

CSE 5400 Interdisciplinary CS — HW3
Due 5pm, March 13, 2014
Submit Server: Class = intercs , Assignment = hw3

To explore how to recommend products in a personalized marketing effort, this assignment attempts to predict how customers would rate movies that they have not watched/rated.

1. Use Java (or C/C++) to implement:
 - (a) basic Intersection algorithm: `Intersection.java` has the `main` method
 - i. if supersets exist, use only supersets (weighted by similarity)
 - ii. if supersets do not exist, but subsets exist, use only subsets (weighted by similarity; for movies not rated, similarity/distance is minimum/maximum)
 - iii. if neither supersets nor subsets exist, use global movie and customer averages weighted by standard deviation
 - (b) k-nearest neighbor algorithm: `KNN.java` has the `main` method (k=5 for `nf`, k=2 for `toy`)
2. Input (data files are on course website):
 - (a) rating file: ratings available to the algorithm: `custID movieID rating`
 - (b) quiz file: ratings to be predicted (correct answers are provided here for measuring predictive performance in RMSE): `custID movieID rating`
3. Output (4 decimal places for RMSE and *similarity*):
 - (a) Intersection algorithm
 - i. screen: number of customers with supersets, number of customers with subsets but not supersets, number of customers with neither supersets nor subsets, RMSE on the quiz file
 - ii. prediction file: `custID movieID correctRating predictedRating`
 - iii. superset file (customers with supersets): `custID supersetCustID similarity`
 - iv. subset file (customers with subsets, but not supersets): `custID subsetCustID similarity`
 - v. disjoint file (customers with neither supersets nor subsets): `custID`
 - (b) k-nearest neighbor algorithm
 - i. screen: RMSE on the quiz file
 - ii. prediction file: `custID movieID correctRating predictedRating`
 - iii. top-k file (top k neighbors): `custID neighborCustID similarity`
4. Provide a report (pdf):
 - (a) Discuss two key differences of the two algorithms. For each difference, discuss:
 - i. the difference and
 - ii. the reason you think can improve the predictive performance
 - (b) Compare the two algorithms:
 - i. RMSE performance
 - ii. time/speed
 - iii. space/memory
5. Provide `readme.txt`
 - (a) how to compile your programs
 - (b) how to run the two algorithms
 - (c) sample output of each algorithm for each input data set
6. Submit: source code, report, and `readme.txt`