RECOMMENDER SYSTEMS COURSE

KAGGLE CHALLENGE - 2017/18

IL POLLO DEL PROFILO

LUCA BUTERA - 878595

GIADA CONFORTOLA - 898540

BACKGROUND AND OBJECTIVES

A Recommender System is a system whose purpose is to predict the ratings a set of users would give to a set of items, given the ratings over a subset of such items.

Our objective is to develop one for songs recommendation over playlists, aiming to the best MAP score possible.

DATASET ANALYSIS

The dataset contains two main information groups:

Playlists data

created_at	playlist_id	title	numtracks	duration	owner

Seemingly not useful, indeed not used.

Songs data

track_id	artist_id	duration	playcount	album	tags

Useful data, needing some preprocessing.

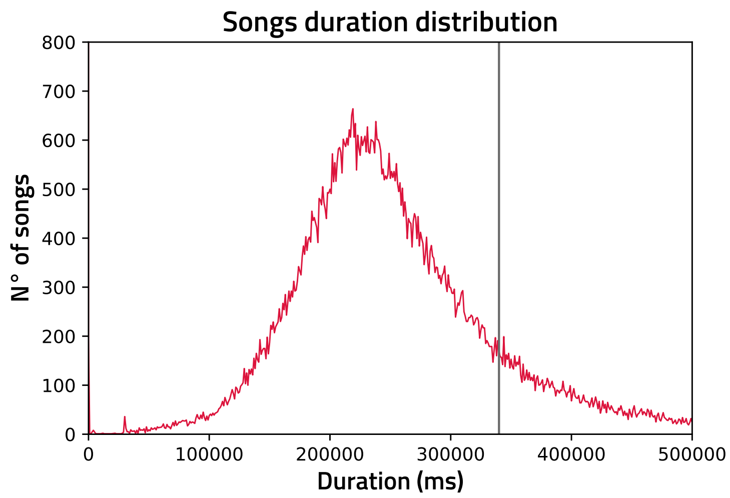
Song-playlist pairs used as they are.

