Bryce Rothschadl
Dr. Hien Nguyen
CS767: Big Data and Data Mining
2025-03-12

# Matching Explanation

## Algorithms

To compare the string similarity of each game's title, I used the Levenshtein python package. Specifically, the ratio function using a similarity threshold of 80%. The ratio function works by calculating how many insertions and deletions it would take to transform one string into the other. This score is normalized to a range of 0 to 1, where 1 indicates an exact match. The formula for this is:

```
1 - (levenshtein_distance / (len1 + len2))
```

If the similarity score of two title strings exceeds the above threshold, I further verify the match by checking if the release dates of the two games are the same. This additional check helps reduce false positives

## Problems

I initially implemented my `is_match()` function using a similarity threshold of 90% and without checking the release dates of the two matched games. This caused a lot of games to be matched with one another when they were obviously different. For example, the games "Final Fantasy IX" (tableA_id:51) and "Final Fantasy XVI" (tableB_id:250) were matched with a Levenschtien ratio of 90.91%, despite one clearly being the 9th game in a franchise while the other is the 16th.

To fix this problem, I added the release date check mentioned in the algorithms section. This change brought the total number of matched games down from 1528 to 738, removing over 50% of the entries in Table C. Most, if not all, of these removed entries were false positives like the Final Fantasy example given above.

Another problem I found was that the similarity threshold of 90% was too strict in some cases. For example, one game's two title strings: "Ratchet & Clank (PS4)" (tableA_id:1289) and "Ratchet & Clank" (tableB_id:394) have a similarity score of 83%. I changed the similarity threshold to 80%, adding 22 more games to TableC for a total of 760 matches.

## Data

| | |
|---|---|
| Table A Size | 2,500 |
| Table B Size | 2,500 |
| Tuple Pairs in the Cartesian Product of A & B | 6,250,000 |
| Tuple Pairs in Table C | 760 |

## Cleaning and Additional Information Extraction

No additional cleaning or information extraction was required for this assignment. The data in Table A and Table B was already in a consistent format, with no missing or inconsistent values for critical attributes like Title and Release Date