Home work until November 8, 2016 (23h59):

Task 1: Data acquisition, representation, and similarity computation for CB (5 pts)
First, discuss and decide on one (or more) data source(s) to acquire external data on music artists, e.g., web pages about the artists returned by a search engine, lyrics of the artists, or microblogs about the artists. Write a **crawler that automatically fetches the respective ("music context") data** for all artists in the collection C1ku.

Create a **representation of the artists inferred from your crawled external data**: e.g., term weight vectors in the vector space model or co-occurrence information. Consider post-processing of your data, e.g., in case of term weight vectors, using a dictionary or performing stopword removal and casefolding. Then, **compute pairwise similarities between the music items**: e.g., cosine similarity on term weight vectors, co-occurrence likelihood, or set-based Jaccard index. This should result in an artist-artist-similarity matrix.

In the report, describe your approaches and choices made for the data acquisition and representation, in particular: Which data source(s) did you use and why? Did you encounter any problems? How did you implement your crawler? How did you perform post-processing of the crawled data? How did you represent artists? And which similarity measure did you use to create the artist-artist-matrix?

Task 2: Content-based recommender (5 pts)
Implement a content-based recommender (as function `recommend_CB`) using the similarity matrix just created and evaluate it in your evaluation framework (average precision, average recall, average F-scores) using *different numbers of recommended artists*.

In the report, particularly describe your aggregation strategy for recommended artists, i.e., how do you deal with artists that are "recommended" by more than one of the artists listened to by the target user? Did you just consider their frequency? Or their similarity to the target user's artists listened to? Or both? Since you should ensure that a certain number of artists is recommended, you have to carefully think about the above questions!

Task 3: Hybrid recommender (3 pts)
Implement at least one way to integrate CF and CB recommendations and describe in the report how you merge the individual recommendations of both approaches in an intelligent way, for instance, so that none of the recommenders is penalized. In your implementation, again ensure that a certain number of artists is recommended.

Task 4: Evaluation (2 pts)
Using the evaluation framework you implemented in the last home exercise and the provided dataset C1ka (on Mediacube), perform **10-fold cross-fold validation** on the user level. Evaluate the CB and the hybrid recommendation algorithms you implemented and compute average **precision, recall, and F1 measure**.

Report these performance measures for the two recommenders (you can of course also include results for your CF and baseline recommenders). To this end, create precision/recall plots (such as the one shown in `precision_recall_plot.pdf` on MediaCube) that allow to easily analyze the trade-off between recall and precision for the different approaches. Create these plots by varying the number of recommended artists.