1.Objasniti pojam dokumenta, papirnog dokumenta i digitalnog dokumenta.
- Dokument predstavlja stvarni pisani ili snimljeni proizvod namenjen za komunikaciju ili čuvanje podatak a.
-Dokument na papiru – rukom napisan,odštampan ili otkucan proizvod za čuvanje informacija napisan na hartiji
-Digitalni dokument – računarski obrađen skup informacija kojim se rukuje kao osnovnom jedinicom obra de

2. Šta su metapodaci i navesti primere metapodataka?
- Metapodaci predstavljaju podatke o podacima tj. informacije koje bliže opisuju naše podatke poput lokac ije naših podataka,veličine
tih podataka,formata podataka …

3. Kakve sve veze između dokumenta i metapodataka mogu postojati?
-Kao što je prethodno rečeno, metapodaci nekog dokumenta su podaci koji više govore o našem dokume ntu.
Metapodaci pisanih dokumenata: -Autor -Datum nastanka -Naslov -Sadržaj -Ključne reči …
Metapodaci fotografija: -Autor -Datum i vreme nastanka -Mesto nastanka - podešavanje kamere …

4. Koje su faze životnog ciklusa dokumenta?
-Faze životnog ciklusa: -Inicijalizacija(formiranje podataka potrebnih za kasniju pripremu)
-Priprema(proizvodnja sadržaja sve do trenutka uspostavljanja)
-Uspostavljanje(odobravanje sadržaja, dodavanje sadržaja za id.podnosioca zahteva,id zateva,)
-Korišćenje(sama upotreba podataka poput čitanja istih i distribucija istih)
-Revizija(promena sadržaja ili namene dokumenta)
-Arhiviranje(ostavljanje starih podataka na sigurno mesto na duže vreme)
-Uništavanje

5. Objasniti životnu fazu dokumenta korišćenje
-Predstavlja fazu same upotrebe dokumenta npr. čitanje ili analiza istog, kao i dodavanje komentara u me tapodatke.

6. Objasniti životnu fazu dokumenta arhiviranje.
-Predstavlja fazu gde se dokumenti koji se više ne koriste ostavljaju na sigurno mesto u arhivu ili bazu na neki duži period.Ti dokumenti više nisu
podložni promeni, ali ih je moguće reprodukovati.

7. Objasniti pojmove upravljanje verzijama dokumenata, sekvencijalno i konkurento važenje verzija
-Prilikom izmene verzije dokumenta podrazumeva se izmena njegovih metapodataka i/ili sadržaja.
Sekvencijalno važenje dokumenata - jedina važeća je poslednja verzija dokumenta,koja podržava sve pr omene od prethodnih verzija
Konkurentno važenje dokumenata - u isto vreme postoji više različitih verzija dokumenata koji su važeći. Nova verzija ne zamenjuje automatski
staru verziju dokumenta.

8. Koja je osnovna namena sistema za upravljanje dokumentima?
-Osnovna namena sistema za upravljanje dokumentima je praćenje i skladištenje digitalnih dokumenata.

9.Koje su funkcije sistema za upravljanje dokumentima?
Osnovne funkcije za upravljanje dokumentima jesu: -Skladištenje dokumenata -Katalogizacija -Pretraži vanje -Zaštita podataka -Oporavak od katastrofe
- Arhiviranje -Distribucija -Upravljanje poslovnim procesima

10. Opisati Dublin Core format metapodataka.
- Dublin Core ili dablinsko jezgro predstavlja skup od 15 osnovnih metapodataka za opisivanje fizičkih ili di

gitalnih dokumenata.Oni uključuju:
-Doprinosioce -Pokriveno gradivo -Glavnog doprinosoca -Datum -Opis -Format -Identifikator :jedinstve ni obeleživač - Jezik -Izdavač -Reference
-Autorska prava -Izvori -Tema -Naslov -Žanr
Može biti napisan u XML format, a i ne mora.

11. Šta su protokoli za razmenu podataka?
Predstavljaju protokole razvijene za svrhu sakupljanja metapodataka iz dokumenata ili arhiviranih podatak a tako da se od tih podataka može napraviti
neka nova vrsta usluga.

