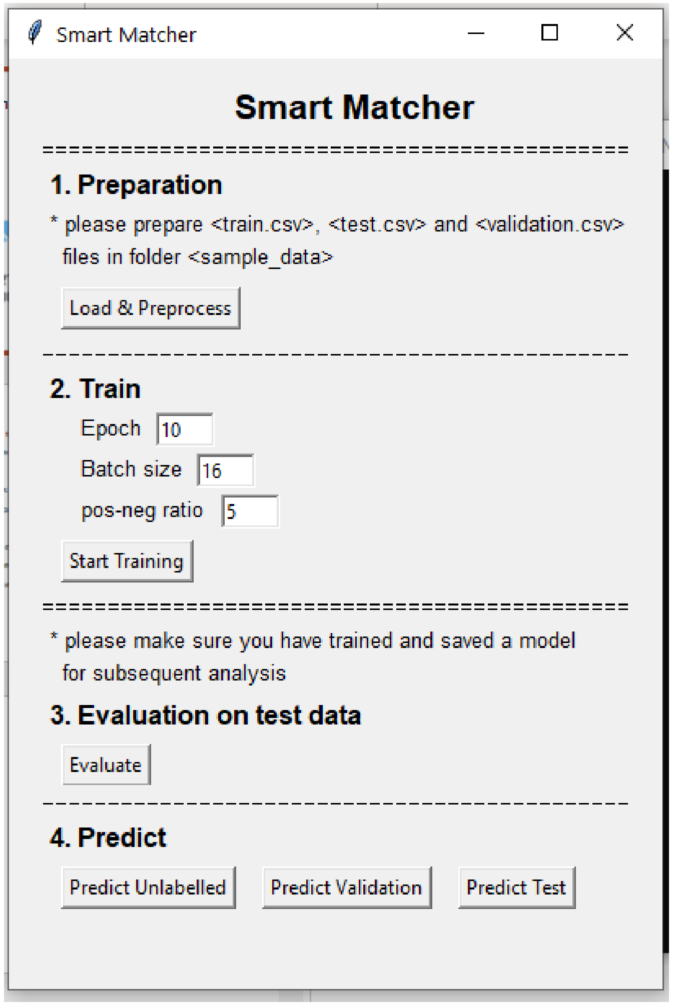
Deep learning-based Entity Matching (Quick Start)



Smart Matcher

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

* please make sure you have trained and saved a model for subsequent analysis

3. Evaluation on test data

4. Predict

Outline

- Introduction

- Quick start

- Appendix

Introduction

- In this software, *Smart Matcher*, we have made the flow extremely simple that everyone can build a deep learning model to match entities.

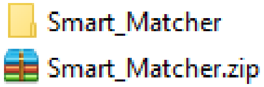
- This software is equipped with the state-of-the-art machine learning algorithm for deep learning-based entity matching.

- This file is a quick introduction to the entity matching software, *Smart Matcher*. This trial software has been simplified so that users can run it smoothly on standard Windows computers.
  Noted that, if better facilities (larger RAM, GPUs) are available, this software can be made more complicated and robust (e.g., by incorporating pre-trained character-level embeddings and training using GPUs to process long text and match entities with many attributes).

- This software has been tested on a collection of datasets which include most publicly available datasets for entity matching.

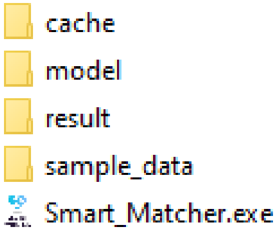
Quick start - preparation

Download the software and unzip it into any directory



Smart_Matcher
Smart_Matcher.zip

Go to folder <Smart_Matcher>, we can see the following files/folders



cache
model
result
sample_data
Smart_Matcher.exe

Quick start - preparation

What are these folders/files about?

cache:
A pre-trained word-level embedding model of relatively small size has been saved in this folder.

model:
A model trained by the user will be saved in this folder.

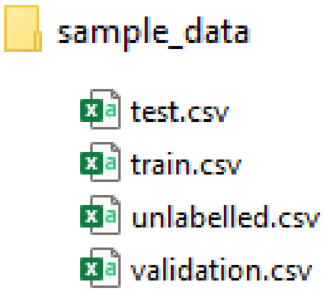
result:
All results will be saved in this folder. More details later.

sample_data:
Sample datasets have been prepared in this folder. More details later.

Smart_Matcher.exe:
The software.

Quick start - sample data

Folder <sample data> contains sample datasets for training, evaluation and prediction. (*csv, utf-8 encoding*)



sample_data

- test.csv
- train.csv
- unlabelled.csv
- validation.csv

The sample data is about matching research articles.

Each article (entity) has attributes: id, title, authors, venue and year.

Quick start - sample data

test.csv, *train.csv* and *validation.csv* have the same formats as below

Column	**Example column**	**Example Entry**
id	id	5428
label	label	**0**
left_A	left_title	transaction timestamping in (temporal) databases
left_B	left_authors	christian s. jensen , david b. lomet
left_C	left_venue	vldb
…	…	…
right_A	right_title	time-parameterized queries in spatio-temporal databases
right_B	right_authors	yufei tao , dimitris papadias
right_C	right_venue	international conference on management of data
…	…	…

Quick start - sample data

test.csv, *train.csv* and *validation.csv* have the same formats as below

Column	**Example column**	**Example Entry**
id	id	5428
label	label	**1**
left_A	left_title	dynamic maintenance of data distribution for selectivity estimation
left_B	left_authors	kyu-young whang , gio wiederhold , sang-wook kim
left_C	left_venue	vldb j.
…	…	…
right_A	right_title	dynamic maintenance of data distribution for selectivity estimation
right_B	right_authors	kyu young whang , sang wook kim , gio wiederhold
right_C	right_venue	the vldb journal -- the international journal on very large data bases
…	…	…

Quick start - sample data

unlabelled.csv has the same columns as *test.csv*, *train.csv* and *validation.csv* except for '*label*', because this is what the trained model will predict later.

