OH OH OH WHOAH!
TOWARDS AUTOMATIC TOPIC DETECTION IN SONG LYRICS

rsa SAT
Studio Smart Agent Technologies
Research Studios Austria
Forschungsgesellschaft mbH

Florian Kleedorfer
Research Studios Austria
florian.kleedorfer@researchstudio.at

Peter Knees
Johannes Kepler Universität Linz
peter.knees@jku.at

Tim Pohle
Johannes Kepler Universität Linz
tim.pohle@jku.at

Department of Computational Perception

Research Goals

Creation of meaningful data from song lyrics for music browsing and retrieval.

Approach: detection of *topics.*

Assessment of the usefulness of the identified topics.

Computation of topic relations for future use in browsing systems.

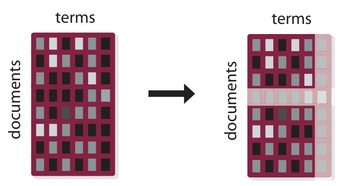
Algorithm

1 Creation of a Term-Document Matrix

All lyrics documents are read and a matrix containing the terms in the columns and the documents in the rows is created.

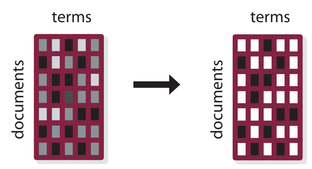
2 Term Selection

Infrequent terms and terms found in stopword lists are removed from the matrix. This can also lead to the removal of whole documents.



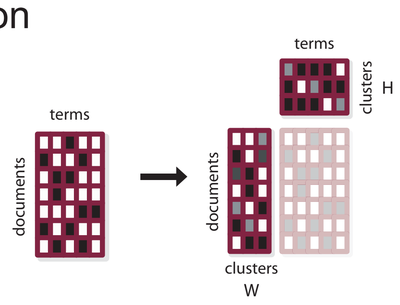
3 Weighting of the Terms

Binary weighting is applied to the term-document matrix, i.e., all nonzero entries are set to 1.



4 Identification of Topics by Non-negative Matrix Factorization

The term-document matrix is approximated by the product of two matrices, which can be interpreted as a clustering of the terms and the documents into topic clusters.

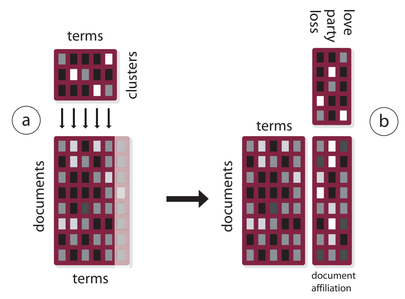


5 Manual labeling of the Topics

The topics, which are essentialy weighted term lists, are given labels by test subjects.

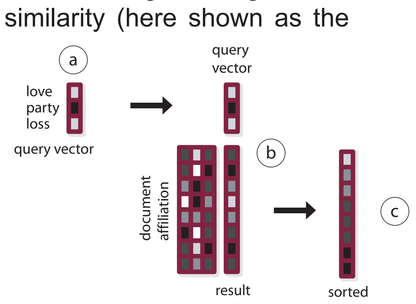
6 Calculation of Topic Membership for each Song

From the original term-document matrix, the columns that have been removed during term selection are dropped. , Then, the matrix product of this matrix with the term cluster matrix yields a topic membership matrix (b).

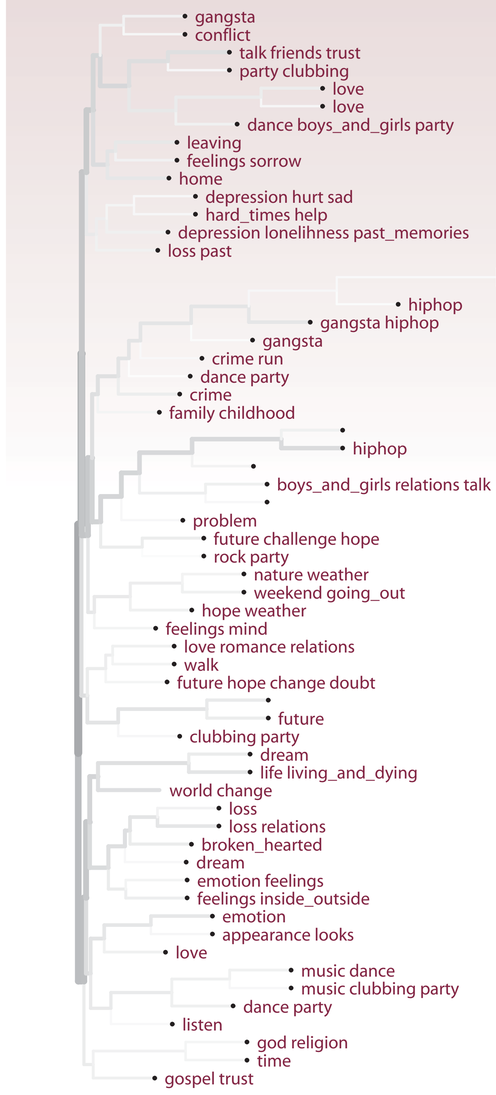


7 Querying by Specification of Topic Weights

For querying, a vector is defined containing the weights of the desired topics (a). The cosine similarity (here shown as the matrix product) between this vector and the row vectors of the topic membership matrix defines the strength of match between each document and the query.(b). The resulting vector is sorted (c).



