Pretraživanje teksta

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Information retrieval

- Pronalaženje informacija (information retrieval)
  (1) reprezentacija, skladištenje, organizacija i pristup informacijama
  (2) pronalaženje materijala (dokumenata) nestrukturirane prirode (tekstualnih) koji zadovoljava potrebe za informacijama u okviru velike kolekcije

Indeksiranje i pretraživanje

- Indeksiranje predstavlja pripremu informacija za efikasno pretraživanje i uključuje tehnike za reprezentaciju, skladištenje i organizaciju informacija
- Pretraživanje predstavlja proces obrade upita i pronalaženje informacija koje korisnik traži, uključuje tehnike za efikasan pristup i pronalaženje informacija u prethodno kreiranim indeksima u procesu indeksiranja

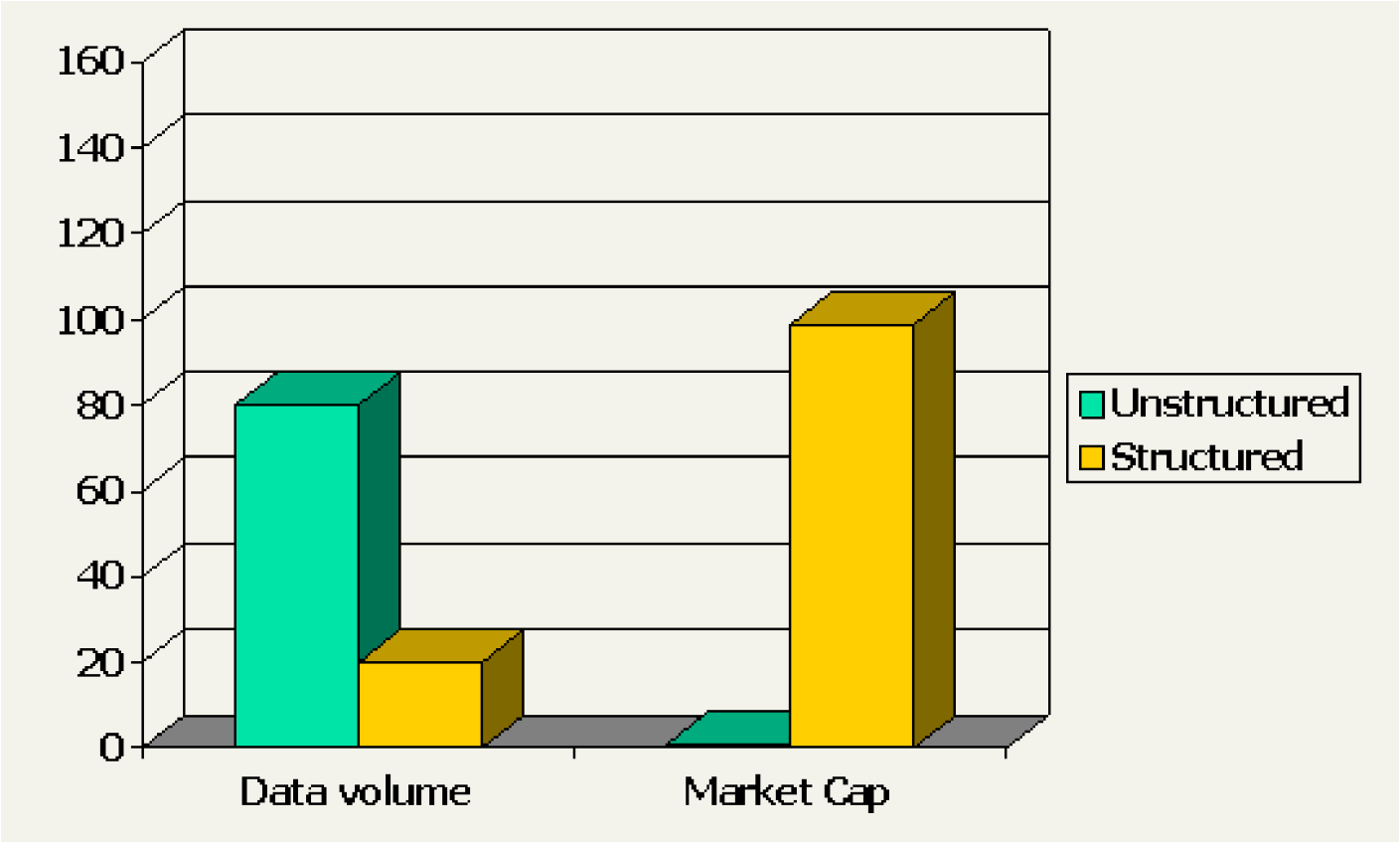
Data retrieval vs information retrieval

- Data retrieval: pronalaženje podataka koji zadovoljavaju precizno definisan kriterijum
  - karakteristično za baze podataka
- Information retrieval: korisnika interesuju informacije o nekoj temi, a ne podaci koji zadovoljavaju upit
  - podrazumeva nepreciznost, može da sadrži greške
  - informacije iskazane prirodnim jezicima mogu biti semantički neprecizne ili višeznačne
  - rangiranje pronađenih rezultata $\rightarrow$ pojam **relevantnosti** pogotka

