1.1

知识：

pairs completeness (recall)

pairs quality (precision)

0: rec_id

1: first_name

2: middle_name

 # 3: last_name

 # 4: gender

5: current_age

6: birth_date

7: street_address

 # 8: suburb

9: postcode

10: state

11: phone

12: email

Attributes selected as blocking keys should: – Not change over time (i.e. be constant) – Be accurate (no errors or variations in them) – Be complete (no missing values) – Have a frequency distribution close to uniform – Have a reasonable number of unique different values.

a. Not change over time (constant)一致性 b. Be accurate (no errors or variations in them)不能有明显的错误或者改变 c. Be complete (no missing values)不能有空缺值 d. Have a frequency distribution close to uniform 频率分布接近均匀 e. Have a reasonable number of unique different values 有合理数量的独特的不同值 （这列有 50 个 tom 不选）

Answers:

1. Using blocking techniques can reduce the number of record pairs that are compared in detail as much as possible, by removing pairs that unlikely correspond to true matches to reduce the comparison space.

2. My choice of method is phonetic (Soundex) and blocking keys are “last name”, “gender”, “suburb”.

3. “middle name”, “phone”, “email”, “birth date” and “stress address” have missing values so that they cannot be blocking keys. “postcode” has incorrect values. “first name”, “state” and “current

age” do not have a reasonable number of unique different values, e.g. maybe they are many people of the same age and first name in the databases. Therefore, I used “last name”, “gender”, “suburb” as my blocking keys.

1.2

知识：

linkage performance (efficiency 效率):

reduction ratio, pairs completeness, and pairs quality (from question)

linkage quality (effectiveness 效力):

pairs completeness, pairs quality, precision, recall, etc(from slides)

Recall and Precision:

前者需要返回的数据多，错杀一百，不放过一个。(PC)

后者要准确，只杀一个，也不能杀错人。(PQ)

Reduction ratio: Measures by how much a blocking technique can reduce the comparison space.

Pairs completeness: Measures how many true matches 'pass' through a blocking.

Pairs quality: Measures how many candidate record pairs generated by blocking are true matches.

F-measure: Calculates the harmonic mean of precision and recall.

课本原话：

Achieving high linkage quality is a main goal of most record linkage projects / applications.

It should be noted that there is a trade-off between precision and recall. Depending upon the data matching or deduplication situation, it might be more important to achieve matching results with high precision but accept a lower recall, while in other data matching situations having a low precision is acceptable but a high recall is required.

For example, for a crime investigation where certain suspect individuals need to be matched with a large database of people, a high recall is desired to make sure that it is likely that the individuals one is looking for are included in the matched record pairs, even if there is a larger number of matches that need to be investigated. On the other hand, high precision is required in many public health studies where each match would correspond, for example, to a patient with certain medical characteristics who needs to be included into a cohort study. In this situation one wants to be sure to only include patients into the set of matched record pairs who do have the medical condition one is interested in. (Page168)

Answers:

Yes, there is a trade-off between performance and quality of the final record linkage results.

Please notes records of blocking evaluation are attached. Pairs completeness and quality are almost same between simple blocking and SLK blocking, but in phonetic (Soundex), pairs completeness(recall) is much higher than them, but pairs quality(precision) is little lower than them. Consider the database, high precision is not required in this situation, however, both of blocking techniques have same the reduction ratio. And high recall is necessary to keep as much as people's records as possible, although it would match more record pairs.

```
Blocking evaluation(Simple):
  Reduction ratio:    1.000
  Pairs completeness: 0.499
  Pairs quality:      0.990


Blocking evaluation(Phonetic):
  Reduction ratio:    1.000
  Pairs completeness: 0.680
  Pairs quality:      0.929


Blocking evaluation(SLK):
  Reduction ratio:    1.000
  Pairs completeness: 0.499
  Pairs quality:      1.000
```

1.3

题目要求：

Trade-off 会发生改变由于 **Datasets with lower-level or high-level data quality 和 Characteristic of data quality:**

How:

Why:

Characteristic of data quality（包括）：

Missing values

Inconsistent values

Invalid values

课本原话：

For databases that contain data of low quality, employing an indexing technique that inserts records into several blocks or clusters will be of advantage compared to employing a technique that inserts each record into one block only. On the other hand, if the data to be matched or deduplicated are of good quality, then using the traditional blocking technique (that inserts each record into one block only) might be appropriate. (Page99)

Answers:

Yes, I think the trade-off would change in different situations.

