Stage3-3: Estimating Accuracy

Density for Each Iteration

- we ran for two iterations:
  - candidate size before first iteration: 41,391
  - density after first iteration: 0.02
  - candidate size before second iteration: 2,426
  - density after second iteration: 0.28

Blocking Rule

- we added one blocking rule after first iteration
- main idea of the blocking rule:
  - if two tuples does not have *similar year*, they are not the same movie;
  - if two tuples has similar year but does not have *similar name*, they are not the same movie.
- description
  - check similar year:
    - if any of the years is null, return similar
    - if abs(tuple_a['year'] - tuple_b['year']) <= 1, return similar
    - else return not similar
  - check similar name:
    - split each name by space to get tokensA and tokensB
      - for any a in tokensA and b in tokensB, if existing (a,b) such that normalized_edit_distance(a,b) < 0.1, return similar
        - normalized_edit_distance = edit_distance / max(length(tokensA), length(tokensB))
      - else return not similar