12.Koje su osnovne karakteristike OAI-PMH protokola?
OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting) protokol je protokol za razmenu pod ataka između sistema.On je HTTP baziran,
definiše 6 vrsta zahteva (identify - preuzimanje informacija o repozotorijumu
getRecord - preuzimanje pojedinačnih metapodataka zapisa iz repozitorijuma
ListMetadataFormats - preuzimanje formata metapodataka dostupnih u repozitorijumu
ListRecords - prikupljanje liste zapisa iz repozitorijuma
ListIdentifiers - skraćeni oblik ListRecords
ima mogućnost definisanja skupova,
moguće je iterativno preuzimane delova teksta.

13. Šta su protokoli za udaljeno pretraživanje?
Predstavljaju međunarodno standardizovane komunikacione protokole aplikativnog sloja za pretraživanje i dobavljanje informacija
iz baza podataka putem TCP/IP mreže. Z39.50 i SRU su predstavnici ovoga.

14.Koje su osnovne karakteristike Z39.50 i SRU protokola?
Koriste se za pretraživanje i dobavljanje informacija iz baza podataka putem TCP/IP mreže.

- Z39.50 - binarni klijent server protokol koji koristi TCP/IP protokol za komunikaciju između klijenta i serve ra, ne zavisi od platforme.
Njegova osnovna funkcionalnost je uspostava komunikacije između klijenta i servera,izvršavanje zahteva za pretraživanje, a zatim vraćanje formirane liste
rezultata pretrage. Omogućava bogat upitni jezik koji omogućava korišćenje Bulovih izraza,skraćivanje re či i ima izbor naprednih opcija pretrage.
Nedostatak ovog protokola je relaksirana specifikacija koja ostavlja mogućnost neimplementiranja osnovn ih mehanizama kao i
implementiranje istih na različite načine.

-SRU protokol - naslednik Z39.50, koristi XML dokumente za transport podataka,koristi CQL (contextual q uery language) za specifikaciju
upita. Većina mogućnosti i mehanizama je preslikana iz Z39.50. Napredak - za komunikaciju se koriste ot voreni i široko prihvaćeni standardi.

15.Čime se bavi oblast pronalaženja informacija (information retrieval)?
Information retrieval -pronalaženje informacija bavi se reprezentacijom,skladištenjem,organizacijom i prist upom željenim informacijama kao i pronalaženjem dokumenata
nestruktuirane prirode(tekstualnih) koji zadovoljavaju potrebe za informacijama u okviru velike kolekcije.

16. Koja je razlika između pronalaženja podataka (data retrieval) i pronalaženja informacija (information re trieval)?
Razlika je u tome što data retrieval pronalazi podatke koji zadovoljavaju samo precizno definisan kriteriju m (daju samo one podatke koji su istovetno),
dok information retrieval pronalazi informacije o nekoj temi, a ne tačne podatke (recimo ako tražite nešto v

ezano za određenu pesmu, biće izbačene informacije
i o izvođaču, albumu i žanru muzike).

17. Kako se razvijala oblast pronalaženja informacija?
Počela je time što se neko davno (pre oko 4000 god) dosetio da napiše sadržaj knjige i time obeleži delov e.Zatim, dodat je indeks pojmova - izdvojene bitne celine,
pa je zatim dodata klasifikacija po temi, piscima, žanru unutar biblioteka. Dolaskom računara se došlo na i deju da se i knjige prebace u digitalni sadržaj i
da se klasifikuju dodatno.Zatim je došao WWW. I tako .....

