1. Objasniti pojam dokumenta, papirnog dokumenta i digitalnog dokumenta

Dokument predstavlja svaki pisani ili snimljeni proizvod namenjen za komunikaciju ili cuvanje podataka.

Papirni dokument je svaki dokument koji je rukom napisan, otkucan ili odstampan proizvod za cuvanje informacija, napisan na papiru.

Digitalni dokument je racunarski obradjen skup informacija i tim informacijama se rukuje kao osnovnom jedinicom obrade.

2. Metapodaci, navesti primere.

Metapodaci predstavljaju podatke o podacima, tj. detaljnije opisuju podatke kao sto su velicina, format, lokacija itd. Postoje metapodaci tekstualnih dokumenata i meta podaci fotografija.

Metapodaci za tekstualni dokument: autor, naslov, datum nastanka, kljucne reci..

Metapodaci za fotografiju: datum, vreme i mesto fotografisanja, autor, objekti na slici, podesavanje fotoaparata...

3. Veze izmedju dokumenata i metapodataka. -> O,T,K,I, R,V

-opisni metapodaci -> opisuju sam dokument

-tehnicki metapodaci -> opisuju tehnicke karakteristike dokumenta

-kontekstualni metapodaci -> podaci o kategoriji ili temi dokumenta

-izvorni metapodaci -> informacije o izvoru metapodataka

-reference -> metapodaci sadrze informacije o povezanim dokumentima

-vremenski -> informacije o vremenskim aspektima dokumenta(datum izrade, izmene..)

4. Faze zivotnog ciklusa dokumenta.

- inicijalizacija -> formiranje podataka

- priprema -> proizvodnja sadrzaja sve do uspostavljanja

-uspostavljanje -> pre koriscenja, dokument se odobrava

-koriscenje -> upotreba dokumenta

-revizija -> promena sadrzaja ili namene dokumenta

-arhiviranje -> ostavljanje starih podataka na sigurno, duze vreme

-unistavanje -> nakon isteka arhiviranja

5. Objasniti zivotnu fazu dokumenta koriscenje.

Faza koriscenje je sama upotreba dokumenta(analiza i citanje), mogu se dodati i iskustva korisnika o koriscenju dokumenta.

6. Faza dokumenta arhiviranje.

Faza gde se dokument koji se vise ne koristi ostavlja na sigurno mesto kao sto je arhiva ili neka baza podataka na duzi period. Arhivirani dokumenti se ne mogu menjati, ali se mogu reprodukovati.

7. Objasniti pojmome upravljanje verzijama, sekvencijalno i konkurentno vazenje verzija.

Prilikom izmene verzije dokumenta, menjaju se i njegovi metapodaci.

Sekvencijalno vazenje -> jedina vazeca verzija dokumenta je poslednja i ona podrzava sve promene od prethodnih verzija.

Konkurentno vazenje-> sa dolaskom nove verzije dokumenta, nova verzija ne postaje jedina verzija dokumenta, nego u isto vreme postoji vise verzija koje su vazece.

8. Osnovna namena sistema za upravljanje dokumentima?

Osnovna namena je pracenje i skladistenje digitalnih dokumenata.

9. Funkcije sistema za upravljanje dokumentima? - nabrojati 8

- skladistenje, katalogizacija, pretrazivanje, zastita podataka, oporavak od katastrofe, arhiviranje, distribucija i upravljanje poslovnim procesima.

10. Dublin core format metapodataka?

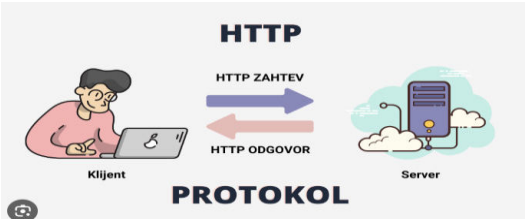
Dublin core ima 2 nivoa: SIMPLE(15 elemenata za opis metapodataka) i QUALIFIED(3 dodatna elementa)

-doprinosioci, glavni doprinosioc, piokriveno gradivo, datum, opis, format, identifikator, jezik, izdavac, reference, autorska prava, izvor, tema, naslov, zanr

-moze biti XML, a i ne mora

11. Sta su protokoli za razmenu podataka?

Protokoli koji omogucavaju prenos podataka izmedju *sistema* ili *aplikacija*. Oni definisu standardizovane formate poruka, komunikacione kanale, pravila za slanje, prijem i obradu podataka. --> HTTP, FTP, SMTP



12. Osnovne karakteristike OAI-PMH protokola. - reći šta je i šta omogućava

To je protokol za razmenu metapodataka izmedju sistema. Http baziran.

