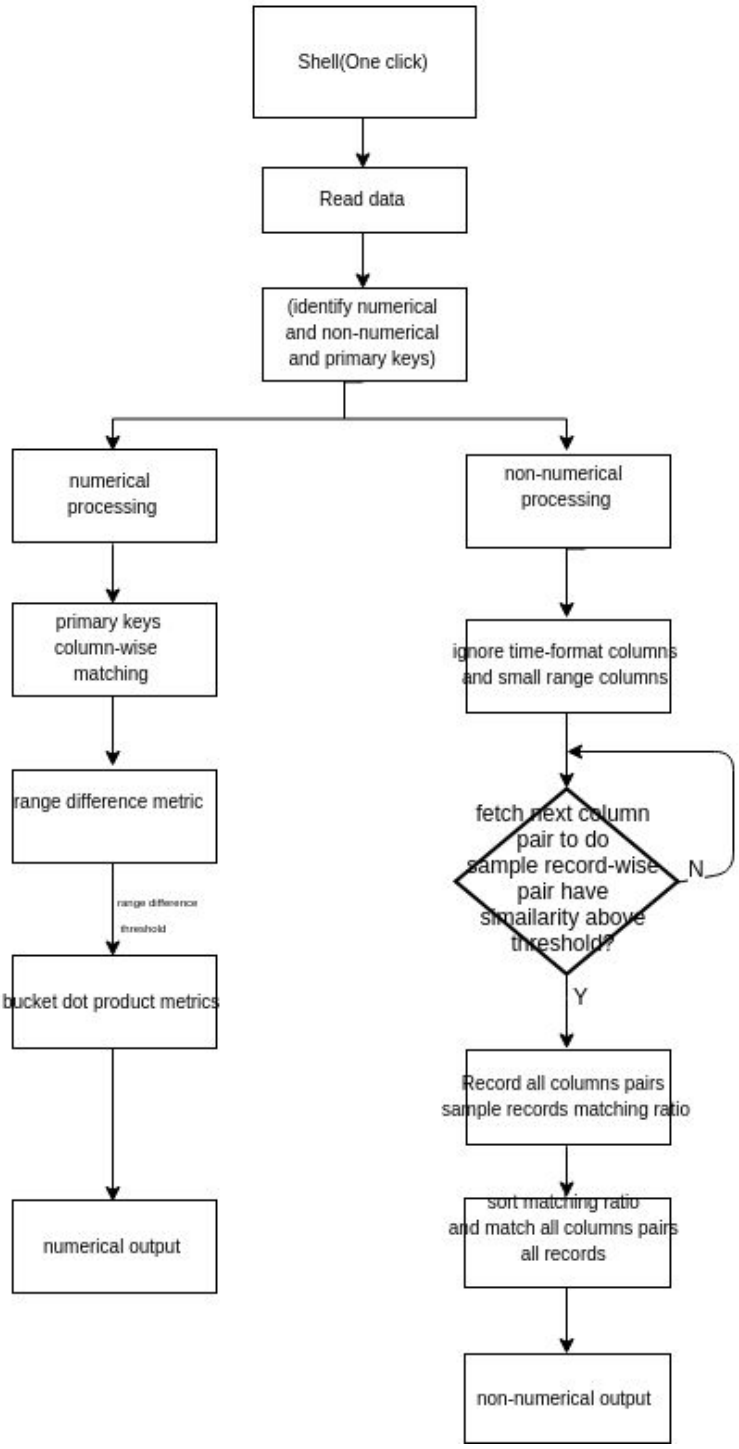
Code sorting out and cleaning documents

1. Code flowchart

Flowchart pic is in GraphMatching/SpydeWks/Codes/sys_docs/Flowchart/overviewFlowchart.png