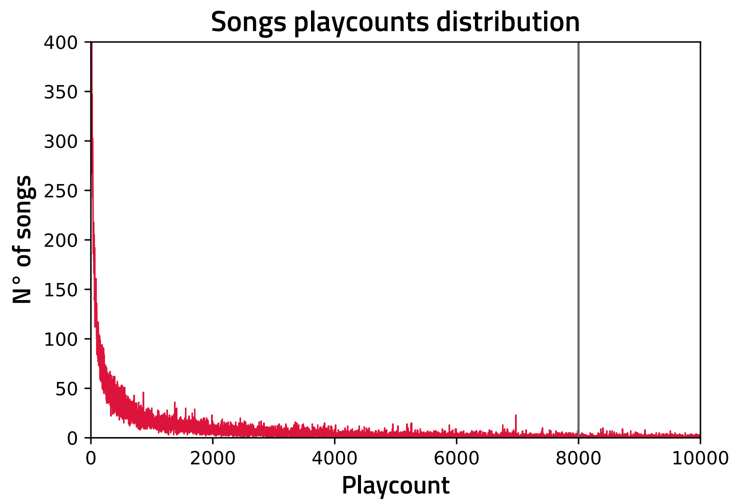
playlist_id	track_id

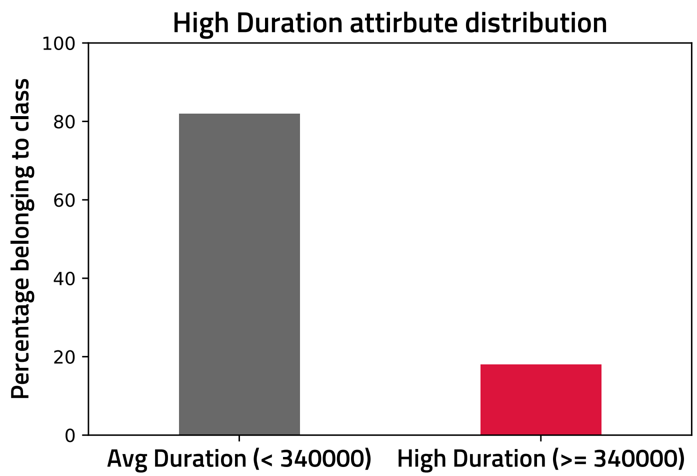
DATASET REFACTORING

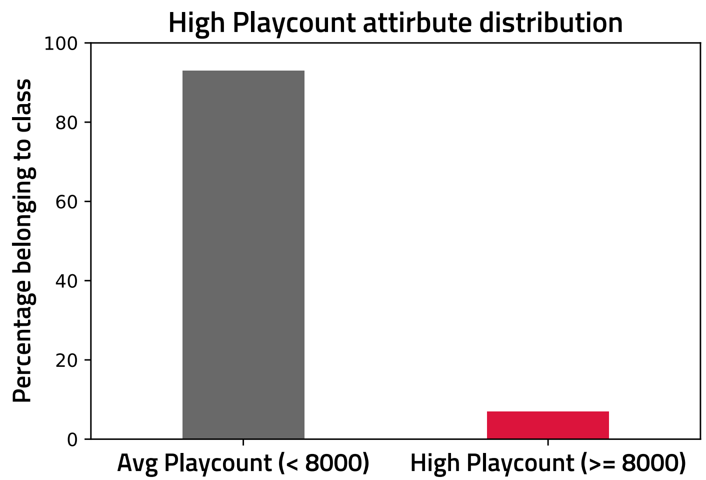
As said, data needed some formatting.

Playcount and Duration have been transformed to be more meaningful.









TESTING

To evaluate recommender performances a train set and a test set have been created from given data.

Playlist containing more than ten tracks

80% | 20%

Remaining songs

Five songs per playlist

Train set

Test set

Cross-validation avoided, due to high computational time and no relevant advantage.

MODELLING CHOICES

We tried different approaches to our problem.

Simple models

Content Based Filtering
User Based Filtering
Item Based Filtering

Complex models

Matrix Factorization
S.L.I.M.

But complex models revealed to be:

Time intensive

A single run required lots of time, as well as writing optimized low level code to speed up computation.

