Home work until November 20, 2016 (23h59):

For the final home work, use the standardized dataset provided on MediaCube in order to be able to compare results between groups (and challenge the other groups ☺). The dataset is available in files C1ku.zip (and thanks to a group from previous year C1ku_users_extended.zip and C1ku_artists_extended.zip). The zip files contain the following files and content:

C1ku_UAM.txt	1100 x 10122 user-artist-matrix containing raw Last.fm play counts
C1ku_idx_users.txt	1100 user-ids corresponding to the rows in C1ku_UAM.txt, which can be mapped to the ids and metadata provided in LFM1b_users.txt
C1ku_idx_artists.txt	10122 artist-ids corresponding to the columns in C1ku_UAM.txt, which can be mapped to the ids and names provided in LFM1b_artists.txt
LFM1b_users.txt	user-id \t Last.fm_user-name \t country_code \t age \t gender
LFM1b_artists.txt	artist-id \t artist-name
C1ku_users_extended.csv	user \t age \t country \t long \t lat \t gender \t usertype
C1ku_artists_extended.csv	artist-id \t artist \t mbid

Task 1: Demographic filtering (5 pts)
Using the demographic information provided in C1ku_users_extended.csv, implement an extension to your collaborative filtering recommender (called `recommender_DF`), that restricts nearest neighbor search to users in the same country, of same gender, and similar age (it is up to you how you define similarity in terms of age). This means, you should implement (at least) three recommenders: one for country, one for gender, and one for age. You can also create a fourth that combines these three aspects. Pay attention that you need a sufficient amount of users in each category in order to make meaningful predictions. Use the same artist aggregation technique as you did in the CF recommender.

Task 2: Comparison of all approaches and discussion of results (10 pts)
Use your evaluation framework to perform cross-fold validation on the user level. Evaluate all your recommendation algorithms on the dataset (C1ku from MediaCube) using 10-fold CV, i.e., for each user, split their unique artists into 90% training artists and 10% testing artists and iterate 10 times to cover all possible 9:1 splits. Compute **average precision, recall, and F1 measure**. Investigate how results change for different numbers of neighbors k and different numbers of predicted artists n.

The recommender systems to evaluate should at least include the following 11 approaches:

- Random user baseline
- Random artists baseline
- Popularity-based RS
- Collaborative Filtering
- Content-based RS
- Demographic Filtering (country, gender, and age)
- Hybrid (set-based, score-based and rank-based) – If you are curious, you can of course also fuse more than CB and CF (e.g., combine CB+CF+PB).

In the report, include:

(i) **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 and

(ii) **figures plotting F1-measure vs. number of recommended artists** to investigate at which number of artists the different approaches perform best overall.

Then **discuss** in a structured and comprehensive manner the **relationship between recall and precision for the different approaches** and the **relationship between the performance measures and number of recommended items**.

Some ideas of topics you may (and should) address in your detailed discussion:

- Why is there a performance difference between your two baseline recommenders?
- Can you outperform the stand-alone CF and CB recommenders using the hybrid approach? Why? Or why not?
- For your CB recommender, did you encounter any problems during data acquisition? How did you resolve them? To which extent do you think the data source influences the recommendation quality?
- Which of the two, CF and its extension DF, performs better? Speculate about the reasons.
- Among the three DF variants, which one performs best? Any idea why?
- How are recall and precision related to each other for the different approaches and why?