* input: dictionary: files, query: 1 song
* First build our dataset = dictionary
  + songselection:
    - scan tracks\_with\_tags; only use this songs
    - songs into our dataset if a song is tagged at least twice with one tag or at least tagged with two tags of the same group (there exists different mood groups). In this step we already delete duplicates.
    - tokenize songs: we create songs as vectors s=(t1…tn). with attributes id, group, quadrant.
    - build dictionary with tags / groups / quadrants : three levels of granularity are defined: tags, mood groups and mood quadrant.
* Next we do some statistics
  + by calculating the term distribution for each document and classes (mood quadrants) by using five different term weighting schemes: a binary term weighting, the term-frequency (tf), inverted document frequency (tf-idf), BM25, and delta tf-idf.
  + Therefore we use the vector-space-model and match every song to a vector s=(t1…tn). Moreover we use stemmed terms. Annotation: we don’t analyze the distribution over tag groups and mood tags as it is not done in the paper.
  + we compute the number of distinct terms, unique terms and number of unique terms per song and compute the standard deviation for them.