18.Kakve arhitekture mogu imati sistemi za pretragu?
Po arhitekturi se dele na osnovu toga kako su indeksi organizovani na:
Centralizovane sisteme za pretragu -Distribuirane sisteme za pretragu(skupljanje sa različitih lokacija i t ipova fajlova)

19.Koje vrste sadržaja mogu biti pretraživane putem sistema za pretragu?
Vrste sadržaja koje mogu biti pretraživane su : tekstualni(-1.Struktuirani sadržaji 2.Nestruktuirani sad ržaji), linkovani tekstualni, multimedijalni sadržaji(slika,zvuk,video) , ostalo (3d objekti,izvorni kodovi)

20. Koje vrste sadržaja mogu biti pretraživane putem sistema za pretragu?
1.Klasični modeli : Bulov model,Vektorski model,model verovatnoće(probabilistički)
2.Alternativni modeli: Prošireni Bulov model, Fuzzy model, Model neuronske mreže Jezički model

21.Koja je razlika između terma i tokena?
Token predstavlja bilo koju reč koja se pojavi unutar tražene rečenice.
Term predstavlja normalizovanu reč (prebačena u nominativ,bez neki sufiksa ...)
U julu ¢u i¢i na odmor u Gr£ku, na ostrvo Krit. - 11 tokena 9 termova (veliko u i malo u isto, na je isto)
-MOŽDA NIJE DOBRO

22.Šta je tokenizacija i koji problemi postoje u ovoj fazi pretprocesiranja?
Tokenizacija predstavlja izdvajanje pojedinačnih tokena iz rečenice. Problemi nastaju kada neka na prime r, pokušamo da razdvajamo reči
koje su složenice i sastoje se iz 2 reči (Novi Sad,New York), ili kada nam smeta znak interpunkcije (don't - don t -nije dobro) ili kod drugačijeg
formata brojeva, datuma,drugog pisma, različite semantike ukoliko su dve reči jedna pored druge,spojenih reči,čitanja sa desna na levo (Arapi ...),
postavlja se pitanje šta je od toga ispravno.

23. Zašto se vrši „normalizacija“ reči?
Vrši se radi smanjenja termova koji se koriste i da bi se pronašle ekvivalentne reči. Normalizacija je svođe nje termova u isti oblik.

24. Šta je to steming?
Steming predstavlja grub heuristički proces kojim se odsecaju krajevi reči sa ciljem postizanja rezultata št o sličnijim onome kojim postiže pravilna
lematizacija bazirana na lingvističkom znanju. Primer : autamote,automatic,automation --> automat

25.Šta je to lematizacija?
Lematizacija predstavlja "pravilnu" redukciju na osnovni oblik (lemu), tj. svođenje reči da budu istog padež a, vremena roda ...

26.Objasniti Bulov model pretraživanja.
Bulov model koristi operatere AND,OR i NOT za kombinovanje termova u upitu.Ceo dokument se posmat ra kao skup termova i koristi
se Bulova algebra.Daje nedvosmislen rezultat : dokument je ili zadovoljio upit ili ne, nema delimičnog pokl

apanja.

27.Šta je to invertovani indeks i kako se kreira?
Invertovani indeks predstavlja indeks unutar baza podataka koji čuva ceo dokument u slučaju da se unuta r teksta pojavi tražena reč i on omogućava
bržu pretragu celog teksta na štetu povećanje količine procesuiranja dokumenta prilikom dodavanja u baz u podataka.
Kreira se tako što se prvo prikupe dokumenta koja trebaju da se indeksiraju.zatim se svaki dokument pret vara u listu tokena pa se uradi pretrpocesiranje
teksta tj. pravi se lista normalizovanih termova koji će biti u rečniku, a zatim se radi indeksiranje dokumen ata tj. proverava se da li dokument ima tražene termove,
koliko puta se ti termovi pojavljuju i na kraju se vraća rezultat.

28. Objasniti procesiranje upita kod Bulovog modela. - na prezentaciji 4.
Povuku se svi podaci iz dokumenata, zatim se pravi matrica incidencije(pojavljivanja unutar) term/dokume nt i onda se izvlači vektor incidencije(jedan red matrice) za svaki term
i vrši se operacije AND.

29. Šta su pointeri za preskakanje? -PrOVERI
Pointeri za preskakanje ili skip pointeri omogućavaju preskakanje pojava koje svakako neće biti u rezultati ma. Pomoću njih omogućavamo ubrzanje
izračunavanja preseka. Česta praksa je da se od liste pojava dužine P napravi koren iz P pointera jednak e dužine.

