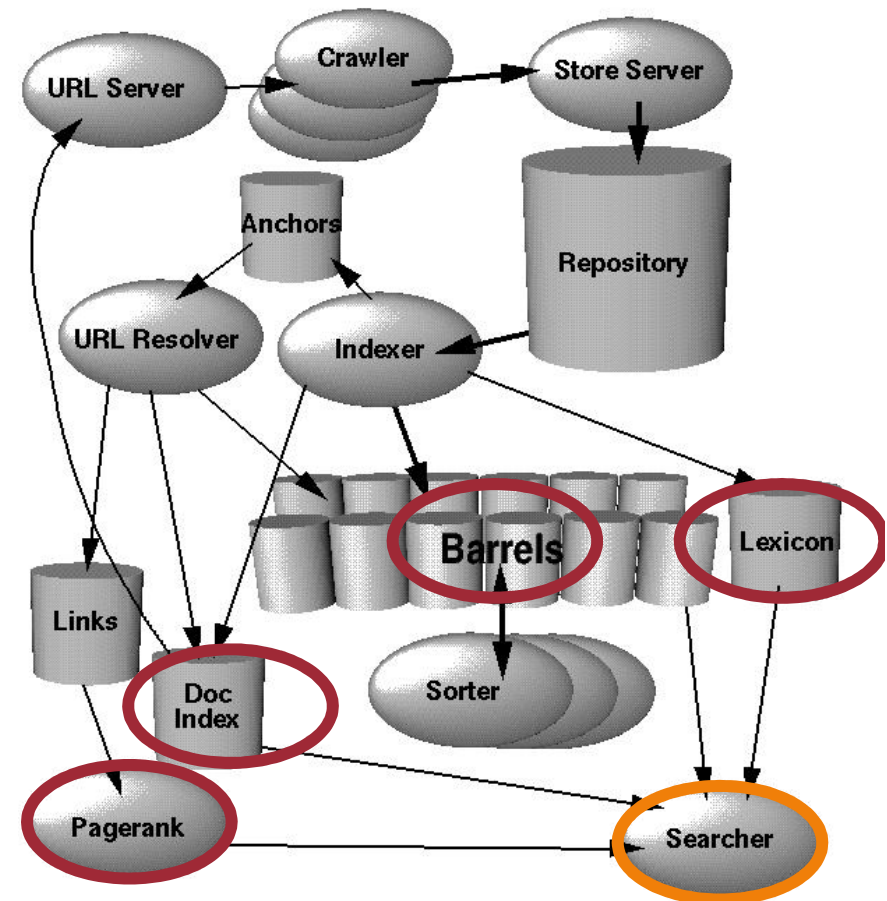
Lexicon: 293MB

Inverted Barrels: 41 GB

wordid	ndocs		docid: 27	nhits:5	hit hit hit h
wordid	ndocs		docid: 27	nhits:5	hit hit hit
wordid	ndocs		docid: 27	nhits:5	hit hit hit h
			docid: 27	nhits:5	hit hit

Kako dejansko dela Google?