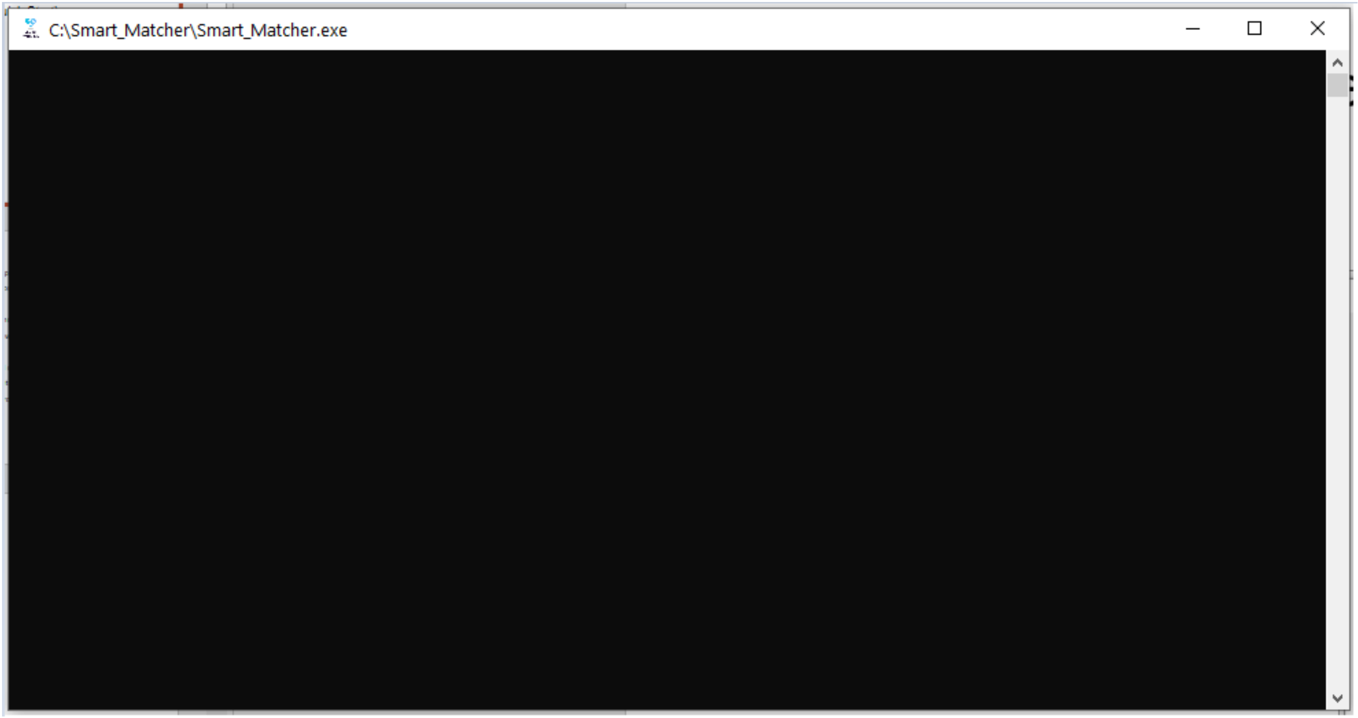
Next let's start using *Smart Matcher*.

Quick start - data preprocessing

Double click *Smart_Matcher.exe* to launch the software

SMART MATCHER
Smart_Matcher.exe

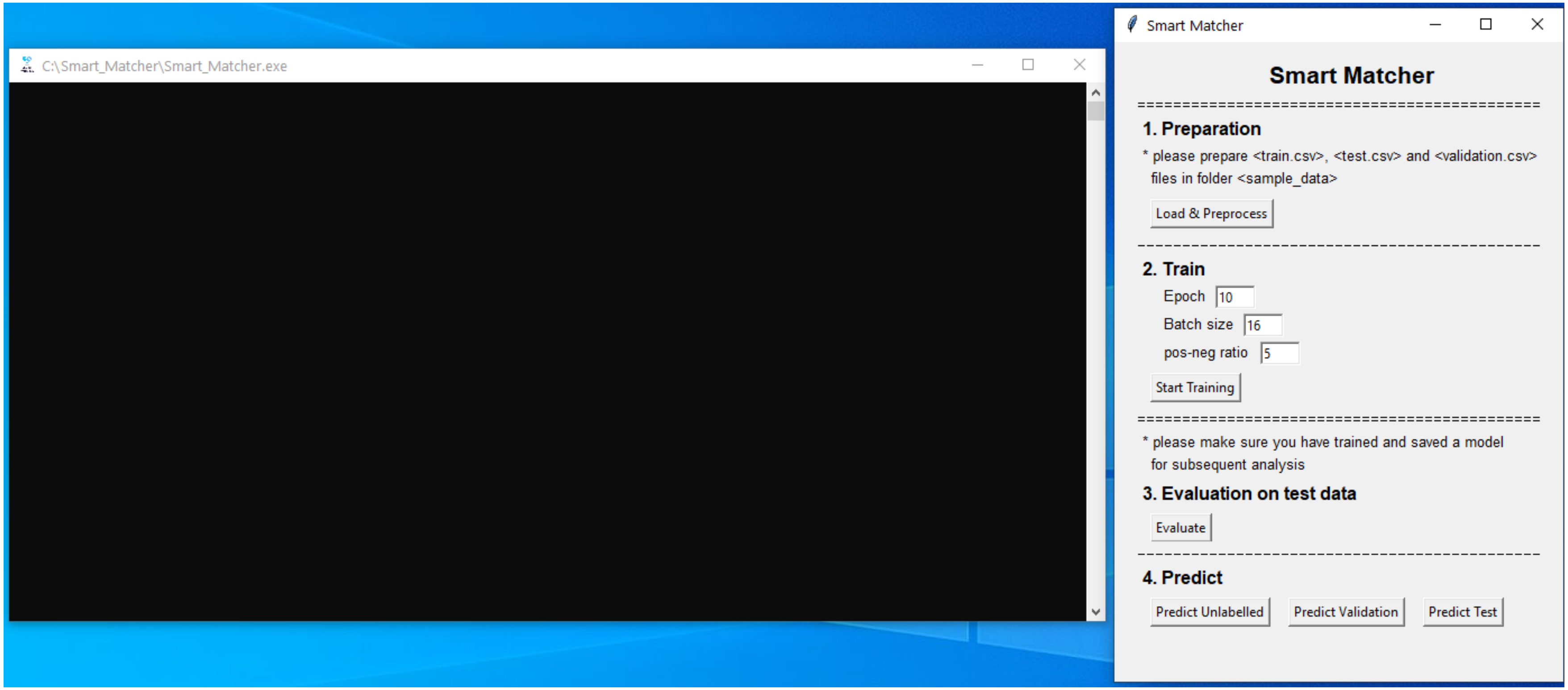
Then we will see a command window which initializes the environment and launches *Smart Matcher*. This may take a minute.



Quick start - data preprocessing

The command window is for us to monitor preprocessing, training, evaluation and prediction (also for developers to debug).

