HOMEWORK 2 – DATA MINING FOR BUSINESS AND SOCIETY

Group members:

- ***Alessandro Taglieri, 1890945***
- ***Guglielmo Lato, 1257406***

Part 1 – Recommendation System Evaluation

- Part 1.1 :
  The result of applying all recommendation system algorithms provided by surprise library on the given dataset (taking 5 folds).

  - **SVDpp:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.8916	0.8925	0.8881	0.8968	0.8935	0.8935	0.0038
Fit time	722.92	723.28	717.12	725.32	729.16	723.56	3.91
Test time	11.20	11.15	11.43	8.99	8.31	10.22	1.30

  - **KNNBaseline:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9052	0.9102	0.9065	0.9101	0.9108	0.9086	0.0023
Fit time	0.55	0.57	0.55	0.60	0.57	0.57	0.02
Test time	5.70	6.06	6.31	6.19	5.66	5.98	0.26

  - **SVD:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9087	0.9085	0.9078	0.9110	0.9130	0.9098	0.0019
Fit time	8.79	8.85	9.23	9.17	8.82	8.97	0.19
Test time	0.29	0.31	0.26	0.22	0.23	0.26	0.03

  - **BaselineOnly:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9176	0.9201	0.9174	0.9209	0.9225	0.9197	0.0020
Fit time	0.08	0.08	0.08	0.08	0.08	0.08	0.0
Test time	0.17	0.15	0.17	0.15	0.16	0.16	0.01

  - **SlopeOne:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9202	0.9256	0.9212	0.9239	0.9255	0.9233	0.022
Fit time	2.37	2.32	2.35	2.41	2.51	2.39	0.07
Test time	8.16	8.30	8.53	8.28	7.81	8.21	0.23

HOMEWORK 2 – DATA MINING FOR BUSINESS AND SOCIETY

- **KNNWithMeans:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9288	0.9348	0.9313	0.9332	0.9345	0.9325	0.0022
Fit time	0.50	0.49	0.49	0.49	0.49	0.49	0.00
Test time	4.99	5.03	5.21	5.27	4.79	5.06	0.17

- **NMF:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9354	0.9371	0.9350	0.9347	0.9378	0.9360	0.0012
Fit time	7.60	7.77	8.02	7.81	7.45	7.73	0.19
Test time	0.23	0.23	0.25	0.20	0.22	0.22	0.02

- **CoClustering:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9357	0.9474	0.9354	0.9411	0.9341	0.9387	0.0050
Fit time	1.33	1.31	1.34	1.34	1.29	1.32	0.02
Test time	0.22	0.18	0.19	0.18	0.18	0.19	0.02

- **KNNBasic:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	0.9468	0.9517	0.9505	0.9530	0.9560	0.9516	0.0030
Fit time	0.50	0.47	0.49	0.46	0.46	0.48	0.02
Test time	4.51	4.67	4.79	4.80	4.51	4.66	0.13

- **NormalPredictor:**

	Fold 1	*Fold 2*	*Fold 3*	*Fold 4*	*Fold 5*	*Mean*	*Std*
RMSE (testset)	1.5131	1.4988	1.5128	1.5159	1.5128	1.5107	0.0061
Fit time	0.14	0.16	0.14	0.14	0.14	0.14	0.01
Test time	0.22	0.22	0.22	0.22	0.22	0.22	0.00

- **Number of CPU-cores used**: 12
- To use all cpu-cores we used the parameter n_jobs = 12 in the cross_validate command. 12 is the number of cpu core of my pc.

HOMEWORK 2 – DATA MINING FOR BUSINESS AND SOCIETY

- Part 1.2:
  Performing hyper parameter tuning:

  - **SVD optimization:**
    - Grid of Parameters used to increase the performaces:
      - N_factors: [50,100,125,150,200];
      - Init_mean: [0.1,0.15]
      - Lr_all – learning rate for all the parameters: [0.005,0.01,0.025]
      - Reg_all – regularization term for all parameters: [0.02,0.005,0.1]

    - Best Configuration:
      - N_factors: 150
      - Init_mean: 0.15
      - Lr_all: 0.025
      - Reg_all: 0.1

    - Mean RMSE: 0.8835
    - Optimization time: 5 min e 55 sec

  - **KNNBaseline optimization:**
    - Grid of Parameters used to increase the performaces:
      - K – max number of neighbours to take into account for aggregation: (1,60,2)
      - min_k – The minimum number of neighbours to take into account for aggregation: [1,2,3,4,5,6,7,8,9,10,11]
      - Similiarity options – a disctionary of options for the similiarity:
        - Name: [“coisine”, ”msd”, ”pearson”, ”pearson_baseline”]
        - User_based: [true, false]

    - Best Configuration:
      - K : 37
      - Min_k: 11
      - Similiarity options:
        - Name: “pearson_baseline”
        - User_based: False

    - Mean RMSE: 0.8864
    - Optimization time: 9 min 45 sec
  - **Number of CPU-cores used:** 12
  - To use all cpu-cores we used the parameter n_jobs = 12 in the cross_validate command. 12 is the number of cpu core of my pc.

HOMEWORK 2 – DATA MINING FOR BUSINESS AND SOCIETY

Part 2 – Local Community Detection with PerosanlizedPageRank:

This following table represents data that *output.tsv* contains. It is sorted by ascending values of the first columns (*Book*) and then by ascending values of the second column (*Character*).

Book	Character	Dumping_factor	Exponent	Conduntance	Baratheon	Lannister	Stark	Targaryen	Total
book_1.tsv	Daenerys-Targaryen	0.9	1.0	0.07801418439716312	0	0	0	3	22
book_1.tsv	Jon-Snow	0.9	1.0	0.07913669064748201	6	6	11	5	166
book_1.tsv	Samwell-Tarly	0.8	1.0	0.07913669064748201	6	6	11	5	166
book_1.tsv	Tyrion-Lannister	0.85	1.0	0.07913669064748201	6	6	11	5	166
book_2.tsv	Daenerys-Targaryen	0.9	1.0	0.09859154929577464	0	0	0	3	18
book_2.tsv	Jon-Snow	0.75	1.0	0.08527131782945736	0	0	1	4	28
book_2.tsv	Samwell-Tarly	0.55	1.0	0.08527131782945736	0	0	1	4	28
book_2.tsv	Tyrion-Lannister	0.95	1.0	0.09865470852017937	8	6	5	6	160
book_3.tsv	Daenerys-Targaryen	0.9	0.8	0.07142857142857142	0	0	0	2	25
book_3.tsv	Jon-Snow	0.95	1.0	0.06198347107438017	0	0	1	1	74
book_3.tsv	Samwell-Tarly	0.95	1.0	0.06079664570230608	0	0	1	1	71
book_3.tsv	Tyrion-Lannister	0.95	1.0	0.07903780068728522	7	9	10	10	204
book_4.tsv	Daenerys-Targaryen	0.95	1.0	0.07157894736842105	5	8	4	11	298
book_4.tsv	Jon-Snow	0.95	1.0	0.06060606060606061	2	0	6	2	181
book_4.tsv	Samwell-Tarly	0.95	1.0	0.08624708624708624	2	0	6	2	177
book_4.tsv	Tyrion-Lannister	0.95	1.0	0.0441025641025641	5	8	4	11	286