On omogucava:

-definisanje skupova podataka, iterativno preuzimanje pomocu resumptionTokena i pretragu, kao i preuzimanje metapodataka.

13. Protokoli za udaljeno pretrazivanje?

To su standardizovani protokoli aplikativnog sloja za pretragu i dobavljanje informacija iz baza podataka putem TCP/IP mreze. Predstavnici su Z39.50 i SRU protokol.

14. Osnovne karakteristike Z39.5 i SRU protokola?

Z39.5 -> komunikacija izmedju klijenta i servera ne zavisi od platfrome. Izvrsava zahtev za pretrazivanje i vraca formirane liste rezultata. Ima bogat upitni jezik i koristi Bulove izraze.

SRU -> naslednik Z39.5, koristi XML i SOAP za komunikaciju(preko HTTP). CQL je upitni jezik i omogucava pretragu, preuzimanje i manipulaciju rezultatima.

15. Cime se bavi oblast pronalazenja informacija(information retrieval)?

Bavi se reprezentacijom, skladistenjem, organizacijom i pristupom informacijama, kao i pronalazenjem tekstualnih dokumenata za velike kolekcije.

16. Razlika izmedju information retrieval i data retrieval?

Information retrieval pronalazi podatke o nekoj temi(pesma -> izvodjac, zanr, autor), a data retrieval pronalazi podatke koji zadovoljavaju samo precizno definisan kriterijum pretrage.

17. Kako se razvijala oblast pronalazenja informacija?

Pre oko 4000 godina, zatim nastaje sadrzaj knjige i njeni delovi, zatim su izdvojene bitne celine, klasifikacija o temi, piscu, zanru i dolaskom racunara knjige idu u digitalni sadrzaj, dolazi i WWW itd.

18. Kakve arhitekture mogu imati sistemi za pretragu?

Mogu imati centralizovane i distribuirane i delimo ih na osnovu nacina skladistenja indeksa.

19. Vrste sadrzaja koje mogu biti pretrazivane?

Imamo tekstualne(*struktuirane i nestruktuirane*), linkovane tekstualne, multimedijalne(*slika,zvuk,video*), ostalo(*3d objekti..)*

20. Koji modeli pretrazivanje se koriste u sistemu za pretragu?

-klasicni -> bulov, vektorski i model verovatnoce

-alternativni -> prosiren bulov, fuzzy, jezicki model, model neuronske mreze.

21. Razlika izmedju termina(terma) i tokena.

Token moze biti rec, broj ili interpunkcijski znak i dobija se razbijanjem teksta na manje delove("Ovo je primer" - "ovo", "je", "primer"). - tokenizacija

Term predstavlja pojedinacnu rec ili frazu(u kontekstu medicinske pretrage term moze biti "srcani udar", "dijabetes" itd.)

22. Sta je tokenizacija i koji problemi postoje u ovoj fazi pretprocesiranja?

Tokenizacija je proces razbijanja teksta na manje delove gde dobijamo tokene.

Problemi koji se mogu desiti su: spajanje reci(Novi Sad, New York..), kada nam zasmeta znak interpunkcije(don't, won't), neke skracenice u jeziku, citanje sa levo na desno(Arapi)..

23. Zasto se vrsi normalizacija reci?

Normalizacija je svodjenje termova u isti oblik i vrši se radi smanjenja termova koji se koriste. Na primer kada zelimo da izjednacimo U.S.A i USA, takodje uklanjanje interpunkcijskih znakova, jos jedan primer cevapi i ćevapi. Vrši se i konverzija svih slova u velika ili mala itd..

24. Sta je steaming?

Steaming je proces svodjenja reci na osnovni oblik. Obicno se postize uklanjanjem sufiksa ili prefiksa iz reci. Cilj je *grupisanje* reci radi manjeg obima reci, brze pretrage i analize. Na primer imamo reci --> *trce, trcao i trcanje - svodi se na trc.*

25. Sta je lematizacija?

Lematizacija je proces slican steamingu, ali ona pretvara rec u njenu lemu. *Primer reči: "trce", "trcao" i "trcanje" se svode na lemu "trcati*". Koristi se kada je potrebno sacuvati gramatičko značenje reči.

26. Objasniti Bulov model pretrage.

Bulov model za pretragu koristi opertatore AND, OR i NOT i ceo dokument se posmatra kao skup termova. U bulovom modelu pretrage dokument ili jeste ili nije zadovoljio upit. On je brz i jednostavan, ali ne uzima u obzir važnost ili rangiranje dokumenta.