2. Codes directory

Codes are in GraphMatching/SpydeWks/Codes

Executing code: Shell script (call python codes) databaseMatching.sh

Run ./**databaseMatching.sh script to run python codes**

```
#!/bin/bash

INPUTDATADIR=/home/fubao/Fubao/CiscoWish/data/test/
RANGEDIFFTHRESHOLD=1.5
BUCKETSIZENUM=200000
OUTPUTNUMERICALRES=/home/fubao/Fubao/CiscoWish/CreateGraph/GraphMatching/SpydeWks/Codes/output/numericalOutput/allNumericalFinalResult.tsv

NONNUMPREFIXLENGTH=2
NONSAMPLERECORDNUM=2000
RECORDPAIRSIMITHRESHOLD=0.5
OUTPUTNONNUMDIR=/home/fubao/Fubao/CiscoWish/CreateGraph/GraphMatching/SpydeWks/Codes/output/nonNumericalOutput/
OUTPUTNONNUMRATIOFILE=/home/fubao/Fubao/CiscoWish/CreateGraph/GraphMatching/SpydeWks/Codes/output/nonNumericalOutput/nonNumericalFinalMatch
INTEROUTFLAG=True

echo "Start Database Matching..."
python3 mainEntry.py -i $INPUTDATADIR -rdt $RANGEDIFFTHRESHOLD -bs $BUCKETSIZENUM -oNum $OUTPUTNUMERICALRES  -pl $NONNUMPREFIXLENGTH -spNum
echo "End"
```

There are **10 input parameters.**

-i: input database directory path

-rdt: Range difference threshold

-bs: Bucket dot product bucket number

-oNum: numerical output file name

-pl: prefixLength: non-numerical prefix length

-spNum: non-numerical sample record num

--recsimt: non-numerical record similarity threshold

-oNonDir: non-numerical output dir

-oNonRt: non-numerical output matchingRatio file name

-interOFlg: indicate output intermediate files flag

The parameter values can be changed in that shell script

Two Sample database running result is shown below:

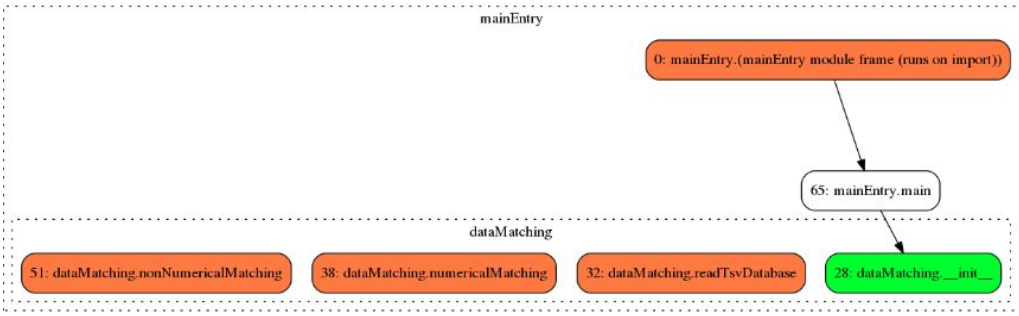
```
fubao@t440p:~/Fubao/CiscoWish/CreateGraph/GraphMatching/SpydeWks/Codes$ ./databaseMatching.sh
Start Database Matching...
Input file: /home/fubao/Fubao/CiscoWish/data/test/
database table name:  tss_service_levels_d_v
database table name:  tss_workgroups_d
------------------------------
begin numerical matching...
len allNumericalValuesMap  19
allNumericalPairsRangeDifferenceScoreMap len:  39
tbFieldAllNonNumericalValuesMap len  36
------------------------------
begin non-numerical matching...
pairsTupleTobeMatched len  315
total runtime:  23.30398154258728
End
```

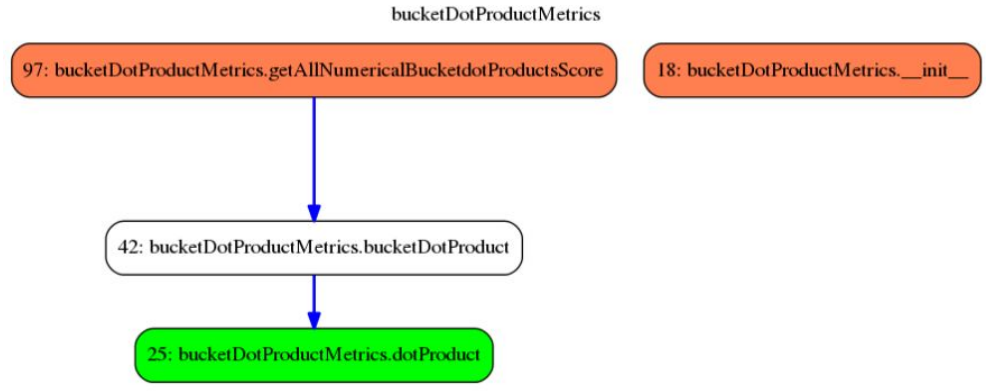
3. Code structures:

