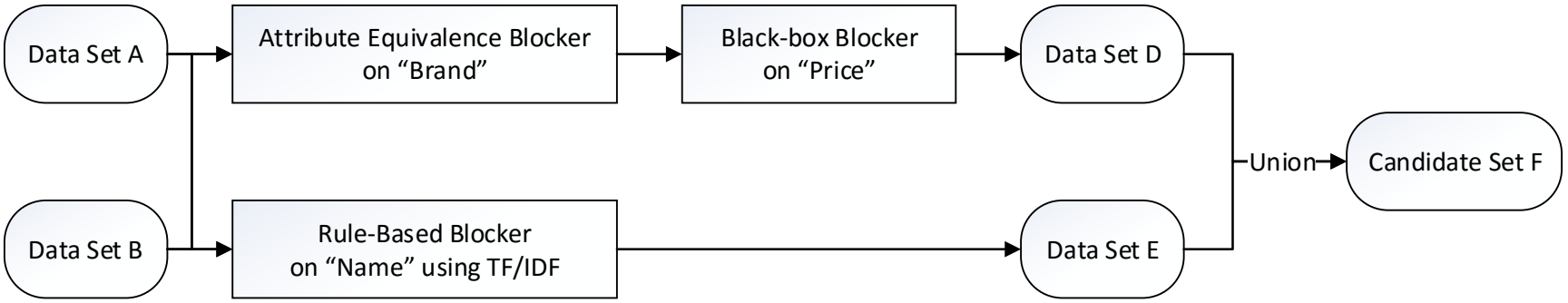
CS 784 – Project Stage 2

Qing Li

Fan Ding

1. Structure of Blockers



The final blocker structure is as above.

We decided to use an "Attribute Equivalence Blocker" on "Brand" first. However, since there are lots of products are laptop accessories, their brand values are all labeled as "Other". To further reduce the size of candidate set, we apply a black-box blocker on "Price" to limit the match range for those products under identical brand.

Since it is using an extra information extraction method to generate the brand value for each product, to avoid mislabeled issue, we also applied a rule-based blocker on name for all tuples.

We also try to implement a rule-based blocker on “Feature”. However, since the length of some features is quite long, to run the blocker is very expensive (estimated to be more than 3 hours). Moreover, there are quite number of null features. Therefore, we drop this blocker in the end.

Finally, we obtain our candidate set by union two output sets as above figure shows.

2. **Debugger**

We use the debugger for rule-based blocker. At first, we used 0.7 as threshold and there were quite few remaining tuples. Then we used the debugger and reset threshold to 0.3 and returned more tuples that might be matched. The debugger would be more useful if it can also show some statistical analysis for output results, like the distribution of similarity.

3. **Running Time**

For AttrEquivalenceBlocker():

29.527 seconds

For BlackBoxBlocker():

48.027 seconds

For RuleBasedBlocker():

3544.701 seconds

Total time elapsed is 3622.255 seconds, nearly one hour.

4. **Size of Results**

Size of Table A (Amazon.csv): 4,559 tuples

Size of Table B (BestBuy.csv): 5,001 tuples

Catersian product of A and B: 22,799,559 tuples

Size of Table C: 823,832 tuples

5. **Pre-processing Step**

Extract “Laptop Brand” from Product Name

Based on “List of Laptop Brands and Manufacturers” in Wikipedia (https://en.wikipedia.org/wiki/List_of_laptop_brands_and_manufacturers), we create a dictionary of laptop brands and corresponding product lines. Based on this dictionary, a new attribute, called “Brand”, is added for each tuple, and applied this attribute into AttrEquivalenceBlocker().

6. **Issues with Magellan**

There is a setup issue in both Windows 7 and Windows 10.

After finishing installing Magellan, when import Magellan in Ipython notebook, it will complain about,

```
ImportError: Building module magellan.cython.test_functions failed: ['Dist
utilsPlatformError: Unable to find vcvarsall.bat\n']
```

It seems that “Microsoft Visual C++ Tool” is undefined in some Windows machine. We think it is better to mention this dependency in the first part of User Manual.