Type of entity:
We have chosen to extract data about smart devices (cell phones, tablets, smart watches etc).

Description of the two tables:
Table A was obtained from GSMArena - https://www.gsmarena.com and
table B from Flipkart - https://www.flipkart.com

Number of tuples per table:
Number of Tuples in Table1(GSMArena):3273
Number of tuples in Table 2(Flipkart):3507

Choice of blocker:
We have used two overlap blockers in series on Name and Resolution.
Number of tuple pairs in the candidate set obtained after the blocking step: 830

Number of tuple pairs in the sample G that you have labeled: 300

Cross Validation Results:

ML algo	***Avg Precision***	***Avg Recall***	***Avg F1***
Decision Tree	***80***	***81***	***0.742***
Random Forest	***85***	***79***	***0.743***
SVM	***70***	***15***	***0.2558***
Naive Bayes	***63***	***81***	***0.704***
Logistic Regression	***83.9***	***81***	***0.745***
Linear Regression	***74***	***67***	***0.643***

Matcher chosen after that cross validation:Logistic Regression

Debugging: In our debugging step we identified some matches that were being missed due to extraneous information present in the name. For example some phone names had colour, internal memory etc as part of the name. We cleaned our data to have names represent purely just the names of the phone.
We now perform cross validation again on the clean data.

Cross Validation Report after debugging

ML algo	***Precision***	***Recall***	***F1***
Decision Tree	***89***	***76***	***0.790***
Random Forest	***97***	***85***	***0.847***
SVM	***40***	***11***	***0.123***
Naive Bayes	***79***	***89***	***0.836***
Logistic Regression	***87***	***89***	***0.863***
Linear Regression	***65***	***68***	***0.746***

Final Matcher chosen: Random Forest(Because Logistic regression does not meet required Precision)

FINAL: PREDICTING PRECISION, RECALL,F1 on TEST SET(SET J)

ML algo	***Precision***	***Recall***	***F1***
Decision Tree	***96***	***83***	***0.896***
Random Forest	***92***	***77***	***0.842***
SVM	***100***	***51***	***0.680***
Naive Bayes	***93***	***93***	***0.935***
Logistic Regression	***88***	***77***	***0.827***
Linear Regression	***87***	***67***	***0.763***

Best Matcher Selected: Random Forest

Precision:92
Recall:77
F1:0.842

Approximate time estimate:
to do the blocking: 2 hours
to label the data: 7 hours
to find the best matcher: 3 hours

Comments on Magellan:

1) Labeller display can be made in such a way so as to avoid scrolling(Maybe vertical side by side display)

2) Make names of rules in rule based blocker more intuitive. (maybe we can use attribute names instead of representing rules in a numerical sequence like rule1,rule2 etc)

3) Once a CSV file is opened in magellan tool, we cannot open the same file again in excel because excel detects it to be an SYLK file.