Hence, maybe a good layout is like this



Smart Matcher

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

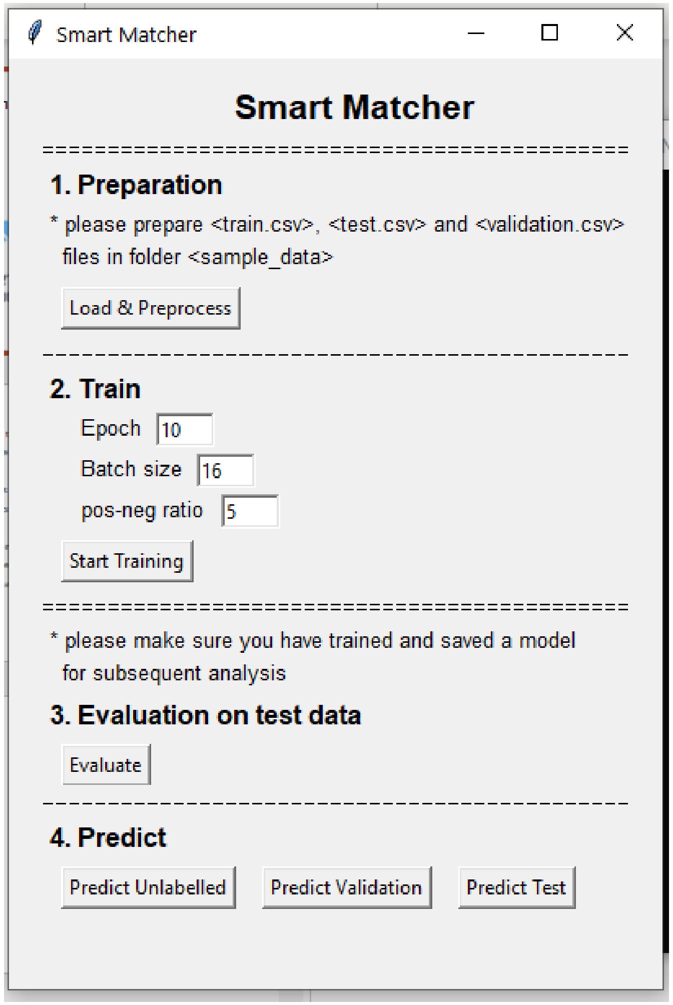
* please make sure you have trained and saved a model for subsequent analysis

3. Evaluation on test data

4. Predict

Quick start - data preprocessing

The GUI of *Smart Matcher* will appear



Smart Matcher

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

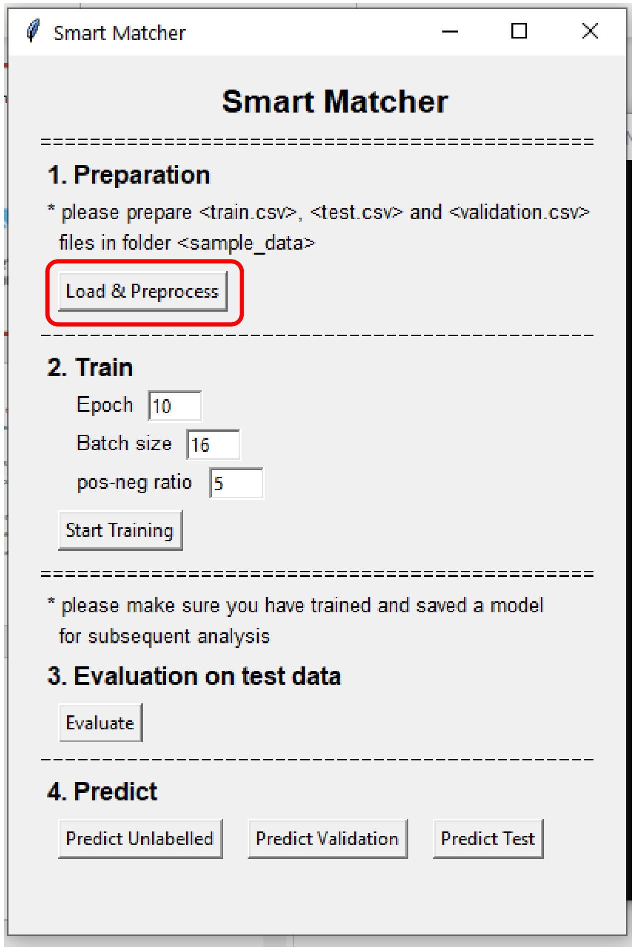
* please make sure you have trained and saved a model for subsequent analysis

3. Evaluation on test data

4. Predict

Quick start - data preprocessing

Since we have prepared sample datasets and pre-trained embeddings in relevant folders, we just click button <Load & Preprocess> to preprocess all datasets



Smart Matcher

Smart Matcher

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

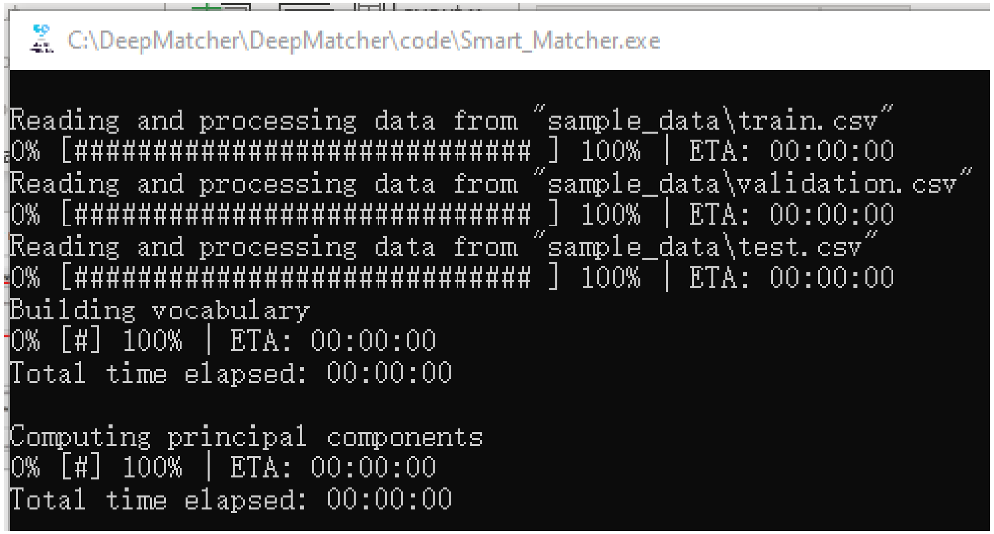
* please make sure you have trained and saved a model for subsequent analysis

3. Evaluation on test data

4. Predict

Quick start - data preprocessing

The progress can be viewed in the command window below

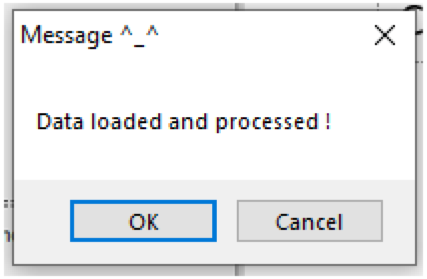


```
Reading and processing data from "sample_data\train.csv"
0% [############################## ] 100% | ETA: 00:00:00
Reading and processing data from "sample_data\validation.csv"
0% [############################## ] 100% | ETA: 00:00:00
Reading and processing data from "sample_data\test.csv"
0% [############################## ] 100% | ETA: 00:00:00
Building vocabulary
0% [#] 100% | ETA: 00:00:00
Total time elapsed: 00:00:00

Computing principal components
0% [#] 100% | ETA: 00:00:00
Total time elapsed: 00:00:00
```

Quick start - data preprocessing

After the preprocessing is done, a pop-up message window will appear



Message ^_^

Data loaded and processed !

