Praktični zadatak iz predmeta Vještačka inteligencija

27. oktobar 2020.

1. (10 bodova) Implementirati u Python-u program koji treba da obezbijedi preuzimanje skupa dokumenata sa adrese https://www.onecle.com/ i njihovo snimanje u txt formatu. Dokumenti su klasifikovani u kategorije.

   Parametri za ovu proceduru su lokacija u koju se dokumenti snimaju (obavezan parametar), lista kategorija iz koje se dokumeti preuzimaju kao i minimalni broj dokumenata u kategoriji.

   Ako lista kategorija nije zadata preuzimaju se dokumenti svih kategorija. Ako je zadat minimalni broj dokumenata preuzimaju se dokumenti iz kategorija koje sadrže najmanje toliko dokumenta.

   Lokacija u koju se dokumenti snimaju treba da bude organizovana kao folder sa podfolderima za svaku kategoriju.

2. Dokumenti se predstavljaju kao N-grami sa sljedećim weighting šemama:
   a. Bag of words
   b. TI-IDF
   c. Positive Pointwise mutual information
3. (30 bodova) Implementirati u Python-u sistem za:
   a. klasifikaciju dokumenata
   b. novelty detection zasnovan na pouzdanosti klasifikatora
   c. simulaciju search engine-a. Korisnik zadaje upit koji se sastoji od proizvoljnog broja riječi a sistem vraća listu najsličnijih dokumenata
   d. first page detection – dokumenti dolaze iz stream-a, stranicu po stranicu, bez jasne odrednice o početku i kraju dokumenta

4. (30 bodova) Napisati seminarski rad koji treba da sadrži:
   a. prikaz osnovnog rješenja za postavljeni zadatak
   b. detaljan opis vašeg rješenja
   c. detaljan opis eksperimentalnog protokola i detaljnu diskusiju rezultata poređenja osnovnog rješenja sa vašim pristupom.

Zadatak se radi samostalno i usmeno se obrazlaže u januarskom ispitnom roku.