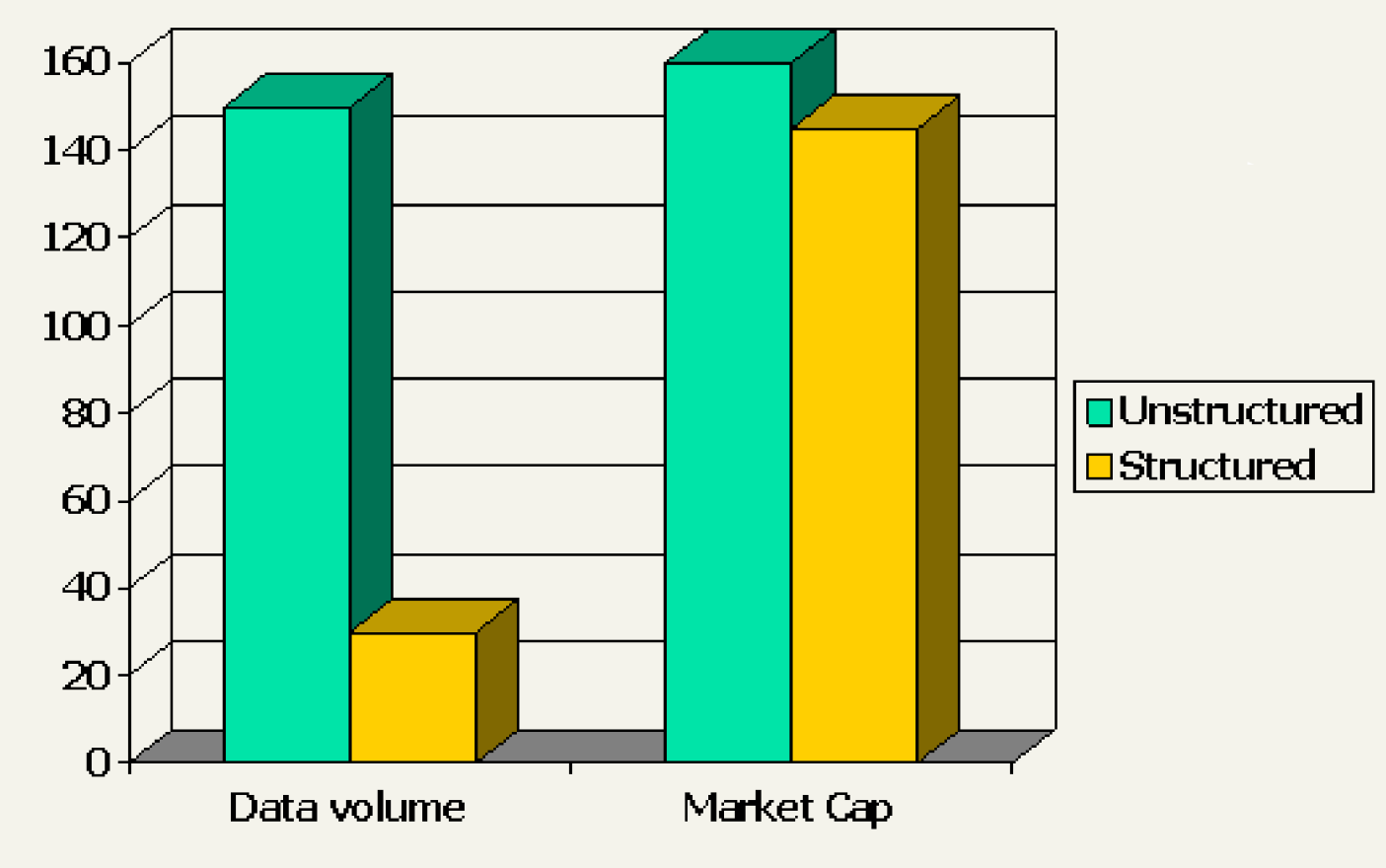
Počeci

- Pre 4.000 godina
- Sadržaji u knjigama
- Indeks pojmova u knjigama
- Indeks na nivou biblioteke knjiga
- Upotreba računara
- Digitalno doba, www, razoj IKT

Strukturirani i nestrukturirani podaci u 1996



Strukturirani i nestrukturirani podaci u 2006



WWW

- Pored bibliotekarstva i WWW je zaslužan za razvoj IR
- Pregled sadržaja, transakcije, društvene mreže, multimedijalni sadržaji
- Popularan izvor informacija zato što je jeftin (gotovo besplatan), pristup sa raznih mesta upotrebom raznih uređaja, sloboda u publikovanju (može biti i loše)
- Pretraga veba

IR danas

- Izučava se na većini fakulteta koji se bave računarskim naukama
- Google je jedna od najmoćnijih kompanija, a osnovna delatnost je IR
- Multidisciplinarna oblast: modelovanje, klasifikacija i klasterovanje, korisnički interfejs, vizuelizacija podataka, arhitektura sistema, prepoznavanje karaktera, semantički veb, itd.

Generacije IR sistema

- Biblioteke su prve imale računarski podržane sisteme za pretragu
- Sistemi su se razvijali na naučnim institucijama, a kasnije su se uključili i komercijalni proizvođači
- Prva generacija
  - Evidencija kataloških kartica
  - Pretraga po imenu autora i naslovu dela
- Druga generacija
  - Pretraga po ključnim rečima i oblasti kojoj delo pripada
  - Naprednije mogućnosti pretrage: kombinovanje upita, unos samo početnih slova, itd.

Generacije IR sistema

- Treća generacija
  - Trenutno aktuelna
  - Unapređen korisnički interfejs, mogućnost interakcije prilikom zadavanja upita
  - Rad sa digitalnim dokumentima, ekstrahovanje teksta, tehnike za analizu dobijenih tekstova (data mining)
- Pretraga veba
  - Krajem 90-tih godina 20. veka