Click OK to continue

Quick start - training

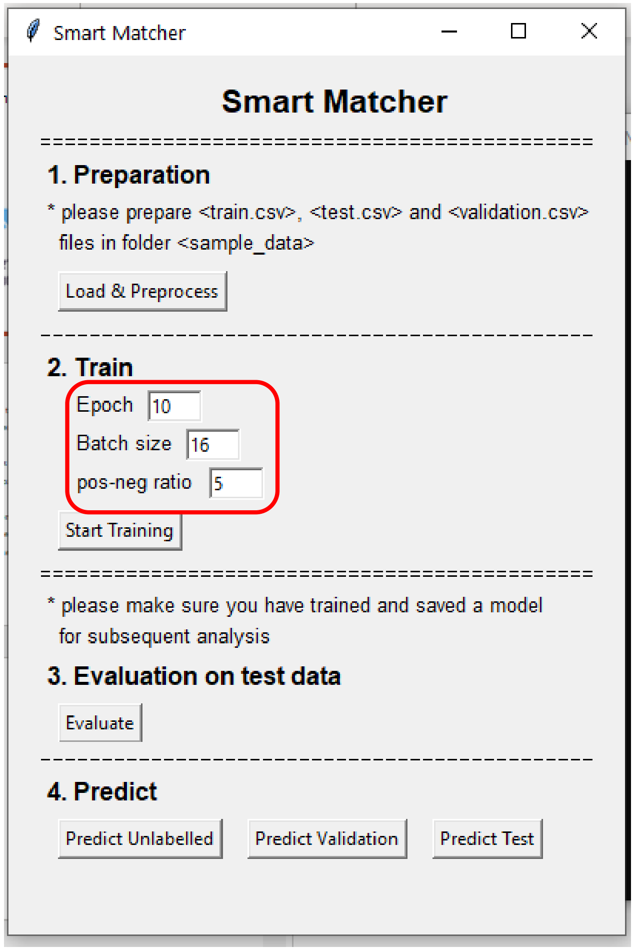
Set training parameters

Epoch: 10 or 15 (recommended)

Batch size: 16 or 32 (recommended)

Pos-neg ratio:

- Sampling ratio between positive and negative examples
- Dataset-dependent
- For example, this value should be increased if we have fewer matches than non-matches in your data
- For the provided datasets, we should put '5'



