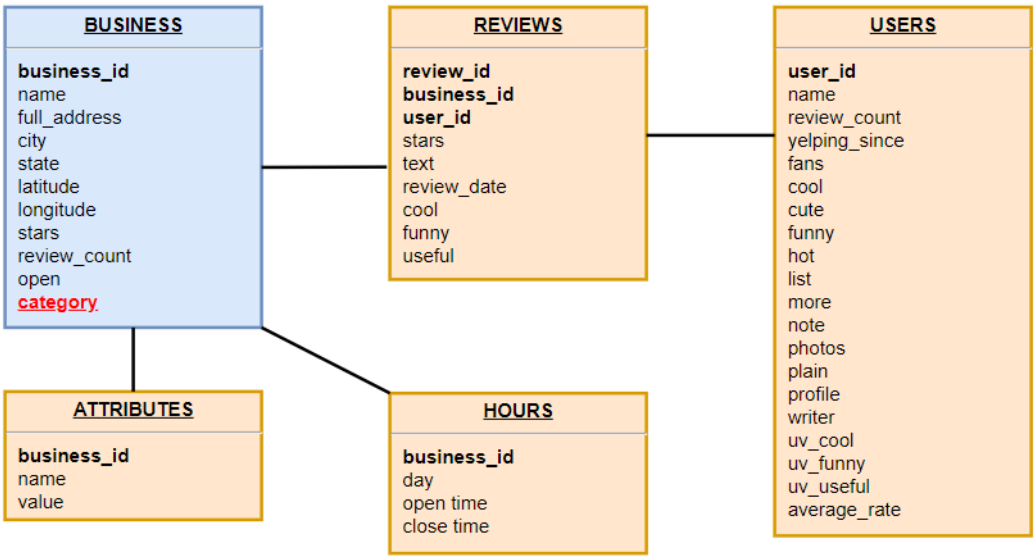
Vaje 14. 4. 2021

Oglejmo si podatke *Yelp*. Pri tem si pomagajmo s shemo na sliki 1.



Slika 1: Shema podatkov *Yelp*. Ciljna tabela je obarvana z modro, ciljni atribut pa z rdečo. Ključi, prek katerih so povezane tabele, so odebeljeni.

Naloga 1. Za učenje najprej uporabimo le tabelo BUSINESS. Katere stolpce je smiselno uporabljati za učenje?

Oglejmo si kodo, ki za učenje uporabi le to tabelo in oceni pripadajočo točnost naključnega gozda (ter jo primerja s točnostjo konstantnega modela).

Naloga 2. Kako bi v podatke iz prejšnje naloge vključili še podatke iz tabel

- ATTRIBUTES,
- HOURS?

Vključite jih in preverite, pri kateri podmnožici vključenih tabel je točnost naključnega gozda največja. Sledite navodilom v predlogi.

Bi znali oceniti pomembnost posameznih stolpcev v tabelah?

Naloga 3. Premislek za naslednjič: vključiti tabeli REVIEWS in USERS je nekoliko težje. Zakaj? Kako bi to lahko storili?