  - Velika količina sadržaja, različiti profili korisnika, ne postoji kontrola ažuriranja sadržaja, veze između dokumenata se koriste prilikom rangiranja

Vrste IR

- Po arhitekturi (kako su indeksi organizovani)
  - centralizovani IR sistemi
  - distribuirani IR sistemi
    - protokoli i standardi: SRU, z39.50, CQL

Vrste IR

- Po sadržajima u kolekciji
  - Pretraga tekstualnih sadržaja
    - nestrukturiranih sadržaja
    - strukturiranih tekstualnih sadržaja koji iako imaju strukturu u nekim poljima svoje strukture imaju velike količine tekstova
  - Pretraga linkovanih tekstualnih sadržaja (pretraga veba)
  - Pretraga multimedijalnih sadržaja: slika, zvuk, video
  - Pretraga ostalih vrsta sadržaja
    - kolekcija programskih izvornih kodova
    - kolekcija 3D objekata

Modeli pretraživanja

- Klasični modeli
  - Bulov model
  - Vektorski model
  - Probabilistički model
- Alternativni modeli
  - Prošireni bulov model
  - Fuzzy model
  - Model neuronske mreže
  - Jezički model

Bulov model

- Zasnovan na teoriji skupova i Bulovoj algebri
- Posmatrani pojam se ili nalazi ili ne nalazi u dokumentu
- Nema rangiranja
- Nema parcijalnog poklapanja upita i dokumenta
- Konjukcija tri terma: jednako se posmatra i dokument koji nema ni jedan term i dokument koji ima dva terma

Vektorski model

- Težinski faktori vezani za pojedine termove u odnosu na dokumente i upite su pozitivne ali ne celobrojne vrednosti
- I upit ima težinske faktore
- Ima rangiranja
- Ima parcijalnog poklapanja upita i dokumenta
- I upit i dokument se predstavljaju kao n-dimenzionalni vektor (n je broj termova u rečniku)
- Ugao koji zaklapaju vektori je obrnuto srazmeran relevantnosti dokumenta za postavljeni upit

