**Analysis**

The completed code operates by completing the matching in three passes:

1. Divide all products into groups based on their manufacturer.
2. Divide all listings into groups based on the product manufacturer list. All listings that do not match a product manufacturer are eliminated
3. Each product is compared to all listings in the corresponding manufacturer list based on the family and model.

This method is far more efficient than blindly comparing every product with every listing. Listings that don’t correspond to a product manufacturer are eliminated from further testing. Additionally, the last two comparisons are only done for matching manufacturers, minimizing unnecessary comparisons.

The table below shows the worst-case and average number of comparisons for each method:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Pass One** | **Pass Two** | **Pass Three** | **Three Pass Total** | **Blind Comparison** |
| **Worst-case** |  |  |  |  |  |
| **Average** |  |  |  |  |  |

Where:  
 is the number of products  
 is the number of listings  
 is the number of unique product manufacturers  
 is the subset of listings remaining after the second pass

To validate, the worst-case and average values were computed for the test data, then the number of comparisons was measured programmatically to compare.

The parameters of the test data are:  
 = 743  
 = 20196  
 = 21  
 = 16399

The table below shows the calculated and measured values:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Pass One** | **Pass Two** | **Pass Three** | **Three Pass Total** | **Blind Comparison** |
| **Worst-case** | 15,603 | 424,116 | 30,011,256 | 30,450,975 | 45,016,884 |
| **Average** | 7,802 | 212,058 | 2,858,214 | 3,780,074 | 30,011,256 |
| **Measured** | 4,008 | 182,242 | 2,296,238 | 2,482,488 | NA |

The measured values for this case were better than the calculated average for all passes. This can be attributed to an approximately equal distribution of products under each respective manufacturer. The results will approach the worst case condition when the majority of products fall under one manufacturer. When there is effectively one manufacturer, this method is functionally comparable to blind comparison. The only exception to this being that organizing the products and listings by the product manufacturers is more efficient than comparing every listing to every product based on their manufacturer. This is why the worst-case for the three pass method is still faster than the blind comparison’s worst-case.

To verify the accuracy of the code, two methods were used:

* Manually review the matches for any discrepancies
* Programmatically check for listings that match with multiple products

By manually reviewing the results, an early implementation of the code was found to be returning false positives for matches. By simply searching for a product’s model within a listing, model names that were a substring of other model names were being incorrectly matched. For example, the following product and listing would be deemed a match:

{"product\_name":"Sony\_Cyber-shot\_DSC-TX1","manufacturer":"Sony","model":"DSC-TX1","family":"Cyber-shot","announced-date":"2009-08-05T20:00:00.000-04:00"}

{"title":"Sony - CyberShot - DSC-TX1H - Appareil photo numérique - 10,2 Mpix - Zoom optique 4x - Ecran LCD 2,7\" - 11 Mo - Gris","manufacturer":"Sony","currency":"EUR","price":"219.99"}

This is due to the fact that the product model “DSC-TX1” is found as a substring of “DSC-TX1H”. To resolve this issue, spaces were prepended and appended to the product model before comparison to the listing’s title string.

The results of checking for duplicate matches identified another issue with the code. Listings were being matched to multiple products in the event that a product’s model name was extended for a newer model. For example, “Pentax WG-1” and “Pentax WG-1 GPS” are two distinct products, yet would both match a listing titled “PENTAX Optio WG-1 GPS – gris”. Nine cases of duplicate matches were found in the sample data.

In the spirit of maintaining accuracy over maximizing the number of matches, both matches were discarded in the event of a duplicate match. While the sample data could be used to generate rules to determine the correct match, the inconsistency in naming conventions between companies could potentially result in the wrong match being selected as correct.

In total, 4,263 matches were found, with the formatted results containing 323 entries. With nine duplicate cases identified and both matches being discarded, nine potentially correct matches were discarded resulting in a known loss of 0.211% of valid matches. For future work, these lost matches can be addressed. Additionally, analysis can be done to determine the best order in which to compare model and family names in order to minimize the number of comparisons.