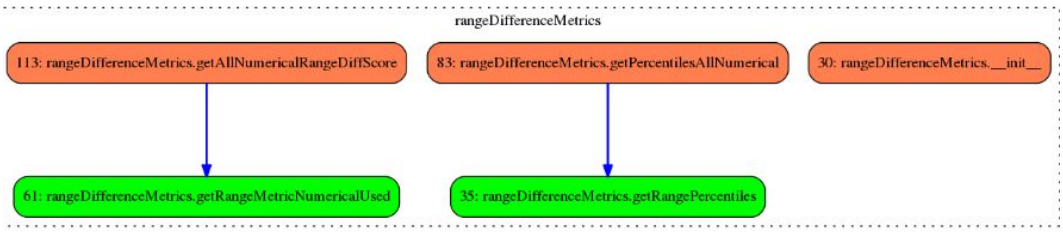
all the pics are in GraphMatching/SpydeWks/Codes/sys_docs/Flowchart/

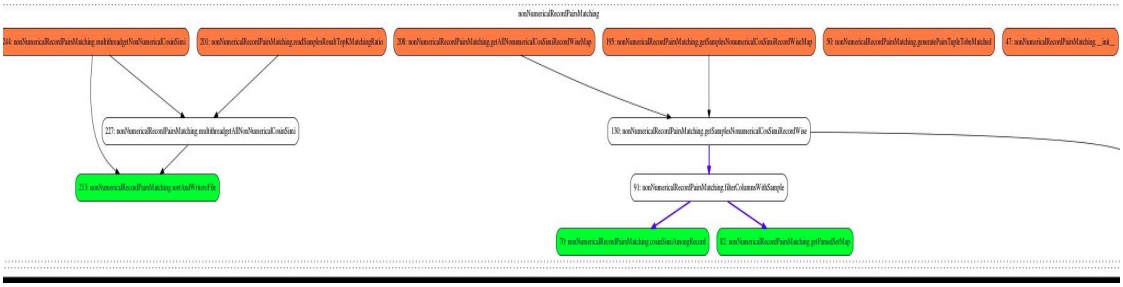
If you can't see the pics below

MainEntry is the main program









4. Code specifications and API documentation

Api documents are in sys_docs folder

For example : “sys_docs/mainEntry” shows the main function entry’s class API

1.3 Class dataMatching

object
└ **mainEntry.dataMatching**

1.3.1 Methods

__init__(*self*)

x.__init__(...) initializes x; see help(type(x)) for signature

Overrides: object.__init__ extit(inherited documentation)

nonNumericalMatching(*self*, *tbFieldAllNonNumericalValuesMap*, *threadNum*, *prefixLength*, *sampleRecordsNum*, *recordPrSimiThreshold*, *finalNonNumericalOutputDir*, *outFileNonNumericalRatioScore*)

numericalMatching(*self*, *rangeDiffThd*, *inputBucketSizeNum*, *primaryKeysSet*, *allNumericalValuesMap*, *outRangeFileFlag*, *finalNumericalOutputFile*)

readTsvDatabase(*self*, *dataInputDir*, *nonNumericalColumnSmallRange*, *InterMediateFileFlag*)

The folder " sys_docs/rangeDifferenceMetrics "shows the rangeDifferenceMetrics class API

1.2 Class rangeDifferenceMetrics

object
└ **rangeDifferenceMetrics.rangeDifferenceMetrics**

1.2.1 Methods

__init__(*self*)

x.__init__(...) initializes x; see help(type(x)) for signature

Overrides: object.__init__ extit(inherited documentation)

getRangePercentiles(*self*, *valsSet*, *percentA1*, *percentA2*, *percentA3*, *percentA4*, *percentA5*, *percentA6*, *percentA7*, *percentA8*, *percentA9*)

getRangeMetricNumericalUsed(*self*, *x2*, *y2*, *x3*, *y3*, *x8*, *y8*, *x9*, *y9*)

getPercentilesAllNumerical(*self*, *allNumericalValuesMap*, *outRangeFileFlag*)

getAllNumericalRangeDiffScore(*self*, *primaryKeysSet*, *allNumericalfieldRangeMap*, *outRangeFileFlag*)

More numerical and non-numerical code html and pdf documentation are in the sys_docs folders...