1. We are using Silicon Valley dataset. We are matching songs and tracks tables. We changed the schema of tracks.csv to match the schema of songs.csv.
    - Number of tuples in songs table = 961594
    - Number of tuples in tracks table = 734484

2. We did downsampling of songs and tracks tables. We did downsampling, blocking and sampling of golden data in 5 iterations.
In each iteration, we took 3000 samples from songs table and approx 18000 samples from tracks table.
For blocking, we are using word level overlap blocker on artist and title columns with the condition of a minimum overlap of one word. To reduce the candidate set even more, we used edit distance of almost 15 characters on both title and artist name.

- Number of tuples pairs in the candidate set after 5 iterations of blocking = 17913.

3. Number of tuples pairs in the labeled sample G (sampled.csv) = 357

4. **F-1**

Name	Num folds	Mean score
DecisionTree	5	0.933930472
RF	5	0.959148357
SVM	5	0.925055917
LinReg	5	0.955003202
LogReg	5	0.928177389
NaiveBayes	5	0.938168912

Precision:

Name	Num folds	Mean score
DecisionTree	5	0.940535117
RF	5	0.968407796
SVM	5	0.892734674
LinReg	5	0.961790317
LogReg	5	0.925045607
NaiveBayes	5	0.968592593

Recall:

Name	Num folds	Mean score
DecisionTree	5	0.929761905
RF	5	0.951190476
SVM	5	0.960714286
LinReg	5	0.951190476
LogReg	5	0.93452381
NaiveBayes	5	0.911904762

5. We selected Random Forest matcher after cross validation since it gave the highest F1.

6. We debugged Random Forest once. We saw 1 false positive out of 42 positive predictions and 1 false negative out of 33 negative predictions.

- False Negative example:- "SOMETHING FOR YOU" and "Something For You" - were detected as different.
- False Positive example:- "Le vilain pays" and "Mr. le PrÃ©siden" were detected as same

We did not add additional features because we did not want to overfit the training data.

7. **Precision/Recall/F-1 on set J**

Random Forest
Precision : 92.86% (52/56)
Recall : 92.86% (52/56)
F1 : 92.86%

Decision Tree
Precision : 87.93% (51/58)
Recall : 91.07% (51/56)
F1 : 89.47%

SVM

Precision : 94.64% (53/56)
Recall : 94.64% (53/56)
F1 : 94.64%

Linear Regression

Precision : 98.11% (52/53)
Recall : 92.86% (52/56)
F1 : 95.41%

Logistic Regression

Precision : 94.55% (52/55)
Recall : 92.86% (52/56)
F1 : 93.69%

Naive Bayes

Precision : 96.15% (50/52)
Recall : 89.29% (50/56)
F1 : 92.59%

8. For the selected final best matcher (**Random Forest**), Precision/Recall/F-1 on set J :

Precision : 92.86% (52/56)
Recall : 92.86% (52/56)
F1 : 92.86%

9. **Approximate time estimates:**

A. Blocking - System takes approx. 30 seconds to do the blocking step. It took up about 2 days to figure out the correct blocking.
B. label the data - 10 mins
C. Find the best matcher - It took up approx. 2-3 hours to figure this out.

10. We did not want to overfit the training data which is why we could not achieve higher recall. In future, we can try converting the entire data into lowercase before the blocking step. That might improve the recall a bit. We can also use a larger golden data sample.

11. **Bonus points:**

- Show Progress=False, then also shows the progress - MAC OS X EI Capitan
- DownSampling = The package allows the user to specify a parameter (k) for each tuple in Table A, which specifies the number of candidate matches to be selected in Table B. Half of the k tuples are potential matches, similar to the tuple and half are randomly chosen from the Table B. There is no tuning parameter that can change this ratio while downsampling.
- Debugging step in blocker can be done only on one column and not multiple columns
- MAC OS X EI Capitan hangs multiple times while running Jupyter Notebooks