

Readme for week 10 Solution

For Task1

Function “PRM_bm25_T1.py” is designed for subtasks (1) to (3).

It generates two files:

PRModel_R102.dat (the Training_set document ranking) and

PTraining_benchmark.txt (a generated relevance judgements by using query "Convicts repeat offenders").

Function “PRM_training_bm25_wk9.py” is updated week 9 Task1 solution. It uses Training_set and the generated relevance judgements “PTtraining_benchmark.txt”. The output is “PModel_w5_R102.dat” (a set of features).

Function “PRM_test_eval_bm25_wk9.py” updates week 9 Solutions for Task 2 and Task 3. It uses the features, “PModel_w5_R102.dat”, generated by “PRM_training_bm25_wk9.py” to rank documents in Test_set and save in file “PR102_test_ranks.dat”. At last, it tests the ranking results as you did in week 9.

For Task 2

Function “IRM_bm25_T2.py” is a BM25 based IR model. It ranks documents in Test_set and saves the result in “IRModel_R102.dat”.

Function “IRM_test_eval_bm25_wk10.py” evaluates the result in “IRModel_R102.dat”.

Other .py functions:

coll.py – document parsing

df.py – document frequency computing