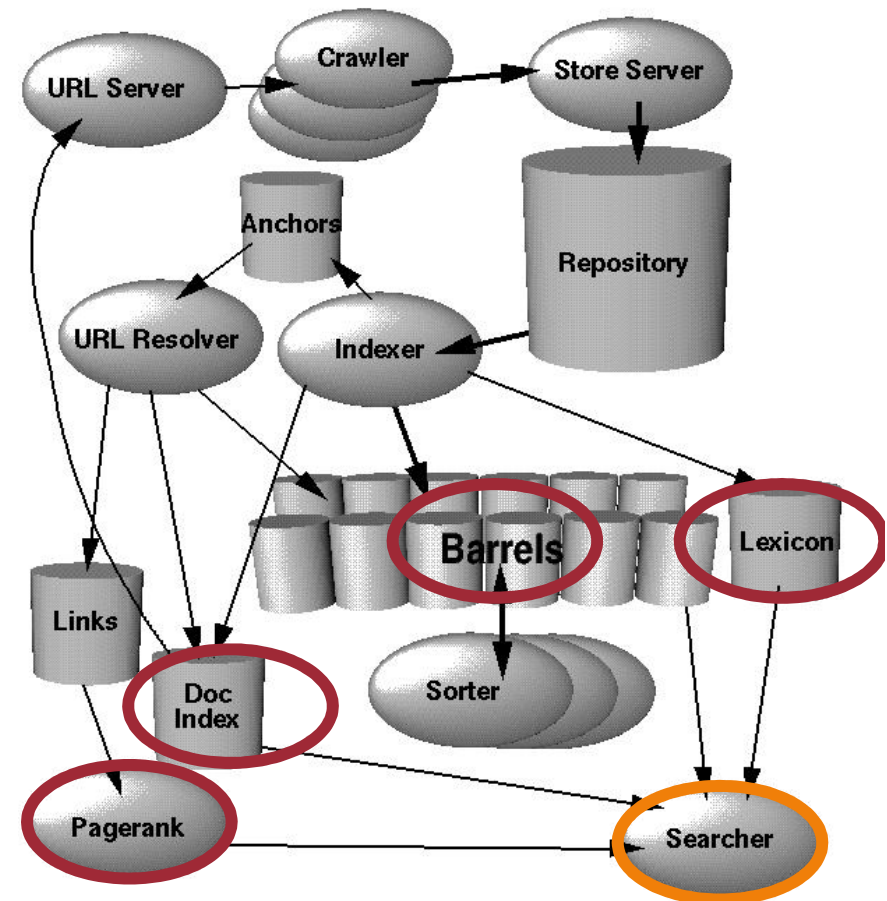
- **Searcher** vrne rezultate na podlagi
 - slovarja
 - obrnjenega indeksa
 - najprej indeksa besed v naslovih in povezavah
 - potem še drugega indeksa
 - rezultatov *pagerank* algoritma
- **Koraki:**
 1. Izločanje simbolov iz povpraševanja
 2. Pretvorba simbolov v wordID-je
 3. Pregled kratkega obrnjenega indeksa, dokler ni najden dokument, ki vsebuje vse simbole
 4. Izračun ranga za dokument
 5. Če je kratek obrnjen indeks pregledan, vzamemo celoten indeks
 6. Če še nismo na koncu indeksa, gremo na 3
 7. Urejanje vrnjenih dokumentov po rangi



Kako dejansko dela Google?

■ Rangiranje dokumentov

- zadetki imajo podatke o tipu (naslov, povezava, URL, velikost pisave)
 - vsak tip ima neko utež
- glede na število in tip zadetkov v dokumentu se izračuna **utež**
 - če je več besed, se upošteva tudi kako blizu so si različni zadetki
- izračunana utež se kombinira z rezultati PageRank algoritma, ki da končen rang dokumenta





ISKANJE V VEKTORSKEM PROSTORU

Iskanje v vektorskem prostoru

- *Vector space model*
- Dokument predstavimo kot **vektor v vektorskem prostoru**
 - dimenzija vektorja je enaka številu simbolov (besed) v vseh dokumentih
- V vektorju so neničelne vrednosti pri simbolih, ki jih dokument vsebuje
- Vsi vektorji tvorijo matriko
 - *term-document matrix*
- Ni važen vrstni red simbolov
 - *bag of words*
- Dokumenti pri spletnih indeksih imajo lahko milijone dimenzij
 - ker je toliko različnih indeksiranih simbolov

	d1	d2	d3	d4	d5	d6
car	1	1	0	1	1	0
drive	1	1	0	1	0	0
banana	0	1	0	0	0	1
eat	0	1	0	0	0	1
sports	0	0	1	0	1	0
food	0	0	1	0	0	1
baseball	0	0	1	0	1	1

Lahko zapišemo le pojavitev besed

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1

V matriko lahko zapišemo tudi število pojavitev besed ali pa vrednosti mere *tf-idf*

Iskanje v vektorskem prostoru

- Dokumente lahko razvrstimo glede na **razdaljo** med povpraševanjem in dokumentom
 - povpraševanje obravnavamo kot dokument - vektor
 - nič več Boolovega povpraševanja
- Kakšna naj bo razdalja?
 - Evklidska:
$$\text{dist}(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_i (a_i - b_i)^2}$$
 - ni dobra izbira
 - dokumenti z različnimi besedami so daleč narazen, čeprav je del besed skupnih
 - dokument, pri katerem se ena beseda večkrat pojavi je lahko dlje, kot dokument, ki nima te besede

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1

	d2	d3	d4	d5	d6
d1	2	3.6	11.4	3.7	5.6

Evklidska razdalja d1 do ostalih dok.



