7. domača naloga: Klasifikacija besedil z jedrnimi metodami

Luka Krsnik (63110179)

14. junij 2015

1 Uvod

Cilj sedme domače naloge je implementacija napovedovanja, ki se od ostalih naloh razlikuje v tem, da nimamo podanih nekih značilk, preko katerih bi lahko iskali podobnost, pač pa delamo na čistem tekstu.

2 Metode

Support vector machine Ta metoda nam omogoča analizo podatkov in prepoznavanje vzorcev, ki se uporabljajo pri klasifikacijski in regresijski analizi. Njegovo delovanje si lahko predstavljamo tako, da točke postavimo v prostor, nato pa med njimi skušamo najti nekakšne navidezne meje. Pri svoji implementaciji lahko uporablja tudi jedra, ki se med seboj razlikujejo v pristopih k specifičnemu problemu.

3 Rezultati

Za implementacijo te naloge sem uporabljal prijateljičine testne podatke, kjer je bila prva množica besedil v angleščini, druga pa v slovenščini. Pričakovano so bile napake na testnih podatkih zelo majhne (izmed šestih testnih podatkov in 34 učnih je napovedni model pravilno napovedal vsa besedila). Ko sem videl da algoritem deluje, sem iz interneta prenesel po dvajset besedil pesmi skupin Queens (kot predstavnik rock glasbe), Nightwish (metal) in Keane (pop rock). Ko sem zagnal algoritem sem dobil zelo slabe rezultate. Zgodilo se je celo to, da je za vse pesmi napovedal enakega izvajalca (vseh 9 napovedi, iz 51 testnih podatkov). Ker so bili rezultati tako slabi, sem poskusil ločiti slovenske pesmi, ki se ne ločijo preko izvajalcev in zvrsti, pač pa preko vsebine. Tako sem zbral 20 ljubezenskih, 20 otroških in 20 napitnic, ki so ponarodele. Tudi tokrat so bili rezultati izjemno slabi, zato sem obupal nad besedili pesmi. Namesto tega sem zbral 20 programskih kod, ki so zapisane v pythonu, 20 v javi in 20 v C-ju, ki je sintaktično podoben javi. Tokrat je algoritem deloval. Sicer rezultati niso bili tako dobri kot pri jezikih, vseeno pa je model pravilno napovedal 6 kod izmed devetih. Po pričakovanjih se je najbolj motil med Javo in C-jem.

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.