In high data quality:

Using blocking techniques to remove more record pairs, at the same time, more true matching pair are removed. Therefore, using traditional blocking technique to inserts data to one block is more useful for database of high quality.

In low data quality:

In real world, data are often dirty like missing values, data changing over time etc. For example, missing values mean BKV cannot be generated. Therefore, for databases have low quality of data, using blocking techniques that insert data into several blocks.

2.1

知识：

Wk8: lec19-20

Comparison techniques:

1. **Q-gram based** (Jaccard(jaccard_comp) and Dice coefficient(dice_comp))

2. **Edit and bag distances** (edit_dist_sim_comp, bag_dist_sim_comp) (the bag distance similarity, the edit distance similarity)

3. **Jaro-Winkler** (jaro_comp, jaro_winkler_comp) (the Jaro comparison function, the Jaro-Winkler modifications)

Answers:

1.Comparison techniques affect the precision of linkage results, and synchronous affect the f-measure.

```
Linkage evaluation:
(Using the Jaro-Winkler and the edit distance)
  Precision:    0.974
  Recall:       0.680
  F-measure:    0.801

Linkage evaluation:
```

```
(Only compare the two given attribute values exactly)
  Precision:   1.000
  Recall:      0.680
  F-measure:   0.810
```

2. Comparison techniques and reasons:

Last name (the Jaro-Winkler modifications): Jaro-Winker is specifically to compare personal name string, taking various heuristics into account that are based on extensive practical experiences of name matching.

Suburb (the edit distance similarity): it widely uses to count how many basic edit operations are needed to convert one string into another.

Gender (Compare the two given attribute values exactly): only have two situations, male and female, therefore, compare the two given attribute values exactly, return 1 if they are the same (but not both empty!) and 0 otherwise.

2.2

名字：

1. Exact
2. Threshold
3. minThreshold
4. weightedSimilarity
5. supervisedML

分类的方法：

C1

```
Linkage evaluation(exact):
  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.492
  F-measure:   0.660
```

C2

```
Linkage evaluation(threshold):
  Accuracy:    1.000
  Precision:   0.929
  Recall:      0.680
  F-measure:   0.786
msim_threshold = 0.01


  Accuracy:    1.000
  Precision:   0.969
  Recall:      0.680
  F-measure:   0.799
```

```
sim_threshold = 0.99

Linkage evaluation:
  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.679
  F-measure:   0.809
sim_threshold = 0.65
```

C3

```
Linkage evaluation(min threshold):
  Accuracy:    1.000
  Precision:   0.998
  Recall:      0.647
  F-measure:   0.785
min_sim_threshold = 0.01


  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.492
  F-measure:   0.660
min_sim_threshold = 0.99
```

C4

```
Linkage evaluation(weighted similarity):
  Accuracy:    1.000
  Precision:   0.929
  Recall:      0.680
  F-measure:   0.786
sim_threshold = 0.01

  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.492
  F-measure:   0.660
sim_threshold = 0.99

  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.679
  F-measure:   0.809
sim_threshold = 0.65
```

C5

```
Linkage evaluation(supervised ML):
  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.680
  F-measure:   0.810
```

Answers:

1.Differnce classification techniques using different parameter settings will affect precision, recall and f-measure. Please notes records of linkage evaluation are attached.

2.I used the supervised machine learning technique (decision tree) because it has the best record of f-measure (0.810), which is combining precision and recall.