Iskanje v vektorskem prostoru

- Razdaljo raje merimo glede na **kot** med vektorji
 - velikostni red uteži ni tako važen
 - npr. če dokument d podvojimo, $d' = dd$, potem je kot med njima še vedno 0

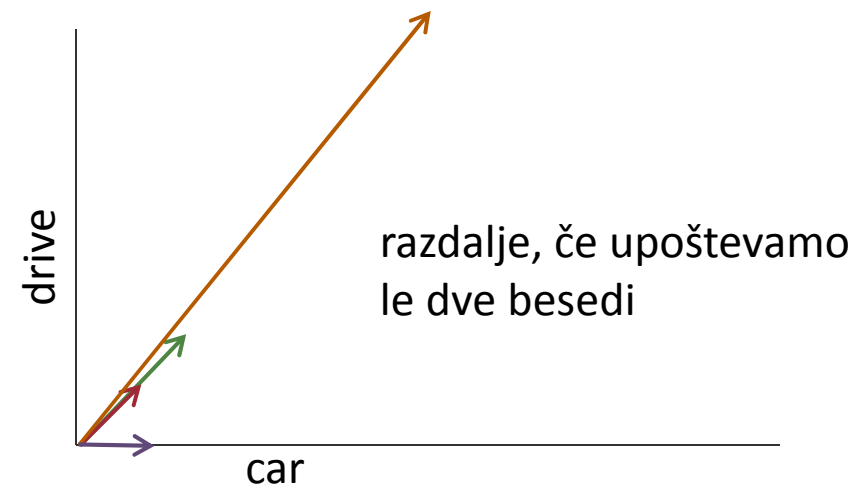
- Kosinusna **podobnost**

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}| |\mathbf{b}|} = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}}$$

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1


	d2	d3	d4	d5	d6
d1	0.11	1	0.01	0.81	1

Kosinusna razdalja d1 do ostalih dok.





Iskanje v vektorskem prostoru

- Povpraševanje predstavimo z tf-idf vektorjem
 - Dokumente tudi
 - Izračunamo kosinusno podobnost med povpraševanjem in dokumenti
 - Razvrstimo rezultate, vrnemo prvih K
-
- 
- V realnosti lahko uporabimo različne uteži za povpraševanja in dokumente
 - dokumenti: npr. logaritmični tf, brez idf, normirani vektorji
 - povpraševanje: logaritmični tf, z idf, brez normiranja



LATENT SEMANTIC INDEXING



Latent Semantic Indexing

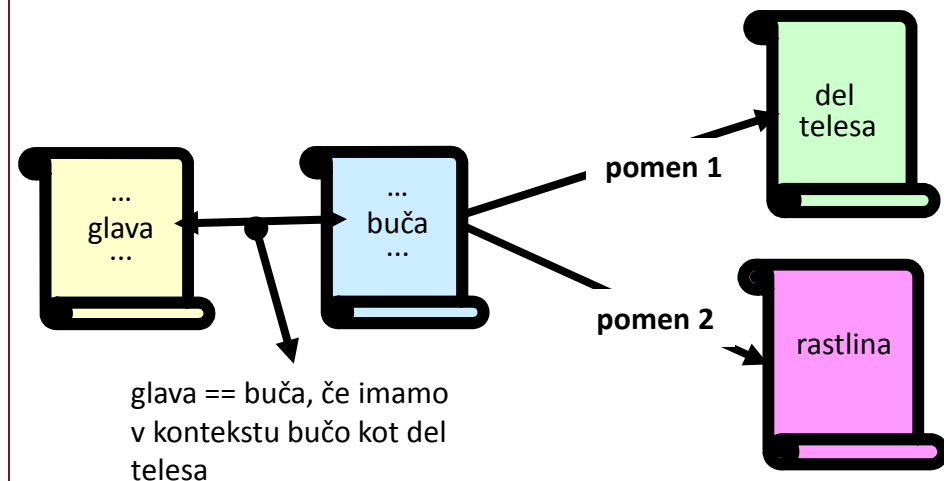
- Matrike, v katerih so dokumenti predstavljeni kot vektorji simbolov so lahko zelo velike
 - vektorji imajo toliko dimenzij kot je simbolov (lahko milijone)
- Po drugi strani lahko dokumenti opisujejo le omejeno število tem
 - šport, politika ...
 - primer, primer
- Bi lahko dokumente predstavili z manj dimenzijami?
 - vsak dokument bi bil vektor, ki opisuje pripadnost nekim temam



• Answers	• Local	• Shine
• Autos	• Maps	• Shopping
• Finance	• Movies	• Sports
• Games	• Music	• Travel
• Groups	• News	• TV
• Health	• omg!	• Voices
• International	• Real Estate	• Yahoo! en Español

Večpomenke, sopomenke

- Standardni model iskanja v vektorskem prostoru ne upošteva večpomenk, sopomenk
 - so vedno različne besede
- Večpomenke
 - beseda ima več pomenov
 - npr. saturn je avto, trgovina, planet ...
- Sopomenke
 - različne besede imajo isti pomen
 - glava, buča
- Bi lahko to kako upoštevali?



