Verslag Practicum 2

Martijn Koenis (3770214) Jordi Vermeulen (3835634)

Samenvatting

1 Probleem en corpus omschrijving

We hebben hetvolgende clasificatieprobleem onderzocht: het indelen van songteksten in de genres Blues, Country, Folk, Gospel, Metal, R&B, Rap en Soul. Om dat te kunnen doen hebben we een corpus verzameld van 50 songteksten per genre (in totaal 400). Deze songteksen hebben we op internet gezocht. Ter inspitatie voor de keuze van nummer hebben we indien nodig gebruik gemaakt van hitlijsten en verzamelCD's. We hebben ervoor gezorgt dat we geen duplicate nummers hebben door een script te schrijven die sterk op elkaar lijkende txt-bestanden detecteerd en rapporteerd.

2 Voorbewerking

Voordat we de teksten gaan analyseren halen we eerst alle whitespace weg zodat die niet interfereren met de uitkomst. Vervolgens vervangen we alle hoofdletters door normale letters omdat die voor ons doeleinde hetzelfde betekenen. Daarna verwijderen we alle engelse stopwoorden omdat deze in alle genres voorkomen en dus niet interessant zijn om te bekijken. Ook verwijderen we alle getallen in de tekst.