Probabilistički model

- Zasnovan na teoriji verovatnoće
- Pretpostavka: za svaki postavljeni upit postoji idealni skup dokumenata
- Proces pretrage je specificiranje osobina idealnog skupa dokumenata
- Problem inicijalnog idealnog skupa od kog se kreće
- Model pokušava da proceni verovatnoću da će korisnik smatrati određeni dokument relevantnim
- U određenim slučajevima efikasniji od vektorskog modela, ali u velikim kolekcijama slobodnog sadržaja obično je vektorski model bolji
- Postoji više vrsta ovih modela koji se razlikuju u načinima računanja verovatnoća i određivanju inicijalnog idealnog skupa - Robertson and Sparck Jones model; Croft and Harper model

Alternativni modeli

- Modifikovane verzije klasičnih modela
- Pokušaj da se prevaziđu neka ograničenja i manjkavosti klasičnih modela

Termovi i dokumenti

- šta je dokument

Termovi i dokumenti

- šta je dokument
- šta je term

Termovi i dokumenti

- šta je dokument
- šta je term
- Oba pojma mogu biti prilično složena

Parsiranje dokumenta

- Pre nego što razmotrimo termove da razmotrimo jezik i format dokumenta:

Parsiranje dokumenta

- Pre nego što razmotrimo termove da razmotrimo jezik i format dokumenta:
- U kom formatu je dokument? pdf, word, excel, html itd.

Parsiranje dokumenta

- Pre nego što razmotrimo termove da razmotrimo jezik i format dokumenta:
- U kom formatu je dokument? pdf, word, excel, html itd.
- Na kom jeziku je dokument?

Parsiranje dokumenta

- Pre nego što razmotrimo termove da razmotrimo jezik i format dokumenta:
- U kom formatu je dokument? pdf, word, excel, html itd.
- Na kom jeziku je dokument?
- Koji kodni raspored se koristi?

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.
  - email na francuskom sa PDF prilogom na španskom

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.
  - email na francuskom sa PDF prilogom na španskom
- Šta smatramo za dokument prilikom indeksiranja?

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.
  - email na francuskom sa PDF prilogom na španskom
- Šta smatramo za dokument prilikom indeksiranja?
- Fajl?

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.
  - email na francuskom sa PDF prilogom na španskom
- Šta smatramo za dokument prilikom indeksiranja?
- Fajl?
- Email poruka?

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.
  - email na francuskom sa PDF prilogom na španskom
- Šta smatramo za dokument prilikom indeksiranja?
- Fajl?
- Email poruka?
- Email poruka sa 5 priloga?

Format/jezik: komplikacije

- Jedan indeks sadrži termove iz više jezika
  - Cross-language IR - upit je na drugom jeziku u odnosu na dokument
  - Multilingual IR - dokumenti su na više jezika
- Nekada dokument ili njegovi delovi sadrže koriste različite jezike i formate.
  - email na francuskom sa PDF prilogom na španskom
- Šta smatramo za dokument prilikom indeksiranja?
- Fajl?
- Email poruka?
- Email poruka sa 5 priloga?
- Grupa fajlova (PPT konvertovan u HTML)?

Definicije

- Reč – Ograničen niz znakova koji se pojavljuje u tekstu

Definicije

- Reč – Ograničen niz znakova koji se pojavljuje u tekstu
- Term – „Normalizovana” reč (padež, morfologija, itd); klasa ekvivalencije reči

Definicije

- Reč – Ograničen niz znakova koji se pojavljuje u tekstu
- Term – „Normalizovana” reč (padež, morfologija, itd); klasa ekvivalencije reči
- Token – Instanca reči ili terma koja se pojavljuje u dokumentu

Definicije

- Reč – Ograničen niz znakova koji se pojavljuje u tekstu
- Term – „Normalizovana” reč (padež, morfologija, itd); klasa ekvivalencije reči
- Token – Instanca reči ili terma koja se pojavljuje u dokumentu
- Tip – U većini slučajeva isto što i term: klasa ekvivalencije reči

Razlikovanje tipa i tokena: primer

- *In June, the dog likes to chase the cat in the barn.*

Razlikovanje tipa i tokena: primer

- *In June, the dog likes to chase the cat in the barn.*
- Koliko ima tokena? Koliko ima tipova?