27. Sta je invertovani indeks i kako se kreira?

Invertovani indeks je *struktura podataka* koja se koristi u sistemu za pretragu kako bismo lakse pristupili odredjenim podacima na osnovu kljucnih reci ili termina. Postupak kreiranja je tokenizacija, normalizacija, indeksiranje*(za svaki normalizovani token se kreira lista dokumenata koji sadrze taj token)* i sortiranje*(zbog brzeg pristupa podacima).*

28. Procesiranje upita kod Bulovog modela?

Procesiranje upita se vrsi tako što: tokenizujemo upit, normalizujemo dobijene tokene, identifikujemo upit(sta se od AND, OR i NOT koristi), evaluacija operatora(sta radi koji opertor), pretraga invertovanog indeksa(na osnovu evaluacije) i na kraju sledi prikaz rezultata.

29. Sta su pointeri za preskakanje?

Pointeri za preskakanje nam omogucavaju preskakanje elemenata u listi koji svakako nece biti u rezultatima pretrage. Zbog njih *imamo ubrzano izracunavanje preseka*. Imamo staticke(fiksni broj pointera) i dinamicke pointere(prilagodjavaju se promenama u strukturi).

30. Sta se moze koristiti ako je potrebno podrzati upite fraze?

Mozemo koristiti dvorecni ili pozicioni indeks.

31. Sta je dvorecni indeks?

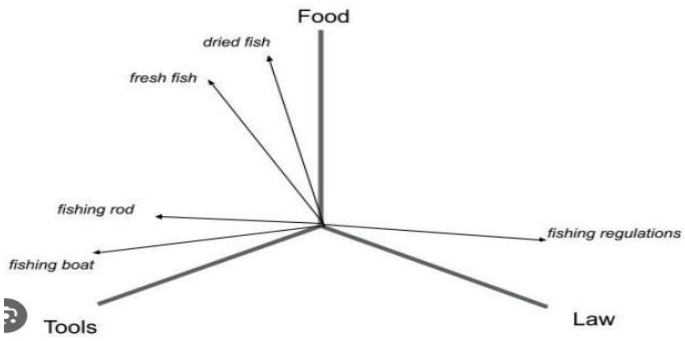
Dvorecni indeks je nacin indeksiranja gde se dve reci stavljaju pod jednim termom ukoliko imaju neko znacenje kada su jedno pored drugog. Te 2 reci se cuvaju kao specifican term. Primer moze biti "sive markirane pantalone" --> "sive markirane" i "markirane pantalone"

32. Sta je pozicioni indeks? - cuva *poziciju* reci

Pozicioni indeks je dobra zamena za dvorecni indeks jer cuva i poziciju reči i zbog toga ne mora da se prolazi kroz ceo dokument.

33. Objasniti vektorski model pretrazivanja - jedan od najcesce koriscenih modela

To je model koji se koristi *za pretragu slicnosti dokumenta u odnosu na zadati upit*. Daje mogucnost da se dokument delimicno poklapa sa upitom. Svaki dokument se predstavlja kao vektor. Za tacnost dokumenta se koristi *kosinusna slicnost* koja meri ugao izmedju **vektora dokumenta** i **vektora upita**. Veci ugao ukazuje na manju slicnost i obrnuto.



34. Sta je ocena relevantnosti?

To je mera koliko se dokument i upit preklapaju. Ukoliko se term ne pojavljuje u dokumentu, ocena je 0, a sto se češće pojavljuje, ocena je veća.

35. Sta je frekvencija terma?

Frekvencija terma je mera koliko se puta odredjeni termin pojavljuje unutar dokumenta ili teksta. Primer moze biti *"ova knjiga je najlepsa knjiga"* -> frekvencija terma "knjiga" je 2 jer se dva puta pojavljuje.

36. Sta je frekvencija dokumenta?

Frekvencija dokumenta je mera koja odredjuje koliko se puta termin pojavljuje u skupu dokumenata. Koristi se zbog *procene vaznosti termina*.

Primer: imamo skup od 100 dokumenata i trazimo frekvenciju za term "pretrazivanje". Ukoliko se pojavi u 50 dokumenata, frekvencija dokumenta za taj termin ce biti 50.

37. Sta je tf-idf?

To je proizvod *tf(frekvencije terma)* i *idf(mere informativnosti terma)* težine nekog terma.

tf-idf težina raste sa brojem pojavljivanja terma u dokumentu i sa retkošću tog terma u kolekciji.