Latent Semantic Indexing

- Poskušamo ugotoviti glavne “teme” v dokumentih?
 - matriko A razstavimo na:
 - $A = U \Sigma V^T$
 - singularni razcep (*singular value decomposition*)
 - A – matrika MxN
 - U – matrika MxM
 - Σ – matrika MxN
 - V – matrika NxN
 - s podrobnostmi izračuna s ne bomo ukvarjali (glej npr. [wiki](#))

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1

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

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix} \begin{bmatrix} \bullet & & \\ & \bullet & \\ & & \bullet \end{bmatrix} \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$$

$\underbrace{\hspace{1.5cm}}_A \quad \underbrace{\hspace{1.5cm}}_U \quad \underbrace{\hspace{1.5cm}}_{\Sigma} \quad \underbrace{\hspace{1.5cm}}_{V^T}$

Latent Semantic Indexing

- Kaj pomeni $A=U\Sigma V^T$?
- Geometrijsko
 - U: stolpci določajo osi, ki najbolj “zajamejo” dane podatke
 - dimenzija je enaka št. besed
 - osi so pravokotne
 - Σ : elementi na diagonalni določajo “razpon” vrednosti na oseh in s tem pomembnost
 - bolj pomembne osi imajo večji razpon
 - elementi so urejeni – najpomembnejši najprej

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1

U – prvi trije stolpci Σ : prvi trije stolpci

-0.63	-0.02	0.09
-0.77	0.05	-0.08
-0.02	-0.33	-0.21
-0.02	-0.49	-0.30
-0.01	-0.26	0.72
-0.00	-0.68	-0.24
-0.01	-0.35	0.53

13.2	0	0
0	5.77	0
0	0	4.35

Latent Semantic Indexing

- Kaj nove osi pomenijo?

- kombinacije besed oz. **teme**
- stolpci U povedo o tem kako močno so posamezne besede zastopane v posamezni temi
- dobimo **prostor tem**

- $A^T U$

- projekcija dokumentov na nove osi

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1

U – prvi trije stolpci Σ : prvi trije stolpci

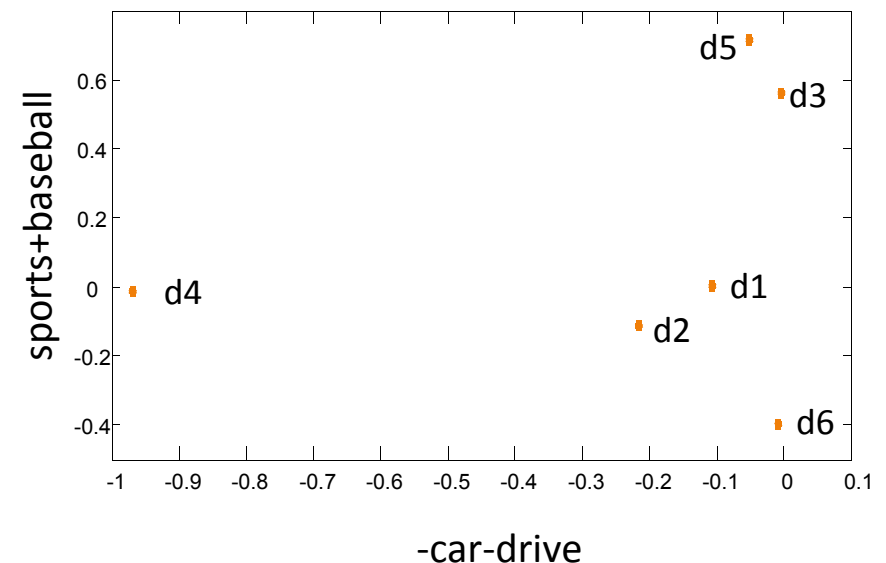
-0.63	-0.02	0.09
-0.77	0.05	-0.08
-0.02	-0.33	-0.21
-0.02	-0.49	-0.30
-0.01	-0.26	0.72
-0.00	-0.68	-0.24
-0.01	-0.35	0.53

