1. Ekstrakcija tabela sa PDF dokumenata

U opštem slučaju treba implementirati modul koji prepoznaje tabelu i generiše listu za svaki red u tabeli

1. Detekcija nove klase, klasifikuju se dokumenti, sistem treba da prepozna da dokument ne pripada ni jednoj postojećoj klasi, konstruisati i implementirati algoritam zasnovan na tehnikama iz poglavlja Finding similar itemsets, uporediti sa algoritmima predloženim u litaraturi
2. Ekstrakcija sadržaja sa PDF dokumenata, implementirati modul koji se obučava na skupu PDF dokumenata jedne kategorije sa zadatkom prepoznavanja određenog broja polja: datum, adresa, total amount itd.
3. Klasifikacija dokumenata kada se predstavljaju sa k-shinglovima, poređenje sa bag-of-words pristupom
4. Klasifikacija dokumenata na nivou layouta