30. Šta se može koristiti ako je potrebno podržati upite fraze? - PROVERI
Moguće je koristiti dvorečne indekse (indeksiranje susednih parova reči)kao i filtriranje pogodaka radi izd vajanja dokumenata koji sadrže celu frazu.

31. Šta je to dvorečni indeks?
Predstavlja način indeksiranja gde se dve reči stavljaju pod jednim termom tj. ukoliko reči imaju specifično značenje kada su jedno pored drugog
čuvaju se kao specifičan term.

32.Šta je to pozicioni indeks?
Pozicioni indeks predstavlja dobru zamenu za dvorečne indekse. On pored toga što čuva samu reč čuva i poziciju te reči. Ovo omogućava bržu i efikasniju
pretragu jer ne mora da se prolazi ceo dokument radi pronalaska neke reči.

33.Objasniti Vektorski model pretraživanja.
Vektorski model pretraživanja je algebarski model koji se koristi za pretragu sličnosti dokumenta u odnosu na zadati upit.
Daje mogućnost da dokument bude delimično poklapanje sa zadatim upitom, pa samim tim sortiranje i ra ngiranje dokumenata.
U prvom koraku sam se dokument pretvara u skup vektora reči, a u drugom se ti vektori prebacuju u num erički format koji se kasnije koristi za izvlačenje informacija.
Tačnost nekog dokumenta rangira se na osnovu ugla koji zaklapa sa upitom ili na osnovu kosinusa u opa dajućem redosledu.

34. Šta je ocena relevantnosti?
Ocena relevantnosti predstavlja meru za tačnost za međusobnu vezu između traženog upita i dobijenih re zultata tj. meru koliko se
dokument i upit poklapaju.Ukoliko se term ne pojavljuje u dokumentu, ocena je 0, što se češće pojavljuje t o je ocena veća.

35.Šta je frekvencija terma?

Frekvencija terma predstavlja broj ponavljanja terma unutar nekog dokumenta i ona se koristi za računanj e upit/dokument ocene.

36.Šta je frekvencija dokumenta?
Frekvencija dokumenta predstavlja broj dokumenata u kolekciji u kojima se pojavljuje dati term koji nije tol iko čest
(ukoliko se traži specifičan pojam, a dokument ga sadrži, on je relevantiji u odnosu na druge).Frekvencija dokumenta predstavlja inverznu meru
informativnosti nekog terma (npr. neka specifična reč se pojavljuje puno puta u nekom dokumentima, ti do kumenti su relevantni)

37.Šta je tf-idf?
tf-idf predstavlja proizvod tf i idf težine nekog terma (tf je frekvecija terma dok je idf mera informativnosti te rma)
tf-idf predstavlja proizvod frekvencije terma i mere informativnosti terma.
Ocena poklapanja upita i dokumenta ????
Frekvencija terma

38.Objasniti kreiranje težinske matrice?
Težinska matrica se formira tako što se od željenih termova i željenih kriterijuma pretrage formira tabela g de se svaki term ocenjuje odgovarajućom ocenom
za željeni kriterijum pretrage tako što se računa njegova tf-idf težina.

Svaki dokument je predstavljen vektorom realnih vrednosti
tf-idf težina. Tako imamo V -dimenzionalni vektorski prostor. Termovi su ose prostora. Dokumenti su tačk e ili vektori u
ovom prostoru. Prostor ima visoku dimenzionalnost: desetak miliona dimenzija kada se primeni na veb pr etraživač. Dokumenti
su vrlo retki vektori - vrednosti na većini osa su 0.
Upite, kao i dokumente, možemo predstaviti kao vektore u
vektorskom prostoru. Kada to uradimo onda možemo rangirati dokumente prema njihovoj blizini u vektors kom prostoru sa upitom.
Dakle, blizina nam zapravo predstavlja sli£nost dokumenta
i upita.

39.Koje su razlike između Bulovog i Vektorskog modela pretraživanja?
Glavna razlika između Bulovog i Vektorskog modela je ta što kod Bulovog modela, neki dokument je ili po godan ili nije dok se kod Vektorskog
modela svakom dokumentu daje određeni nivo važnosti u zavisnosti od toga koliko je pogodan za dati upi t. MOŽDA IMA JOŠ

40. Da li se relevantnost odgovora meri u odnosu na informacionu potrebu ili upit?
Relevantnost odgovora se meri u odnosu na željenu informaciju jer većinu korisnika zanima informacija o nečemu, a ne samo
odgovor na prosleđen upit.

