Dataset Statistics

- Table A (OpenLibrary) size: 10000 books (updated)
- Table B (Google Books) size: 6455 books (Updated crawler to get 10000 (max), and it came back with this number)
- Total possible pairs (Cartesian product): 64,550,000 pairs
- Actual matches in Table C: 541 pairs

Matching Algorithms Used

1. Blocking Algorithm
    - Purpose: Makes the program go a little bit more faster when running. I noticed that a previous algorithm I used took so long I wasn't sure if it was executing correctly.
    - Method: Group records by first 3 characters of normalized title and author. This helped with speed. There maybe some false positives of course, but the majority are accurate matches (probability wise).

2. Field-Specific Matching
    - Title matching: SequenceMatcher with 0.85 threshold incase of extra characters
    - Author matching: SequenceMatcher with 0.85 threshold ^
    - ISBN matching: Direct comparison after normalization (even though there are different ISBN's (10, 13))

Weighted scoring system:

```python
field_weights = {
    'Title': 0.4,    # Primary identifier
    'Author': 0.4,   # Primary identifier
    'ISBN': 0.2      # Secondary verification
}
```

3. Similarity Calculation
    - Text normalization (lowercase, remove punctuation)
    - Fuzzy string matching using difflib.SequenceMatcher package/library
    - Overall match threshold: 0.8 (80% confidence)

Problems Encountered

1. Data Format Inconsistencies
    - Inconsistent ISBN formats (ISBN-10 vs ISBN-13)
    - Different date formats (YYYY vs. MM/DD/YYYY)
    - Mixed character encodings
    - Inconsistent field names between sources

2. Content Variations
    - Author name variations ("J.K. Rowling" vs "Rowling, J.K.")
    - Title differences ("Harry Potter and the Philosopher's Stone" vs "Sorcerer's Stone")
    - Edition-specific variations
    - Multiple authors in different orders

3. Missing Data
    - Incomplete ISBNs
    - Missing publication years
    - Absent author information
    - Empty title fields

4. Technical Challenges
    - Performance issues with $O(n^2)$ comparisons
    - False positives with similar titles
    - Duplicate entries in source data
    - Not enough data (1,000 each previously)