Klasifikacija tekstova temeljena

na skupovima riječi

KAO IZVOR TEKSTOVA sam koristio članke ONLINE dnevnog časopisa 24SATA??. U početku sam htio da PROGRAM KAO IZVOR TEKSTA može primiti članke iz više različitih časopisa, no bilo je teško pronaći i prilagoditi zajedničke KATEGORIJE TEMA časopisa. Također izvlačenje samog teskta članka bi bilo kompleksnije jer svaka MARKA ČASOPISA ima svoj različiti dizajn HTML koda, te bi uklanjanje irelevantnih KOMPONENTI: headere, footere, reklame... bila VELIKA GLAVOBOLJA. NO, KAO ŠTO je već spomenuto, program će EVALUIRATI UNEŠENI HTML kod, uklanjajući SKRIPTE I STLEOVE, ali će zadržati riječi vidljive korisniku na stranici koje nemaju veze sa samom temom članka: reklame, footere...

Metode što veće preciztnosti: NTORKE, uklanjanje NEKONSTENSTULNIH RIJECI, stemming, kombinacije

Za uklanjanje riječi koje nemaju kontenstualnu važnost napravio sam tekstualnu datoteku koja sadrži što više zamjenica, BROJEVA?, PRILOGA?, prijedloga, veznika, čestica i usklika. Tekstualna datoteka je ručno pregledana te su izostavljene riječi koje bi ipak mogle imati KONTEKSTUALNU VAŽNOST. Jedan primjer je riječ „oko“ koja je i prijedlog ali i imenica.

Jedan od problema na koji sam naišao je odluka gdje neka „rečenica“ prestaje i gdje sljedeća počinje. Problem nije bio u klasičnim rečenicama nego u