

Pitanja iz oblasti Text Mining i Prepoznavanje entiteta u tekstu

Pitanja iz prezentacije “Osnove Text Mining-a”

1. Kako definišemo Text Mining?
2. U čemu se ogleda složenost nestrukturiranog teksta?
3. Navesti osnovne karakteristike Bag of Words modela.
4. Navesti osnovne karakteristike Vector Space modela.
5. Šta je to Term Document Matrix? Ukratko objasniti.
6. Zašto se i kako radi normalizacija teksta (u kontekstu Text Mining-a)?
7. Objasniti smisao *Zipf*-ovog zakona i njegovu primenu u Text Mining-u.
8. U čemu se sastoji i koja je svrha lematizacije teksta (*lemmatization*)?
9. U čemu se sastoji i koja je svrha stemovanja teksta (*stemming*)?
10. Koja je osnovna ideja IDF (Inverse Document Frequency) metrike i kako se izračunava?
11. Koja je osnovna ideja TF-IDF (Term Frequency - Inverse Document Frequency) metrike i kako se izračunava?
12. Navesti prednosti i nedostatke Vector Space modela.

Pitanja iz prezentacije “kNN klasifikator”

13. Navesti osnovne karakteristike tzv. *instance-based* metoda mašinskog učenja.
14. Navesti osnovne karakteristike kNN (*k Nearest Neighbors*) metode mašinskog učenja.
15. Kako se procenjuje blizina instance prilikom primene kNN (*k Nearest Neighbors*) metode?
16. Kako se određuje parameter k prilikom primene kNN (*k Nearest Neighbors*) metode?

Pitanja iz prezentacije “Klasifikacija – KNN” (sa vežbi)

17. Zbog čega se koristi kNN algoritam sa težinskim faktorima i koji su najčešći pristupi određivanju težinskih koeficijenata?
18. Kada koristiti kNN metodu klasifikacije, koje su prednosti, a koji nedostaci?
19. Šta je Cost Sensitive klasifikacija? Objasniti kako može uticati na preciznost i odziv.

Pitanja iz prezentacije “Prepoznavanje entiteta u tekstu; Semantičko indeksiranje”

20. U čemu se sastoji razlika između prepoznavanja imenovanih entiteta u tekstu (*named entity recognition*) i semantičkog indeksiranja (*semantic indexing, entity linking*)?
21. U čemu se sastoji razlika između semantičkog indeksiranja (*semantic indexing, entity linking*) i identifikacije ključnih tema/konceptata u tekstu ?
22. Koji se atributi (*features*) obično koriste (kao sastavni deo modela nadgledanog m. učenja) za prepoznavanje entiteta u tekstu?
23. Koji su to osnovni oblici procesiranja teksta koji se primenjuju pri kreiranju skupa atributa za prepoznavanje entiteta u tekstu? Ukratko opisati.
24. Šta je glavna prepreka primeni metoda nadgledanog m. učenja za potrebe identifikacije entiteta u tekstu? Koji su alternativni pristupi i zašto?
25. Šta je *Bootstrapping*? Ukratko objasniti kako funkcioniše u kontekstu prepoznavanja entiteta u tekstu.
26. Kako se baza znanja može iskoristiti za kreiranje trening seta za semantičko indeksiranje? Ukratko objasniti.
27. Koji su osnovni koraci procesa semantičkog indeksiranja u slučaju pristupa koji kombinuju mašinsko učenje i baze znanja?
28. Proces semantičkog indeksiranja sastoji se iz tri osnovne faze. Ukratko opisati prvu fazu ovog procesa: koji je njen zadatak, i kako se obično realizuje?
29. Proces semantičkog indeksiranja sastoji se iz tri osnovne faze. Ukratko opisati drugu fazu ovog procesa: koji je njen zadatak, i kako se obično realizuje?