Smart Matcher

Smart Matcher

==

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

----

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

==

* please make sure you have trained and saved a model for subsequent analysis

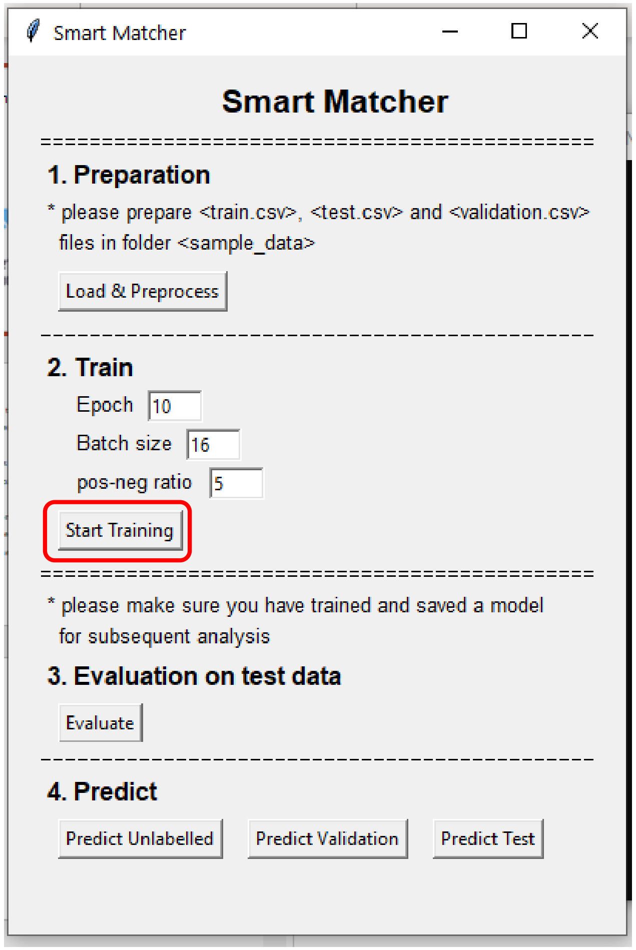
3. Evaluation on test data

----

4. Predict

Quick start - training

Click <Start Training>



Smart Matcher

Smart Matcher

===

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

----

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

===

* please make sure you have trained and saved a model for subsequent analysis

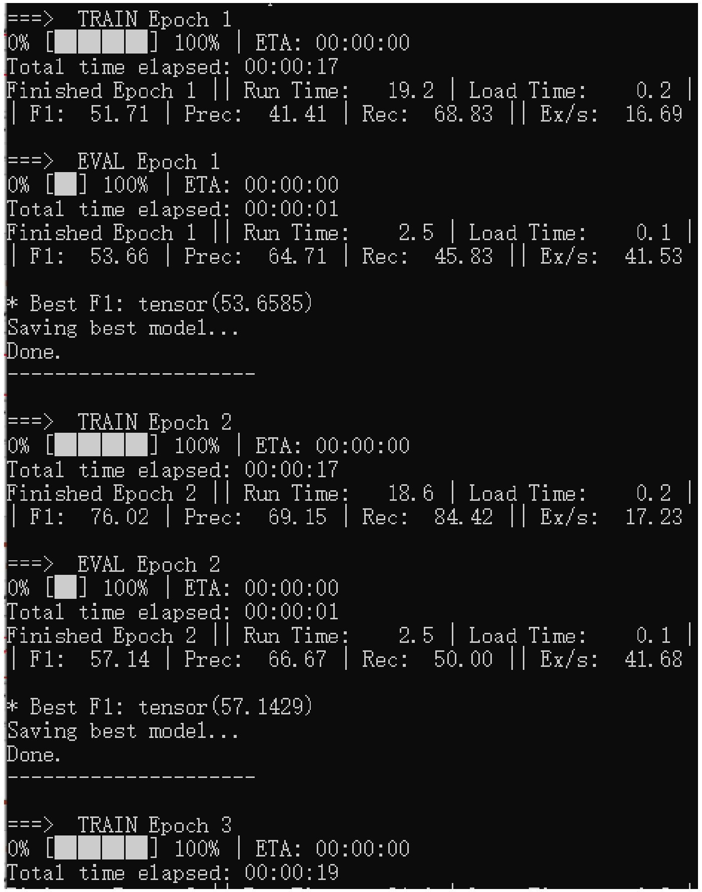
3. Evaluation on test data

----

4. Predict

Quick start - training

The training progress can be monitored in the command window



```
===>  TRAIN Epoch 1
0% [█████] 100% | ETA: 00:00:00
Total time elapsed: 00:00:17
Finished Epoch 1 || Run Time:   19.2 | Load Time:    0.2 |
| F1:  51.71 | Prec:  41.41 | Rec:  68.83 || Ex/s:  16.69

===>  EVAL Epoch 1
0% [█] 100% | ETA: 00:00:00
Total time elapsed: 00:00:01
Finished Epoch 1 || Run Time:    2.5 | Load Time:    0.1 |
| F1:  53.66 | Prec:  64.71 | Rec:  45.83 || Ex/s:  41.53

* Best F1: tensor(53.6585)
Saving best model...
Done.
---------------------

===>  TRAIN Epoch 2
0% [█████] 100% | ETA: 00:00:00
Total time elapsed: 00:00:17
Finished Epoch 2 || Run Time:   18.6 | Load Time:    0.2 |
| F1:  76.02 | Prec:  69.15 | Rec:  84.42 || Ex/s:  17.23

===>  EVAL Epoch 2
0% [█] 100% | ETA: 00:00:00
Total time elapsed: 00:00:01
Finished Epoch 2 || Run Time:    2.5 | Load Time:    0.1 |
| F1:  57.14 | Prec:  66.67 | Rec:  50.00 || Ex/s:  41.68

* Best F1: tensor(57.1429)
Saving best model...
Done.
---------------------

===>  TRAIN Epoch 3
0% [█████] 100% | ETA: 00:00:00
Total time elapsed: 00:00:19
```

Parameters to monitor

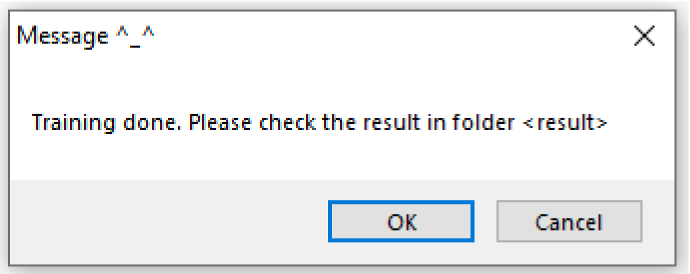
Precision (P): the fraction of match predictions that are correct

Recall (R): the fraction of correct matches being predicted as matches

F1 score (F1): 2PR/(P + R)

Quick start - training

After the training is done, a pop-up message window will appear

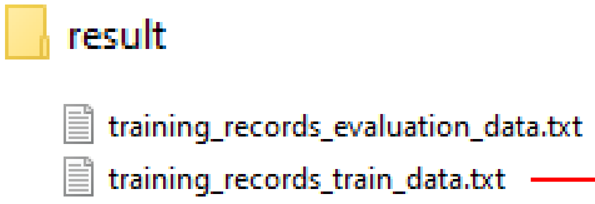


Message ^_^

Training done. Please check the result in folder <result>

Click OK to continue

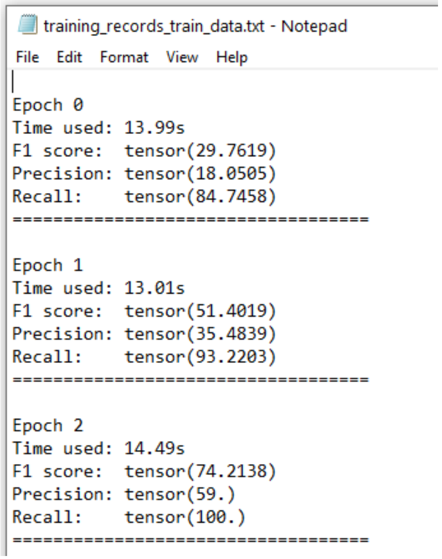
All training records about training and evaluation datasets can be found in folder <result>



result

training_records_evaluation_data.txt

training_records_train_data.txt



training_records_train_data.txt - Notepad

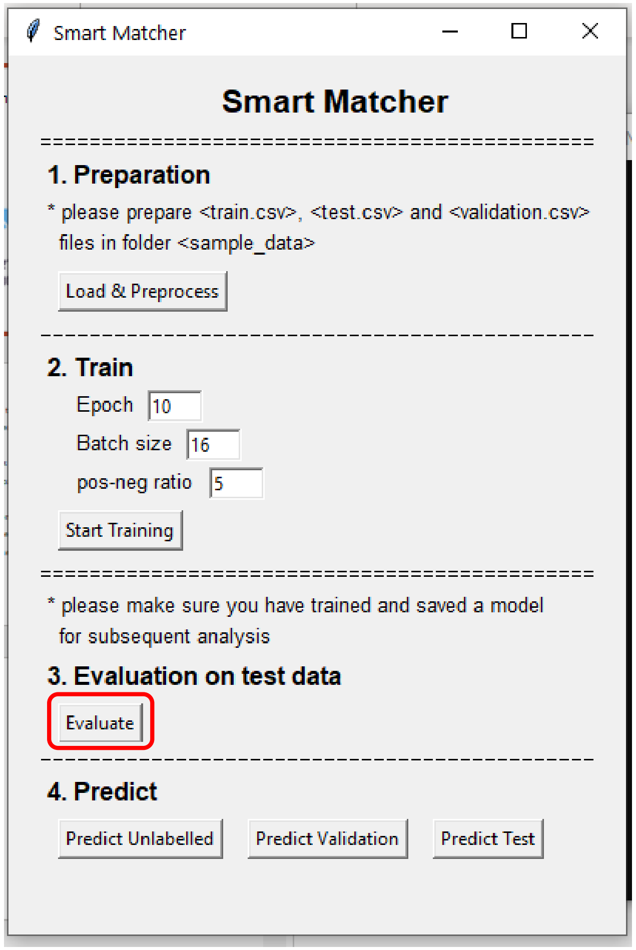
```
Epoch 0
Time used: 13.99s
F1 score:  tensor(29.7619)
Precision: tensor(18.0505)
Recall:    tensor(84.7458)
===================================

Epoch 1
Time used: 13.01s
F1 score:  tensor(51.4019)
Precision: tensor(35.4839)
Recall:    tensor(93.2203)
===================================

Epoch 2
Time used: 14.49s
F1 score:  tensor(74.2138)
Precision: tensor(59.)
Recall:    tensor(100.)
===================================
```