Kreiranje liste termova

- Ulaz:

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

Kreiranje liste termova

- Ulaz:

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

- Izlaz:

  friend | roman | countryman | so | ...

Kreiranje liste termova

- Ulaz:

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

- Izlaz:

  friend | roman | countryman | so | ...

- Svaki token je kandidat za stavku u listi pojava.

Kreiranje liste termova

- Ulaz:

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

- Izlaz:

  friend | roman | countryman | so | ...

- Svaki token je kandidat za stavku u listi pojava.
- Koji su validni tokeni?

Tokenizacija je složena – čak i za engleski

Primer: *Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.*

Tokenizujte ovu rečenicu

Jedna reč ili dve?

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company
- cheap San Francisco-Los Angeles fares
- York University vs. New York University

Brojevi

- 3/12/91
- 12/3/91
- Mar 12, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333
- Stariji IR sistemi ne indeksiraju brojeve ali to je često korisna stvar

Kineski: nema whitespace znakova

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。

Kineski: višeznačna segmentacija

Ova dva znaka mogu biti tretirani kao jedna reč „sveštenik“ ili sekvenca dve reči „i“ i „još“.

Još neki slučajevi bez razmaka

- Složenice u holandskom i nemačkom
- Computerlinguistik → Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter
- → leben + versicherung + gesellschaft + angestellter
- Inuitski: tusaatsiarunnanngittualuujunga (Ne čujem dobro)
- Švedski, finski, grčki, urdu, mnogi drugi jezici

Japanski

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ＭＯＴＴＡＩＮＡＩ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。

4 različita „alfabeta“: kineski znaci, hiragana za glavne gramatičke varijante, katakana za transkripciju stranih reči i latinica. Nema razmaka (kao i u kineskom).
Upit se može izraziti kompletno pomoću hiragane!

Arapsko pismo: s desna u levo + ligature

ك ِ ت ا ب ٌ ⇐ كِتَابٌ

un b ā t i k

/kitābun/ *‘a book’*

Arapsko pismo: dvosmernost

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

← → ← → ← START

'Algeria achieved its independence in 1962 after 132 years of French occupation.'

Dvosmernost nije problem ako je tekst zapisan u Unicode rasporedu.

Normalizacija u engleskom jeziku

- Potrebno je „normalizovati“ termove u indeksiranom tekstu i u upitima u isti oblik
- Primer: želimo da izjednačimo *U.S.A.* i *USA*
- Najčešće se definišu klase ekvivalencije termova
- Alternativno: asimetrično proširivanje
  - window → window, windows
  - windows → Windows, windows
  - Windows (bez proširivanja)
- Moćnije ali manje efikasno

Normalizacija u engleskom jeziku

- Potrebno je „normalizovati“ termove u indeksiranom tekstu i u upitima u isti oblik
- Primer: želimo da izjednačimo *U.S.A.* i *USA*
- Najčešće se definišu klase ekvivalencije termova
- Alternativno: asimetrično proširivanje
  - window $\rightarrow$ window, windows
  - windows $\rightarrow$ Windows, windows
  - Windows (bez proširivanja)
- Moćnije ali manje efikasno
- Zašto ne želimo da stavimo *window*, *Window*, *windows*, i *Windows* u istu klasu ekvivalencije?

Normalizacija u drugim jezicima

- Akcenti: résumé vs. resume, ćevapčići vs. cevapcici (prosto izostavljen akcent)
- Umlauti: Universität vs. Universitaet (zamena nizom slova "ae")
- Najvažniji kriterijum: kako će korisnici najverovatnije pisati upite za ovakve reči?
- I u jezicima koji redovno koriste akcente korisnici ih retko pišu (srpski, češki)
- Normalizacija i detekcija jezika su međuzavisni
- *PETER WILL NICHT MIT.* $\rightarrow$ MIT $=$ mit
- *He got his PhD from MIT.* $\rightarrow$ MIT $\neq$ mit

Velika i mala slova

- Svesti sva slova na mala slova
- Mogući izuzeci: reči u sredini rečenice koje počinju velikim slovom
- MIT vs. mit
- Fed vs. fed
- Često je najbolje sve prebaciti u mala slova jer će i korisnici tako pisati upite, bez obzira na ispravno pisanje

