

# Verslag Practicum 2

Martijn Koenis (3770214)

Jordi Vermeulen (3835634)

## Samenvatting

### 1 Probleem en corpus omschrijving

We hebben het volgende classificatieprobleem 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 songteksten hebben we op internet gezocht. Ter inspiratie voor de keuze van nummer hebben we indien nodig gebruik gemaakt van hitlijsten en verzamel-CD's. We hebben ervoor gezorgd dat we geen dubbele nummers hebben door een script te schrijven die sterk op elkaar lijkende txt-bestanden detecteert en rapporteert.

### 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.

NB, 25 [1] [2] [3] [4] [5] [6] [7] [8] [1,] 0.0 73.6 88.8 95.6 92.4 95.6 100.0 80.4 [2,] 66.8 0.0 75.6 86.4 90.0 92.4 100.0 70.8 [3,] 56.4 58.0 0.0 68.4 61.2 88.8 99.6 82.8 [4,] 85.6 74.0 82.4 0.0 61.2 86.8 97.2 80.4 [5,] 89.6 81.2 84.8 74.4 0.0 88.8 98.0 84.4 [6,] 69.6 60.8 73.2 73.6 71.6 0.0 80.0 58.0 [7,] 90.8 88.0 91.2 88.8 88.4 81.2 0.0 90.4 [8,] 60.8 62.0 73.6 79.2 82.8 81.6 100.0 0.0

KNN, 25 [1] [2] [3] [4] [5] [6] [7] [8] [1,] 0.0 53.2 59.6 89.2 80.8 99.6 100 86.0 [2,] 74.4 0.0 58.8 90.4 78.8 98.4 100 79.2 [3,] 77.2 54.4 0.0 93.2 69.2 99.2 100 71.6 [4,] 76.8 43.2 20.4 0.0 54.8 91.6 100 78.0 [5,] 85.2 62.8 54.0 66.0 0.0 99.6 100 88.8 [6,] 26.0 27.6 32.0 54.0 39.2 0.0 100 24.4 [7,] 22.4 21.2 23.2 34.0 27.6 16.4 0 14.4 [8,] 36.4 36.8 53.2 68.4 54.8 87.6 100 0.0

SVM, 25 [1] [2] [3] [4] [5] [6] [7] [8] [1,] 0.0 69.2 78.0 92.8 96.0 86.8 100.0 70.0 [2,] 67.6 0.0 75.2 82.0 89.6 86.0 100.0 66.0 [3,] 76.0 51.6 0.0 78.8 69.6 88.0 96.8 76.0 [4,] 80.8 80.8 77.2 0.0 64.0 78.8 94.8 80.8 [5,] 83.2 83.6 75.2 71.2 0.0 81.2 98.8 82.8 [6,] 72.4 65.2 80.0 69.6 77.6 0.0 84.8 56.8 [7,] 90.4 86.4 92.0 90.0 93.2 68.8 0.0 86.4 [8,] 59.6 62.0 76.8 88.4 79.2 72.0 99.2 0.0