2.3

题目：

for suitable **linkage quality measures**, describe how the final record linkage quality **changes** with the choice of **different parameters and techniques**。

Is the record linkage quality **particularly sensitive** to certain parameters, or choice of comparison or classification techniques?

Answers:

1.In threshold and weighted similarity classification techniques have the same best parameter of similarity threshold is 0.65 which has the highest f-measure (0.809). However, in min threshold classification technique, the best parameter of minimum similarity threshold is 0.01 which has the highest f-measure (0.785).

2.Yes, the recall is sensitive when use different parameters and techniques. In weighted similarity classification techniques, the recall is 0.492 when similarity threshold is 0.99, however, the recall is 0.680 when similarity threshold is 0.01. As we known, recall is true positive rate (|TP| / |TP|+|FN|), the reason of recall is sensitive because it is the proportion of matches that have been classified correctly. On the other word, in binary classification, recall is called sensitivity.

2.4

题目：

哪些评估方法没有用，为什么

Answers:

1. Reduction ratio

Reduction ratio is calculated as 1 – (The number of candidate record pairs) / (TP + FP + FN+ TN).
The number of candidate record pairs are much smaller than the total number of comparisons betw
een all record pairs, therefore, reduction ratio always is 1.

2. Accuracy
Accuracy is calculated as (TP + TN) / (TP + FP + FN + TN).
The sum of FP and FN is much smaller than the sum of TP and FN, therefore, accuracy always is 1.

2.5

题目：

Provide the **numerical linkage evaluation results（没有优化的）**

```
Linkage evaluation:
  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.680
  F-measure:   0.810

Linkage evaluation:
  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.492
  F-measure:   0.660
```

3.1

Answers:

1.My best linkage quality result

Blocking step: phonetic (Soundex)

Classification step: machine learning technique (decision tree)

```
Blocking evaluation:
  Reduction ratio:    1.000
  Pairs completeness: 0.680
  Pairs quality:      0.929

Linkage evaluation:
  Accuracy:    1.000
  Precision:   1.000
  Recall:      0.680
  F-measure:   0.810
```

2.Reasons for why techniques worked well for my data set pair

It is a trade off with pairs completeness and pairs quality, my choices keep a good balance between efficiency and effectiveness, it compared 7323 records pairs and classified 68050 records which only have 4 wrong in tot al. On the other hand, my confusion matrix is TP=6804, FP=1, FN=3196, TN=399989999, and the f-measure is 0.810.

3.2

Answers:

About measure matching complexity:

The results are good for recall and f-measure. However, they are not good for accuracy.

About measure matching complexity:

The results are good for pairs completeness and pairs quality. However, they are not good for reduction ratio.

Because TP=6804, FP=1, FN=3196, TN=399989999, accuracy is calculated as (TP + TN) / (TP + FP + FN + TN) and reduction ratio is calculated as 1 – (The number of candidate record pairs) / (TP + FP + FN+ TN). The most important thing is the sum of FN and FP is much smaller than the sum of TP and TN. B y the way, we do not involve Clerical Review in this assignment.

4.1

Answers:

1. birth date: it has “q986”, “1i89” and “|999” etc.
2. postcode: it has “082z” and “435o” etc.

I think the data sets for this assignment are much dirty than data sets we use in lab 3 to 7.

4.2

Answers:

My check list for data quality:

1. Missing data (null values in different types, e.g., in data of birth date, “#######”)
2. Out-of-date data (people move house so that they change address, postcode, state)
3. Data variations in different sources (e.g., in data of gender, 0 means male, 1 means female)
4. Error (wrong spelling or incorrect data, e.g., 0 is not acceptable in data of weight, 200 is not acceptabl e in data of age)
5. Misinterpretation of data (two or more datasets use different attributes for the same data)

My methods:

For basic data information, like null values and data type, I use Rattle. For some complex inspections, I write python programs in jupyter notebook.