38. Objasniti kreiranje težinske matrice.

To je proces kojim se vrši transformacija skupa dokumenata u matricu koja predstavlja težine ili važnosti termina u dokumentima.

Koraci su: tokenizacija, izracunavanje tf(u svakom dokumentu za svaki term), izracunavanje idf(za svaki term), izracunavanje tf-idf i
kreiranje tezinske matrice: svaki dokument se predstavlja kao vektor koji sadrži težinske vrednosti tf-idf za svaki termin. Svaki red matrice predstavlja jedan dokument, a svaki element u redu predstavlja težinu termina.

39. Razlika izmedju Bulovog i vektorskog modela pretrage?

Kod Bulovog upita dokument je pogodan ili nije pogodan, a kod vektorskog se svakom dokumentu daje odredjeni nivo vaznosti u zavisnoti koliko je pogodan za upit. Bulov model prikazuje listu dokumenata, a vektorski model rangiranu listu dokumenata.

40. Da li se relevantnost odgovora meri u odnosu na informacionu potrebu ili na upit?

Meri se u odnosu na zeljenu informaciju jer korisnike zanima informacija o necemu a ne samo odgovor na prosledjen upit.

41. Sta je preciznost?

Predstavlja udeo pronadjenih relevantnih rezultata medju svima dokumentima. Primer: nasao 15 od kojih je 10 relevantno, preciznost je velika.

Racuna se -> pronadjeni relevantni/svi PRONADJENI

42. Sta je povrat?

Povrat je udeo pronadjenih relevantnih rezultata u svim relevantnim dokumentima u kolekciji. Racuna se -> pronadjeni relevantni/svi RELEVANTNI

43. Sta je F mera i zasto je relevantnija od preciznosti i povrata?

F mera omogucava da se meri kompromis izmedju preciznosti i povrata. F = 2PR/PR . Ona tezi ka postizanju balansa.

44. Kako se moze vrsiti evaluacija performansi sistema za pretragu?

Vrsi se na osnovu preciznosti, clickthrough-a(prvi klik na sajt kod velikih pretrazivaca), na osnovnu laboratorijskih studija(nadgledanje ponasanja korisnika) i A/B testiranje.

45. Sta je kapa mera?

Mera koliko se medjusobno ocenjivaci slazu i pomocu nje se meri konzistentnost medju ocenjivacima. Ide od -1(neslaganje) do 1(savrsena saglasnost). Formula je: k= P(A) - P(E) / 1-P(E)

46. Opisati A/B testiranje

Testiranje kod velikih pretrazivaca. 1% korisnika u vecini slucajeva tajno dobije novu verziju sa "unapredjenjima" i tako na osnovnu rezultata korisnika gledaju da li promene treba sprovesti.

47. Sta je Lucene?

To je open-source biblioteka visokih performansi za full-text pretragu sadrzaja. Pisana u Javi i koriste je Wikipedia, Eclipse..

48. Klase Document i Field u Lucene biblioteci.

Klasa Document obuhvata nas dokument, njegove podatke, metapodatke i sluzi za indeksiranje i pretragu. Sastoji se od polja(field) gde svako polje ima svoje ime i sadrzaj.

Polja mogu biti oznacena kao: -indexed -> obavezna za pretragu i sortiranje

-tokenized -> podela na tokene

-stored -> cuva se originalan sadrzaj polja u tekstu

-stored termVectors -> uz dokument sacuvan je i invertovan indeks tipa: TextField, StringField, IntF, Long.

49. Osnovne karakteristike Lucene.

Upiti unutar jezika se izrazavaju kao stringovi, parsiranje teksta obuhvata i analizu, postoje dva tipa termova(fraze i posebne reci). Termovi su vezani za polje -> "naslov: Bela Griva".

Ukoliko se polje ne navede, podrazumeva se default. Moze se povezati logickim operatorima. Primer:

"naslov: titanik AND korice: tvrde". Postoji i dzoker znak * koji menja jedno slovo.

50. Kako se implementira analiza (procesiranje) teksta pomoću Lucene-a?

Analiza teksta pomoću Lucene-a se implementira kroz upotrebu analizatora(Analyzer) i filtera (TokenFilter).

1. tokenizacija - deljenje teksta na tokene
2. normalizacija - normalizacija tokena
3. filtriranje - uklanjanje nepozeljnih elemenata
4. steaming ili leming

STEVAN I JOVAN DEGENERICI.