41. Šta je preciznost (eng. precision)?
Preciznost predstavlja udeo pronađenih relevantnih dokumenata među pronađenim dokumentima. (našao 15 dokumenata, od toga
10 je relevantna - velika preciznost.
Preciznost = #(prona đeni relevantni)

#(svi prona đeni) = P(relevantan | prona đen)

42. Šta je povrat (eng. recall)?
Povrat predstavlja udeo pronađenih relevantnih dokumenata u odnosu na sve dokumente u kolekciji. (od 1000 dokumenata, vratio 700 - velik povrat).

Preciznost = #(prona đeni relevantni)

#(svi pronađ ni) = P(relevantan | pronađ en)

https://en.wikipedia.org/wiki/Precision_and_recall

43.Šta je F mera i zašto je ona relevantnija od korišćenja preciznosti i povrata?
F mera nam omogućava da merimo kompromis između preciznosti i povrata.

F= (2*A*B)/(A+B)

44.Kako se može vršiti evaluacija performansi sistema za pretraživanje?
Evaluacija performansi se može vršiti na osnovu preciznosti, clickthrough-a(prvi pogodak kod velikih pretraživača ukoliko velik broj
korisnika klikne na prvi rezultat pretrage) , na osnovu laboratorijskih studija ponašanja korisnika (nadgleda se korisnik
tokom pretrage,njegovo ponašanje,zadovoljstvo rezultatima na upit...) ili na osnovu A/B testiranja.

45. Šta je kapa mera?
Kapa mera je mera koliko se međusobno ocenjivači slažu i pomoću nje se meri konzistentnost među ocenjivačima.

k= (P(A) - P(E))/1-P(E)

46. Opisati A/B testiranje?
Testiranje sistema za pretraživanje kod velikih pretraživača namenjeno za unapređenje sistema pretrage. Potreban je velik broj korisnika za testiranje i puno svakodnevnih upita. 99% korisnika dobija staru verziju pretraživača, dok
1% dobija novu verziju sa nekim unapređenjima. Sagledava se vrednovanje stare i nove verzije pomoću neke automatske mere
poput clickthrough-a za prvi pogodak i gleda se ima li poboljšanja.

47.Šta je to Lucene?
Lucene predstavlja open-source skalabilnu biblioteku visokih performansi za full-text pretraživanje sadržaja. Lucene je napisana u Java jeziku i koriste je mnoge
poznate stranice(Wikipedia) i aplikacije(Eclipse)

48.Šta predstavljaju klase Document i Field u Lucene biblioteci?
Lucene dokument je klasa obuhvata naš dokument,njegove podatke i metapodatke i služi za indeksiranje i pretragu.
Dokument se sastoji od polja-Field - gde svako ima ime i sadržaj.
Polja mogu biti označena kao: Indexed -obavezna za pretragu i sortiranje
Tokenized -podela na tokene, analiza te
Stored: sačuvaj originalan sadržaj polja u indeksu
Stored TermVectors: uz dokument sačuvan i invertovani indeks
i mogu biti različitih tipova : TextField,String Field,StoredField,LongField,IntField ...

49.Navesti osnovne karakteristike upitnog jezika Lucene-a.
Upiti unutar upitnog jezika se izražavaju kao stringovi.Parsiranje teksta obuhvata odmah i analizu teksta.
Postojanje dva tipa termova: posebne reči i fraze. Termovi su vezani za polje, gde se pole navodi ispred t

erma i razdvaja se dvotačkom naslov : Bela Griva
Ukoliko se polje ne navede, podrazumeva se default polje, gde se default polje definiše prilikom konstrukc
ije parsera. Termovi mogu da se povežu logičkim
operatorima - naslov : Bela Griva AND korice : Tvrde
Postoje džoker znaci - ? koji menja jedno slovo * zamenjuje više slova,ali ne sme biti na prvom mestu

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