Stop reči

- Stop reči = česte reči koje nisu korisne prilikom pretraživanja
- Primeri (en): *a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with*
- Primeri (sr): *i, ili, kod, ka, sa, za, ...*
- Eliminacija stop reči je bila uobičajena u klasičnim IR sistemima
- Stop reči su nam potrebne za fraze, npr. "King of Denmark"
- Većina web pretraživača ne izbacuje stop reči

Klase ekvivalencije

- Soundex algoritam: fonetska ekvivalencija, Tchebyshev = Chebysheff
- Tezaurus: semantička ekvivalencija, car = automobile

Šta radi Google?

- Stop reči
- Normalizacija
- Konverzija u mala slova
- Koren reči (stemming)
- Umlauti
- ...

Leme

- Redukovanje raznih gramatičkih oblika na baznu formu
- Primer: *am, are, is* $\rightarrow$ *be*
- Primer: *car, cars, car's, cars'* $\rightarrow$ *car*
- Primer: *the boy's cars are different colors* $\rightarrow$ *the boy car be different color*
- Lematizacija podrazumeva „pravilnu“ redukciju na osnovni rečnički oblik (lemu).
- Inflekciona morfologija (*book* $\rightarrow$ *books*) - ista vrsta reči samo u drugom obliku (jednina, množina, lice, glagolski oblik)
- Derivacione morfologije (*destruction* $\rightarrow$ *destroy*) - nova vrsta reči ali sa istom osnovom

Definicija

- Grubi heuristički proces koji odseca krajeve reči sa ciljem da postigne rezultat što sličniji onome koji postiže pravilna lematizacija bazirana na lingvističkom znanju.
- Zavisna od jezika
- Često je i inflekciona i derivaciona
- Primer derivacione: *automate, automatic, automation* svi se redukuju na *automat*

Definicija

- Ideja steminga je da poveća performanse IR sistema (pre svega povrat - recall)
- Transformiše u isti oblik varijante jedne reči
- Nekad je bio baziran samo na književnom jeziku, poslednjih godina potreba da se steming bavi i slengom
- Steming nije konceptualno primenljiv na sve jezike
  - Nije primenljiv na Kineski jezik
  - Primenljivo je na jezik koji imaju uobičajene paterne kreiranja varijanti jedne reči - Indo-Evropski jezici
- Ne mora svoditi reči na root (koren), svodi ih na stem
- Obično ne skida prefikse, samo sufikse
- Za dve reči koje se svode na isti stem kažemo da su spojene (*conflated*)

Stemeri zasnovani na algoritmu

- Skupom pravila, konvencija i definisanjem redosleda njihove primene se definiše steming
- Vrlo su značajni za IR sisteme, često se koriste jer ih je lakše kreirati nego stemere zasnovane na rečniku
- Na žalost nema puno opisanih stemera zasnovanih na algoritmima, pa čak i kada su opisani često je taj opis nejasan i tumači se pogrešno
- Rade brzo i začuđujuće dobro - *Why does it do so well? (Krovetz, 1995 - page 89)*

Stemeri zasnovani na rečniku

- Postoje i stemeri zasnovani na rečniku, ne koristi se Snowball, teži su za kreiranje, ali mogu imati bolje performanse
- Ove dve vrste stemera nisu striktno razdvojene
  - Stemeri zasnovani na algoritmu mogu imati duge liste izuzetaka (praktično rečnika) koji se koriste da bi se smanjile greške
  - Stemeri zasnovani na rečniku obično skidaju određene nastavke pre nego što izvrše *look-up* u rečniku (uklanjanje množine, svođenje na osnovni padež, itd.) - ovo ukljanjanje je zasnovano na algoritmu
- Izgradnja rečnika je vremenski zahtevna
- Kad je konačno izgrađen već je star nekoliko godina (jezici se menjaju), a treba ga i u budućnosti ažurirati - konstantan posao

Vrste sufiksa

- a-sufiksi (*attached suffixes*)
  - Jedna reč nakačena na kraj druge
  - Ima ih u Italijanskom, Španskom, Portugalskom
  - mandar**gli** - to send + **to him**
  - Obično se skidaju
- i-sufiksi (*inflectional suffixes*)
  - Inflekciona morfologija
  - Ista vrsta reči u različitom obliku - množina, padeži, glagolska vremena, itd.
  - Uvek se skidaju
- d-sufiksi (*derivational suffixes*)
  - Derivaciona morfologija
  - Različite vrste reči sa istim korenom - radnik, raditi, radan (vredan), itd.
  - U retkim situacijama se skidaju

