Block0

- Our candidate set has 464671 rows.
- Our data is movie information from IMDB and TMDB, it has attributes: genres, movie_name, directors, stars, release_year, runtime.

Block1

Blocking rule based on release_year:

- Two movie names are not blank, and one movie name is a substring of the other one.
- Or at least one of the release_year is blank
- Or the two release_year is the same

After this step:

- Our candidate set has 49689 rows.
- 99.6% of the match tuples in prediction set has been kept.

Block2

Blocking rule based on the director:

- Two movie names are not blank, and one movie name is a substring of the other one.
- Or at least one of the directors from both movies is blank
- Or the directors from both movies have at least one in common

After this step:

- Our candidate set has 16085 rows.
- 98.5% of the match tuples in prediction set has been kept.

Block3

Blocking rule based on the star:

- Two movie names are not blank, and one movie name is a substring of the other one.
- Or at least one of the stars from both movies is blank
- Or the stars from both movies have at least one in common

After this step:

- Our candidate set has 14007 rows.
- 98.2% of the match tuples in prediction set has been kept.

Block4

Blocking rule based on weighted score:

Intuition: I find that the candidate set has many replicates in one of the IDs, which means it finds multiple matches from Table_B for each movie in Table_A. I decide to keep the best match in Table_B among all possible matches for each movie in Table_A.

Weighted Score: One for a match and 0 for non-match. We add them up and select the largest score.

- score_name: If one of the movie names is the substring of the other movie name
- score_director: If two director sets have a non-empty intersection
- score_star: If two star sets have a non-empty intersection
- score_year: If two release_years are the same
- score_runtime: If two runtimes are the same

After this step:

- Our candidate set has 2306 rows.
- 97.7% of the match tuples in prediction set has been kept.