Quick start - evaluation

Click <Evaluation> to check the performance on the test dataset



Smart Matcher

Smart Matcher

==

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

----

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

==

* please make sure you have trained and saved a model for subsequent analysis

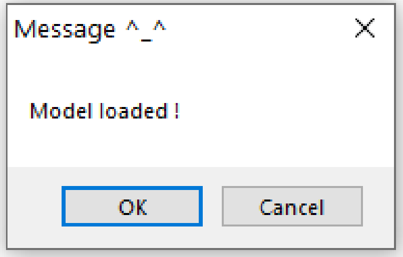
3. Evaluation on test data

----

4. Predict

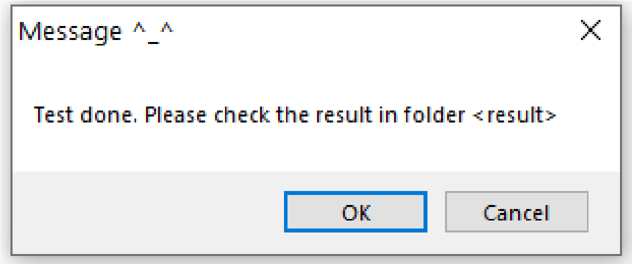
Quick start - evaluation

The evaluation will load the trained model first
Click OK to continue


Message ^_^

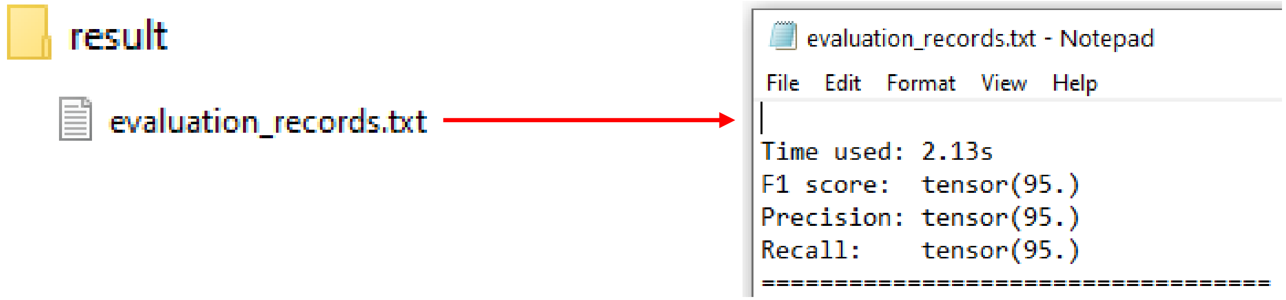
Model loaded !

After the evaluation is done, click OK to continue


Message ^_^

Test done. Please check the result in folder <result>

The performance can be found in folder <result>



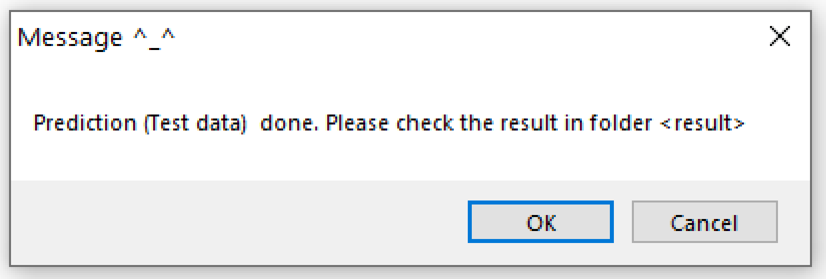
```
Time used: 2.13s
F1 score:  tensor(95.)
Precision: tensor(95.)
Recall:    tensor(95.)
====================================
```

Quick start - prediction

Click the highlighted two buttons to check predictions for validation and test datasets.

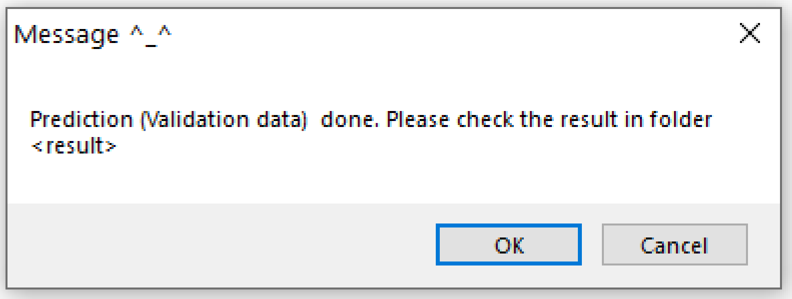
This is how we can ‘feel’ the performance of the model

Click OK to continue



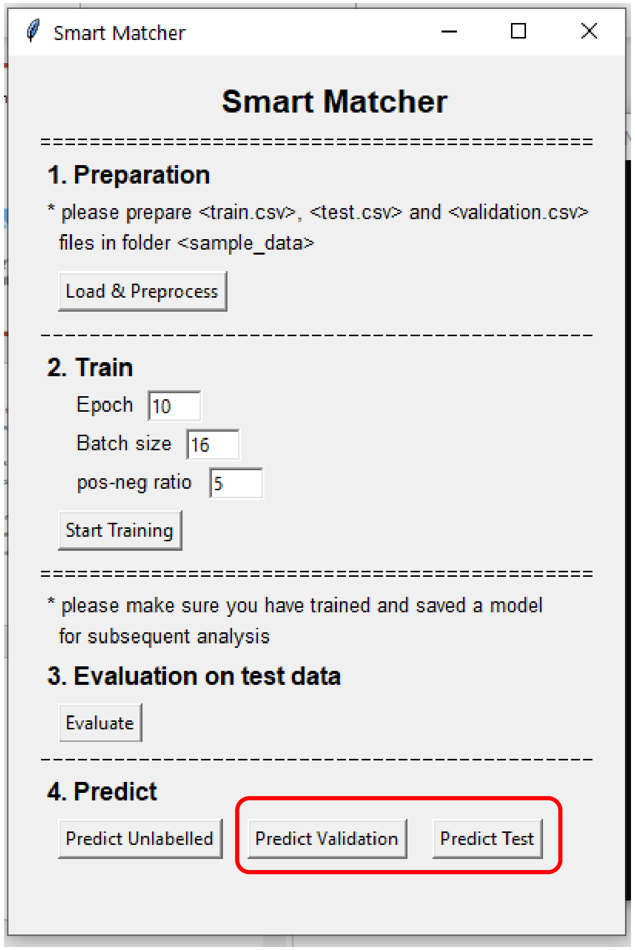
Message ^_^

Prediction (Test data) done. Please check the result in folder <result>



Message ^_^

Prediction (Validation data) done. Please check the result in folder <result>



Smart Matcher

Smart Matcher

1. Preparation

* please prepare <train.csv>, <test.csv> and <validation.csv> files in folder <sample_data>

