

Skupovi za učenje

Skup za učenje generiran je korištenjem komentara o 203 najbolje rangirane pive i 100 najgore rangiranih piva sa Beer Advocate stranice. Komentara u skupu za učenje je bilo oko 250 000. Za svaki od tih komentara poznata je ocjena za miris, okus, izgled i osjećaj te ukupna ocjena. Kako bi se smanjio šum unutar skupa za učenje komentare smo podijelili u tri klase (positive, negative i average). Budući da ocjena uz komentar ne mora biti relevantna u odnosu na to što je napisano u komentaru. Samo komentari sa jako visokom ukupnom ocjenom (veća od 4.75) su označeni kao pozitivni, a komentari sa jako niskom ocjenom (manja ili jednaka 2.5) su označeni kao negativni. Komentari koji nisu upali u niti jednu od ovih klasa su označeni kao srednji(average).

Zbog posebnog formata nekih komentara, gdje su korisnici dali oznaku o kojoj kategoriji pričaju, napravljen je parser koji razdvaja komentar na rečenice koje pričaju o svakoj od kategorija (okus, miris, izgled i osjećaj). Format komentara je sličan slijedećem formatu:

T - dio komentara vezan uz okus(taste)
S - dio komentara vezan uz miris(smell)
F - dio komentara vezan uz osjećaj(feel)
L - dio komentara vezan uz izgled(look)
O - dio komentara vezan uz sve kategorije(overall)

Na temelju takvih komentara, naučeni su klasifikatori vezani uz kategorije. Budući da je dobiven skup rečenica vezanih uz pojedinu kategoriju. Ocjene za pojedine kategorije određivane su na prethodno opisani način, ali se za određivanje klase (positive, negative, average) unutar kategorije koristila ocjena vezana uz tu kategoriju.