13.2	0	0
0	5.77	0
0	0	4.35

Latent Semantic Indexing

- **Zmanjšanje** števila dimenzij
- Glavna poanta LSI
- Obdržimo le **prvih k** dimenzij – tem - z dovolj veliko singularno vrednostjo (Σ)
 - U_k – prvih k stolpcev U
 - S_k – prvih k vrstic/stolpcev v S
 - V_k – prvih k stolpcev V
- Projiciramo dokumente na izbrane osi – **prostor tem**
 - dobimo **pripadnost** dokumenta **temi**
 - za dokumente na katerih smo računali LSI je ta projekcija kar enaka V_k
 - sicer jo za poljuben dokument izračunamo kot $d^T U_k S_k^{-1}$

	d1	d2	d3	d4	d5	d6
car	1	2	0	8	1	0
drive	1	2	0	10	0	0
banana	0	1	0	0	0	2
eat	0	1	0	0	0	3
sports	0	0	3	0	2	0
food	0	0	1	0	0	4
baseball	0	0	1	0	3	1



Latent Semantic Indexing

- Računanje podobnosti med dokumenti
 - uporabimo **kosinusno podobnost** v prostoru tem
- Če imamo povpraševanje q
 - ga **preslikamo** v prostor tem: $q' = q^T U_k S_k^{-1}$
 - izračunamo **kosinusno podobnost** med dokumenti in q'
 - rangiramo dokumente

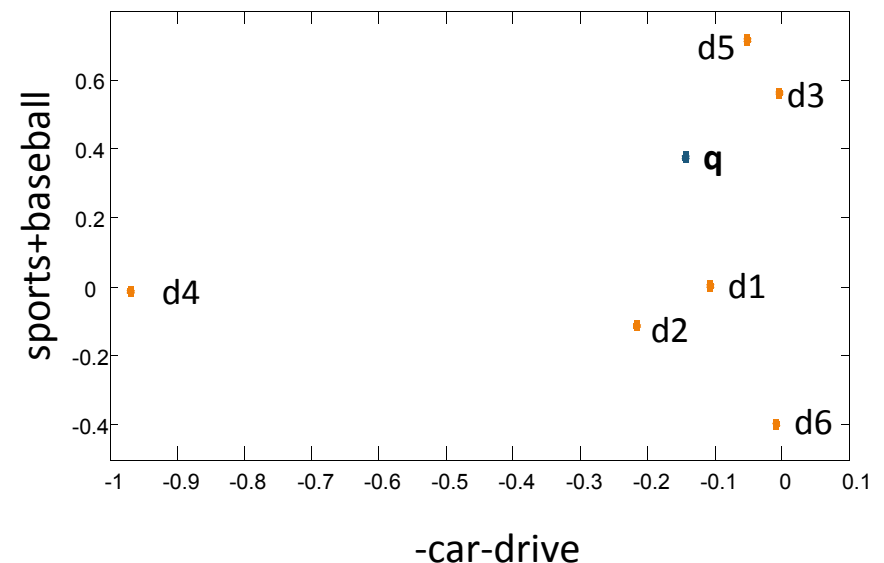
če nas zanimata šport in baseball

$$q = [0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1]$$

Kosinusna podobnost:

najbolj d3 (0.5), d5 (0.2)

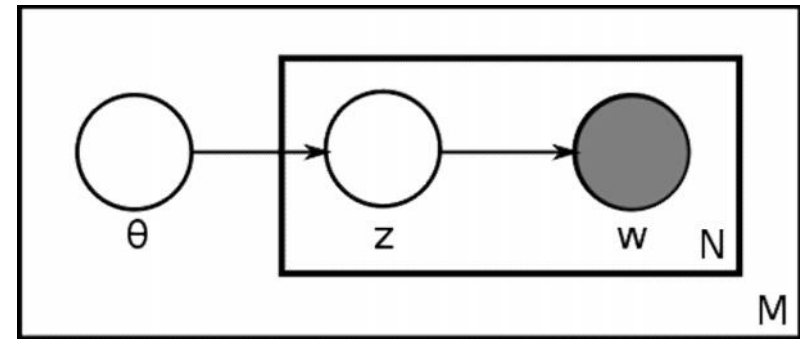
najmanj d4 (-0.74), d2





- Povpraševanje z LSI je **približno**
 - odkrije tudi dokumente, ki uporabljajo **sinonime** – besede z enakim pomenom
 - ker se navadno pojavljajo v enakem kontekstu in bodo zato del iste **teme**
- Obstaja več podobnih tehnik
 - Probabilistic LSA – pLSA
 - LDA – latent dirichlet allocation
 - ...
 - odpravljajo nekatere pomanjkljivosti, uvajajo izboljšave

Podobne tehnike



REFERENCE

- C.D. Manning et al. Introduction to Information Retrieval (book and slides), 2008
- P. Norvig: How to write a spelling corrector
- C. Whitelaw et al. (Google) Using the Web for Language Independent Spellchecking and Autocorrection, 2009
- S. Brin, L. Page The Anatomy of a Large-Scale Hypertextual Web Search Engine