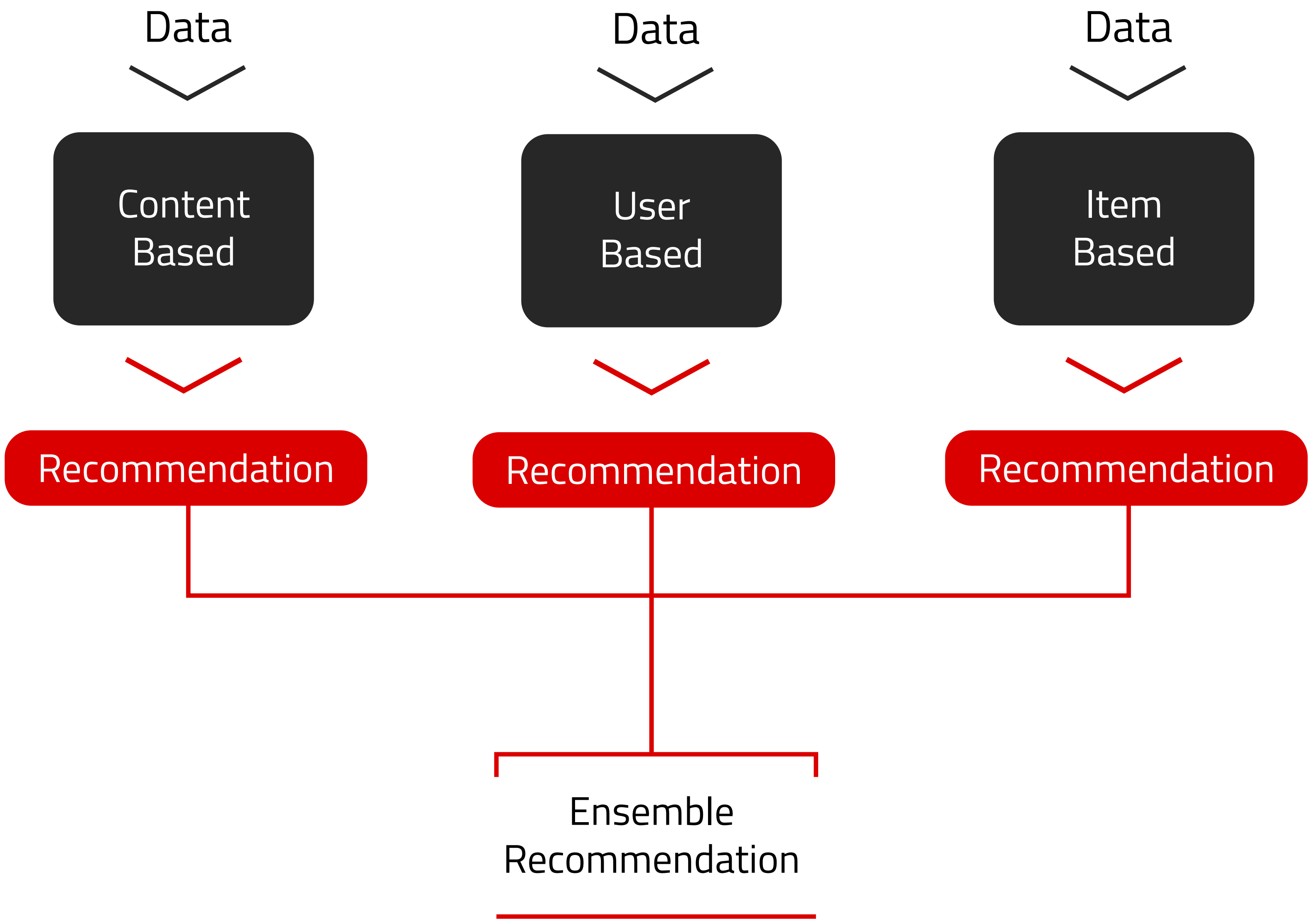
Hard to tune

Obtaining satisfying results relied on tuning a vast number of parameters. This, coupled with the time intensive aspect, was a major disadvantage.

So our choice was to focus on simple models and their optimization.

OUR SOLUTION

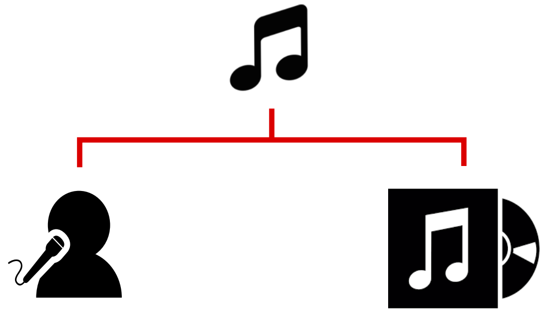
For our final solution we chose an ensemble combining three models, in order to extend their expressive power and achieve better results.



CONTENT BASED

Content Based approach is mostly caracterized by a set of attributes for each item and a similarity measure.

Chosen attributes

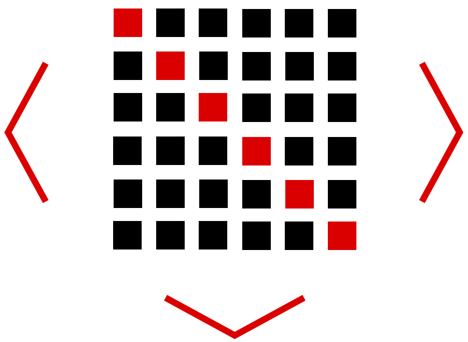


Artist ID

Album

Similarity matrix

Built with dot product



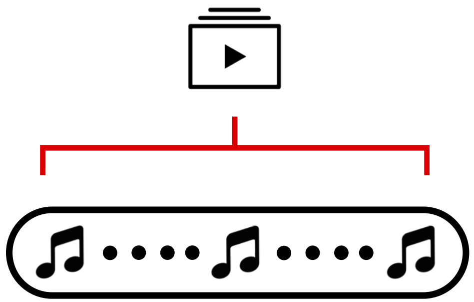
Regularized with IDF

Thresholded to top 120 similarities per item

USER BASED

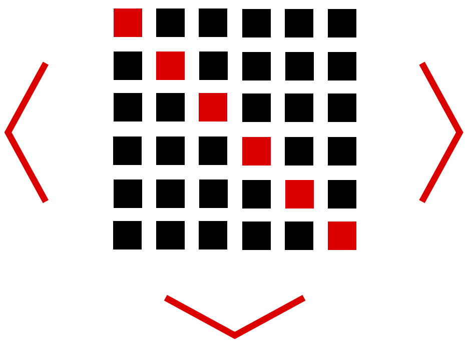
User Based approach takes into account the similarity between users (playlists) to extract recommendations.