Visualizations of the Topic Space

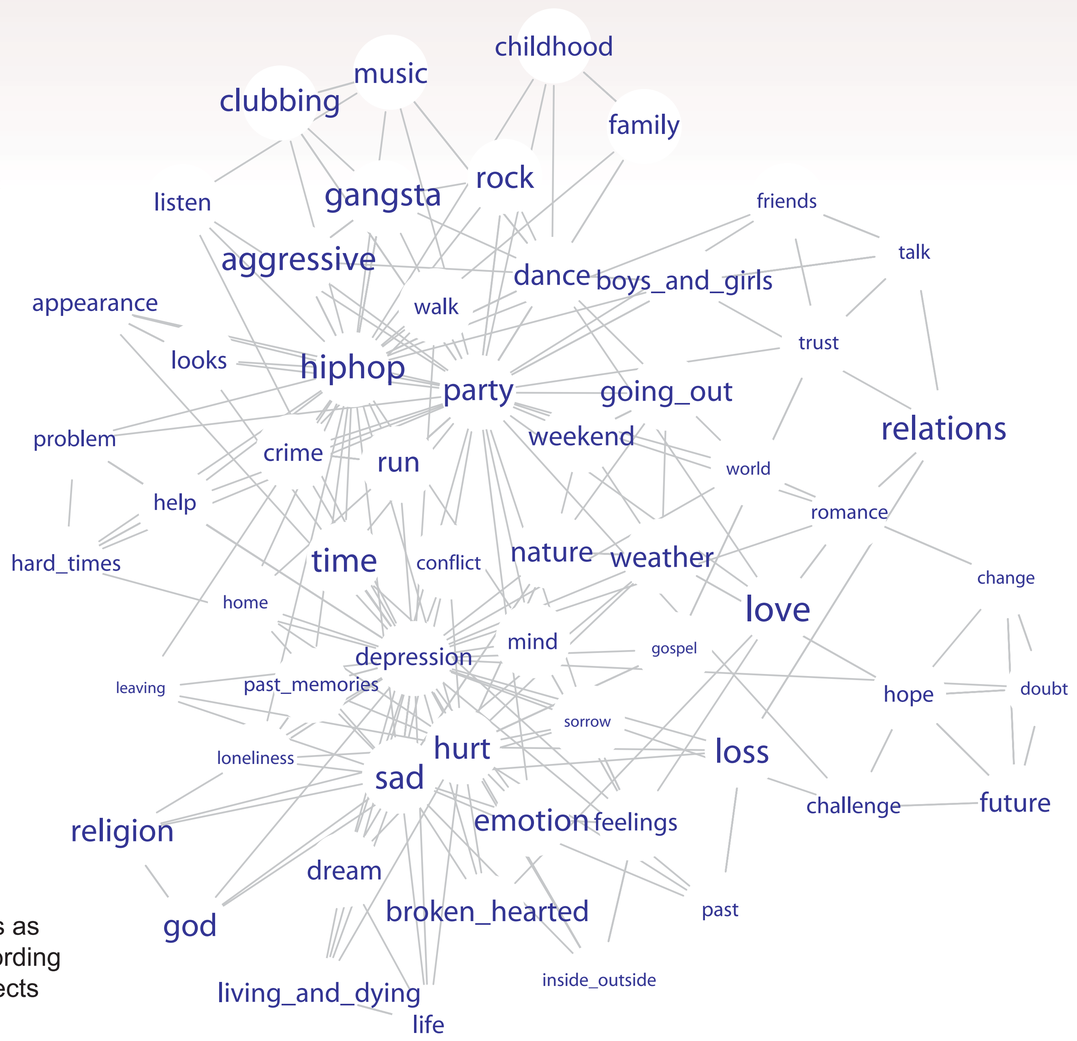


Cluster Hierarchy

Above: Hierarchy of the topic clusters as they are successively combined according to their similarity. Line Thickness reflects the number of songs in the clusters.

Tag Graph

Below: Graph visualization of the tags provided by the test subjects. Tag relatedness is computed from the cluster membership values of the songs. Related tags are connected. Font size reflects importance.



Evaluation

Question: Are Identified Topics Plausible?

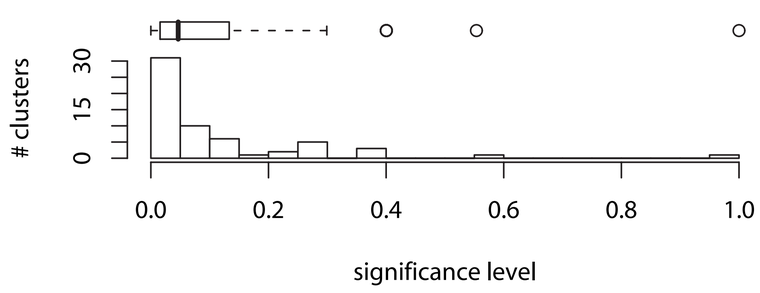
Tagging of Topics by Test Subjects

Phase 1: Test subjects provide tags for each cluster.

Phase 2: Test subjects select best 2 from the tags collected in first phase. Tags thus receive a score.

Comparison with Random Baseline

For the winning tag of each cluster, the prior probability of reaching its score assuming random user choices is computed.

Result: Topics are Plausible

For more than half of the topics, the most important tag is selected at a significance level of less than 5%, and almost 3/4 of the clusters are below the 10% significance level.