2. Train

Epoch 10

Batch size 16

pos-neg ratio 5

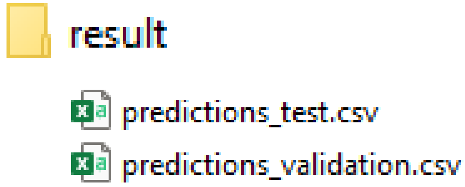
* please make sure you have trained and saved a model for subsequent analysis

3. Evaluation on test data

4. Predict

Quick start - prediction

Check the performance in folder <result>

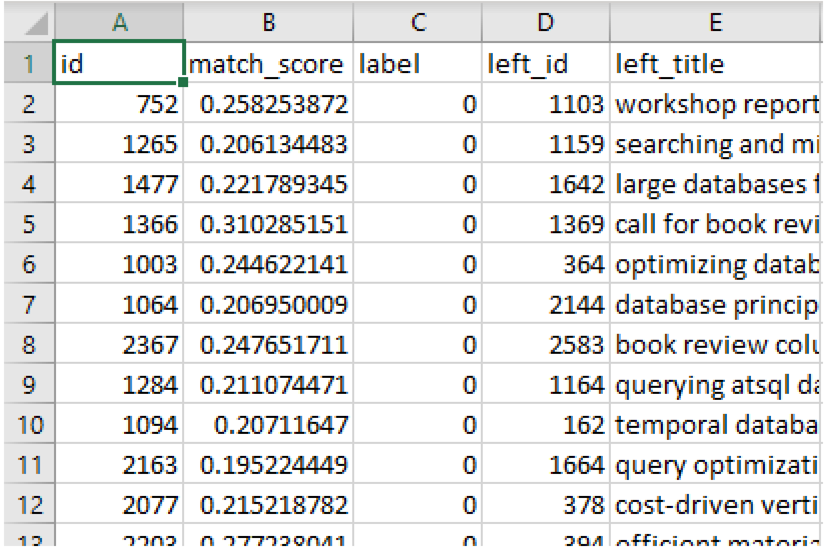


result

predictions_test.csv

predictions_validation.csv

A new column <match_score> will be added to indicate the ‘confidence’



	A	B	C	D	E
1	id	match_score	label	left_id	left_title
2	752	0.258253872	0	1103	workshop report
3	1265	0.206134483	0	1159	searching and mi
4	1477	0.221789345	0	1642	large databases f
5	1366	0.310285151	0	1369	call for book revi
6	1003	0.244622141	0	364	optimizing datab
7	1064	0.206950009	0	2144	database princip
8	2367	0.247651711	0	2583	book review colu
9	1284	0.211074471	0	1164	querying atsql da
10	1094	0.20711647	0	162	temporal databa
11	2163	0.195224449	0	1664	query optimizati
12	2077	0.215218782	0	378	cost-driven verti
13	2203	0.277238041	0	394	efficient materia

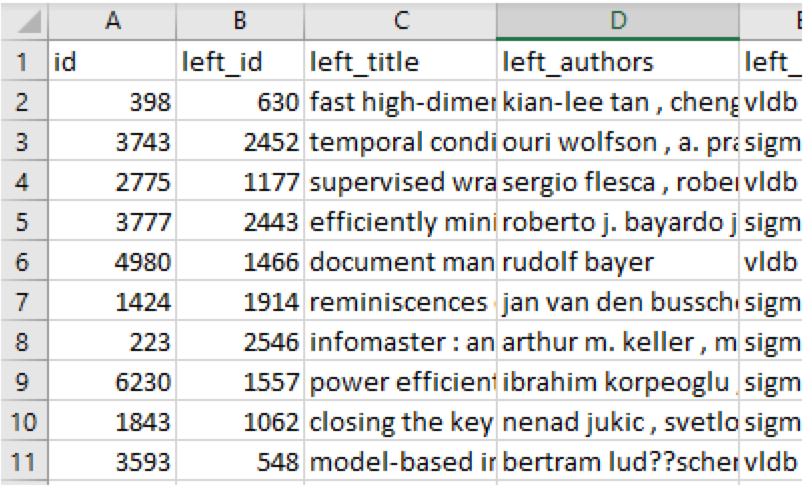
** In general, ‘confidence’ above 0.5 is considered a match*

Quick start - prediction

The unlabelled means a dataset without column <label>

For example,

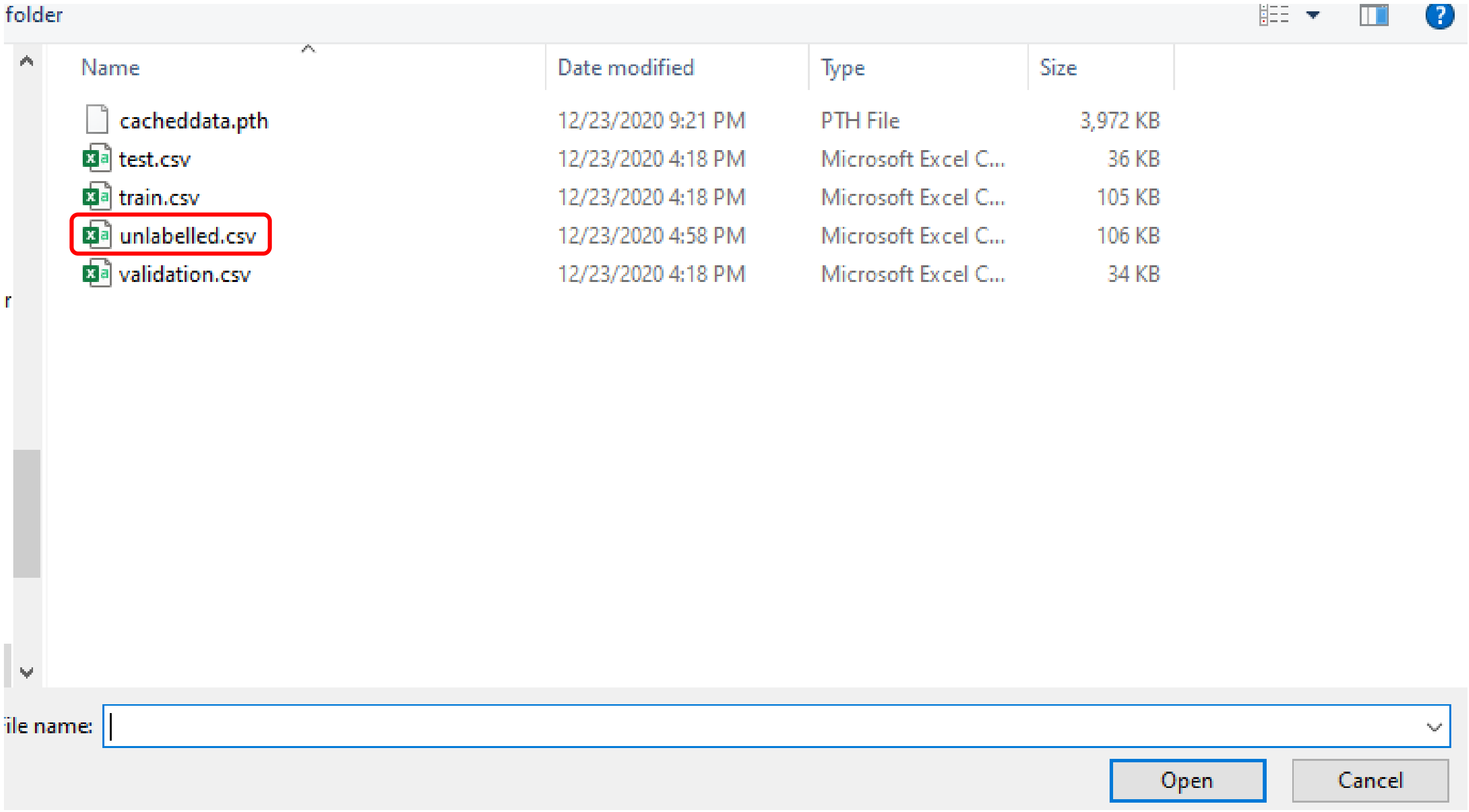
- id
- left_A
- left_B
- …
- right_A
- right_B
- …