Greške u stemovanju reči 1/2

- *Under-stemming*
  - Skidanje previše malog sufiksa
  - Problem je što se ne spajate sa rečima koje imaju isto značenje
  - Dakle, možda imate isti efekat kao da nemate steminga, a možda i gori
  - Ne mora da izazove probleme, ako nema spajanja sa stemovima reči drugačijeg značenja
  - U najboljem slučaju ne poboljšava povrat, a nadamo se da ne pogoršava preciznost, odnosno ne unapređujemo IR sistem, ali se nadamo da ga ne unazađujemo
- *Over-stemming*
  - Skidanje previše velikog sufiksa tako da se reč spaja sa stemom reči drugačijeg značenja
  - Smanjuju preciznost
  - Uvodi nove homonime u jezik - reči (odnosno stemove) koji se isto pišu a ne znače isto
  - Najopasnija vrsta greške
  - Često se pokušava rešiti uvođenjem rečnika

Greške u stemovanju reči 2/2

- *Mis-stemming*
  - Skidanje nečega što je deo stema što se činilo kao sufiks
  - Ne mora da izazove probleme, ako nema poklapanja sa drugim stemovima
  - Može da poboljšava povrat, a nadamo se da ne pogoršava preciznost
  - Nije bitno što je dobijeno nešto što ljudima ne izgleda kao stem te reči
- Nepravilni gramatički oblici - rečnik ili posebna pravila u algoritmima
- Stop reči - nekad se izbacuju pa u nekim stemerima nema pravila za njih
- Retki oblici reči - rečnik, posebna pravila u algoritmima ili ništa (to su greške stemera)

Deo pretprocesiranja teksta

- Steming je deo pretprocesiranja teksta
- Često se očekuje da su pre steminga izbačena velika u mala slova, izbačene stop reči, itd.
- Apostrofi i znaci interpunkcije obično ne stižu do stemera
- Pre ili posle stemera se vrša pretprocesiranje teksta upotrebom rečnika
- Krajnji rezultat ne moraju biti reči koje postoje u jeziku, bitno je da su povećane performanse IR sistema - nije mnogo pogoršana preciznost, i **povrat** je popravljen

Šta je Snowball

- SNOBOL (StriNg Oriented and symBOlic Language) - link
- Prost jezik za obradu stringova
- M. Porter - 2001. godina
- String paterni odlučuju o toku izvršavanja programa
- Pogodan jezik za pisanje stemera zasnovanih na algoritmu, lako ga je razumeti, odnosno pravilno tumačiti
- Stemer opisan Snowball jezikom se lako može prevesti u programske jezike Java i ANSI C, a postoje načini i da se koriste u Python-u, Objektnom PASCAL-u i drugim jezicima

Porterov algoritam

- Najpoznatiji stemming algoritam za engleski jezik
- Po rezultatima je dobar barem koliko i druge alternative
- Primena konvencija + 5 faza redukcije
- Faze se primenjuju sekvencijalno
- Svaka faza predstavlja skup komandi
  - primer komande: Obriši zadnji *ement* ako je ono što ostaje duže od 1 znaka
  - replacement $\rightarrow$ replac
  - cement $\rightarrow$ cement
- Primer konvencije: Od skupa pravila primeni onu koja se odnosi na najduži sufiks

Porter stemmer: Nekoliko pravila

Pravilo			**Primer**		
SSES	$\rightarrow$	SS	caresses	$\rightarrow$	caress
IES	$\rightarrow$	I	ponies	$\rightarrow$	poni
SS	$\rightarrow$	SS	caress	$\rightarrow$	caress
S	$\rightarrow$		cats	$\rightarrow$	cat

Tri stemmera: poređenje

Uzorak teksta: 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 a 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

Da li stemming unapređuje efikasnost?

- U opštem slučaju stemming popravlja efikasnost za neke upite; za neke druge pogoršava
- Klasa ekvivalencije Porter stemmera *oper* sadrži sve ovo: *operate operating operates operation operative operatives operational*.
- Primer upita gde stemming smeta: operational AND system i operating AND system