Playlist characterization



Similarity matrix

Built with simplified cosine for implicit data sets



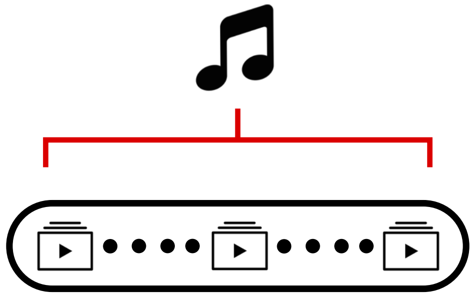
Shrinkage factor of 10

Thresholded to top 10 similarities per item

ITEM BASED

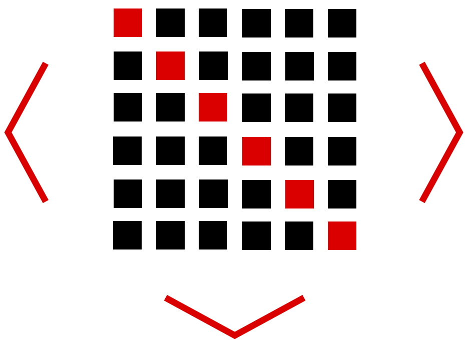
Item Based approach takes into account the similarity between items (songs) to extract recommendations.

Song characterization



Similarity matrix

Built with dot product



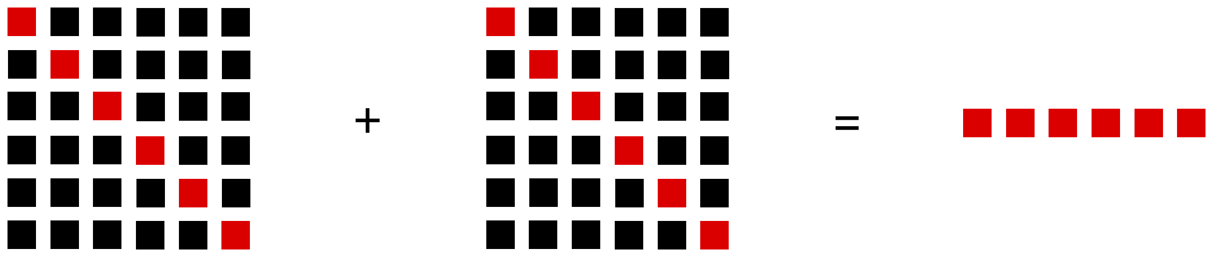
Regularized with IDF

Thresholded to top 140 similarities per item

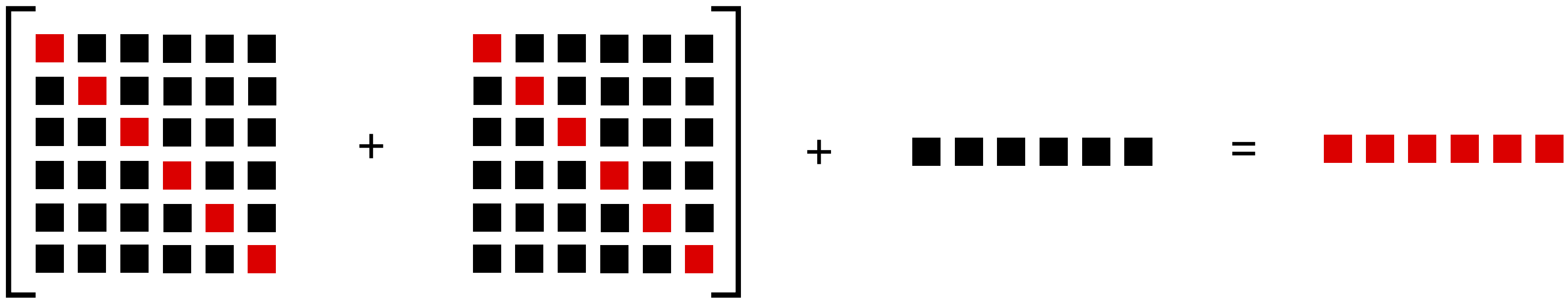
Background and Objectives
Dataset Analysis
Dataset Refactoring
Testing
Modelling Choices
Our Solution
Content Based
User Based
Item Based
Ensembling Choices
Final Solution
Parameters and Results

ENSEMBLING CHOICES

First ensemble idea was to combine similarity matrices.



