Verslag Practicum 2

Martijn Koenis (3770214) Jordi Vermeulen (3835634)

Samenvatting

1 Problem- and corpus description

We have investigated the problem of classifying song texts of the following genres to their genre: Blues, Country, Folk, Gospel, Metal, R&B, Rap and Soul. In order to do this we have gathered a collection of 50 song texts for each of the 8 genres. We have gathered these from the internet by hand. We used hit-lists and top-n collection albums for inspiration if needed. We made sure that no duplicate songs were used by running a script that detects and reports similar txt-files.

2 Pre-processing

We strip the whitespaces from all the song texts because we do not want whitespaces interfering with the test results. After that we change all capitals characters to lower characters because for our purpose both variants are the same. Thereafter we removed all English stopwords from the texts since they are common in all genres. Subsequently we remove all numbers from the texts.