BLOCKING EXPLANATION

- How did you develop the final blocker? What blocker did you start with? What problems did you see? Then how did you revise it to come up with the next blocker? In short, explain the *development process*, from the first blocker all the way to the final blocker (that you submit in the Jupyter file).

  Ans: Started with equivalence blocking based on 'Runtime' (Only 1.62% of values were missing, so it was a good candidate). Then did overlap blocking with overlap size = 2 on 'Title' column. This returned 251 matching pairs.

  Then ran the debug blocker and found that there were many false negatives. This was expected, so tried black box blocking based on runtime instead of equivalence blocking on runtime. We faced some issues here (Explained later in the issues section)

  Finally, we used overlap blocking based on 'Directors' with overlap size of 1, then overlap blocking based on 'Title' with overlap size of 1. This gave a reduced candidate set of size 4730 (which is decent). Ran debug blocker after this and went through top results. Found that there was one movie as a false negative. The other entries suggested by debug blocker were true negatives.

- If you use Magellan, then did you use the debugger? If so, where in the process? And what did you find? Was it useful, in what way? If you do not use Magellan, you can skip this question.

  Ans: We used debug blocking and found that there were many false negatives (actual matching pairs getting blocked off because of not having exactly equal runtimes across the two tables). While this was expected (as we were planning to change equivalence blocking on 'Runtime' to black box blocking based on 'Runtime'), it was certainly useful as a confirmation.

- How much time did it take for you to do the whole blocking process?

  Ans: Blocking process alone: 1 to 1.5 hours. Preparing table (refining) for blocking + getting familiar with Jupyter notebook system took about 2-3 hours

- Report the size of table A, the size of table B, the total number of tuple pairs in the Cartesian product of A and B, and the total number of tuple pairs in the table C.

  Ans: Size of table A: 7415 * 10

  Size of table B: 5008 * 10

Size of A * B: 37134320

No. of tuple pairs in table C: 4730

- Did you have to do any cleaning or additional information extraction on tables A and B?

  Ans: Had to do some data refinement (standardization) in Table B (had to make 'Runtime' column format consistent with that of table A)

- Did you run into any issues using Magellan (such as scalability?). Provide feedback on Magellan. Is there anything you want to see in Magellan (and is not there)? If you do not use Magellan, you can skip this question.

  Ans: While running the black box blocking on 'Runtime', Magellan did not give any errors but also did not give any result. The progress bar remained at 0% even after waiting for 10 minutes. Observed that there was CPU usage for python process at around 30-40% on latest gen i5 laptop. Tried this twice, then went with an alternate blocking.

  Other than that, everything was fine with Magellan.

- Any other feedback is appreciated.

  Ans: Good work!