But not all similarity matrices can be combined this way.
So we tried hybrid combination.

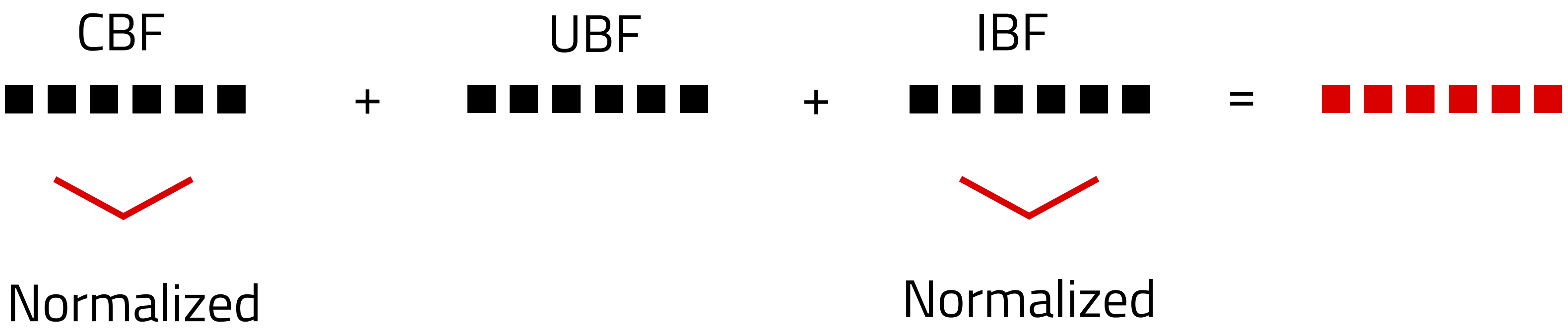


But in the end the best solution was to
directly combine recommendations.

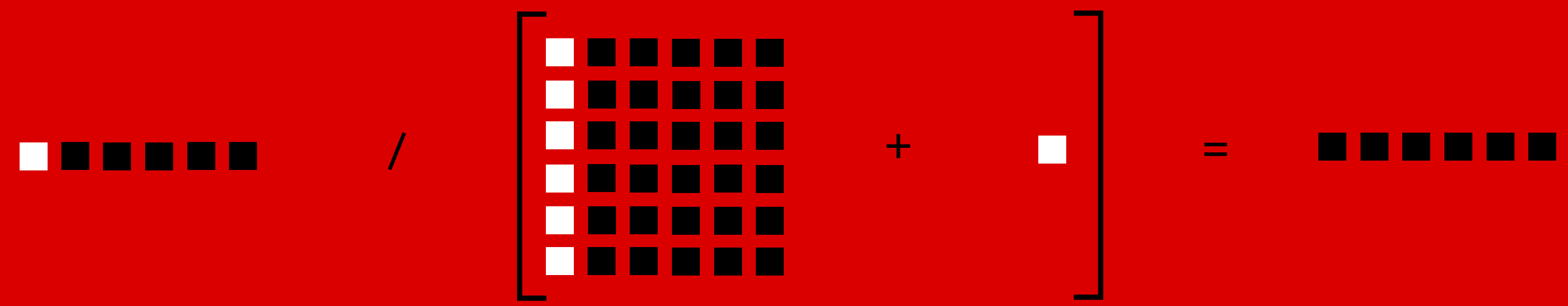
Background and Objectives
Dataset Analysis
Dataset Refactoring
Testing
Modelling Choices
Our Solution
Content Based
User Based
Item Based
Ensembling Choices
Final Solution
Parameters and Results

FINAL SOLUTION

Final results were obtained by directly combining the normalized recommendations of each model.



Applied normalization:
similarity column sum plus shrinkage



PARAMETERS AND RESULTS

Best score parameters

CBF

Attributes: artist_id, album
S measure: dot product
S threshold: 120
IDF: True

UBF

S measure: implicit cos
Shrinkage: 10
S threshold: 10
IDF: False

IBF

S measure: dot product
S threshold: 140
IDF: True

Ensemble

CBF coefficient: 0.4
UBF coefficient: 0.1
IBF coefficient: 0.5

CBF Shrinkage: 60
IBF Shrinkage: 10

Test MAP score

0.0899

Kaggle MAP score

0.097

THANKS FOR WATCHING

▶ Next team