	A	B	C	D	E
1	id	left_id	left_title	left_authors	left_
2	398	630	fast high-dimer	kian-lee tan , cheng	vldb
3	3743	2452	temporal condi	ouri wolfson , a. pra	sigm
4	2775	1177	supervised wra	sergio flesca , rober	vldb
5	3777	2443	efficiently mini	roberto j. bayardo j	sigm
6	4980	1466	document man	rudolf bayer	vldb
7	1424	1914	reminiscences	jan van den bussch	sigm
8	223	2546	infomaster : an	arthur m. keller , m	sigm
9	6230	1557	power efficient	ibrahim korpeoglu	sigm
10	1843	1062	closing the key	nenad jukic , svetlo	sigm
11	3593	548	model-based ir	bertram lud??scher	vldb

Quick start - prediction

Click <Predict Unlabelled> to select an unlabelled dataset

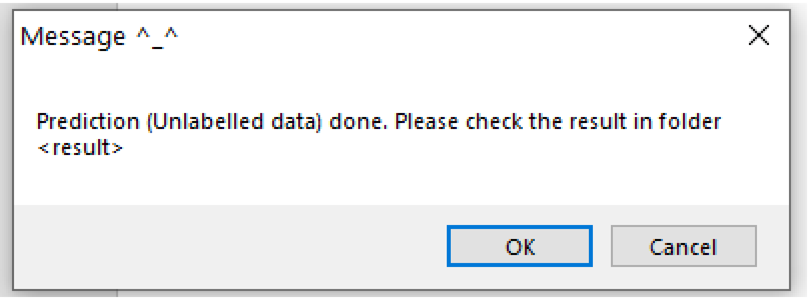


Name	Date modified	Type	Size
cacheddata.pth	12/23/2020 9:21 PM	PTH File	3,972 KB
test.csv	12/23/2020 4:18 PM	Microsoft Excel C...	36 KB
train.csv	12/23/2020 4:18 PM	Microsoft Excel C...	105 KB
unlabelled.csv	12/23/2020 4:58 PM	Microsoft Excel C...	106 KB
validation.csv	12/23/2020 4:18 PM	Microsoft Excel C...	34 KB

Double-click the file to continue

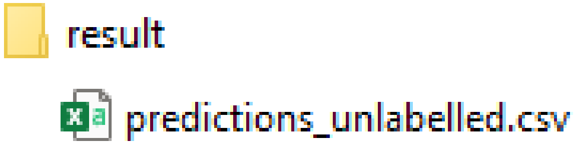
Quick start - prediction

After the prediction is done, click OK to continue



Prediction (Unlabelled data) done. Please check the result in folder <result>

Check the performance in folder <result>

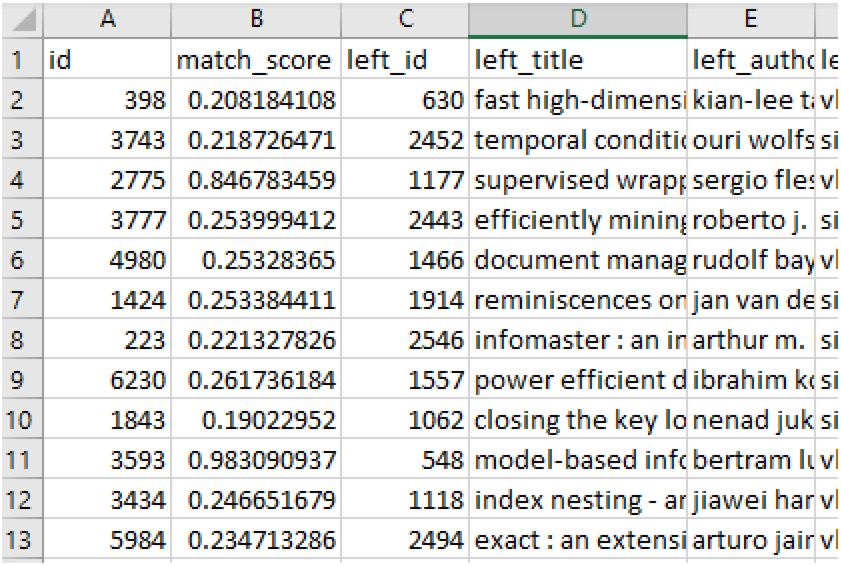


result

predictions_unlabelled.csv

Quick start - prediction

A new column <match_score> will be added to indicate the ‘confidence’



	A	B	C	D	E	
1	id	match_score	left_id	left_title	left_authc	le
2	398	0.208184108	630	fast high-dimensi	kian-lee t	vl
3	3743	0.218726471	2452	temporal conditi	ouri wolfs	si
4	2775	0.846783459	1177	supervised wrapp	sergio fle	vl
5	3777	0.253999412	2443	efficiently mining	roberto j.	si
6	4980	0.25328365	1466	document manag	rudolf bay	vl
7	1424	0.253384411	1914	reminiscences on	jan van de	si
8	223	0.221327826	2546	infomaster : an in	arthur m.	si
9	6230	0.261736184	1557	power efficient d	ibrahim k	si
10	1843	0.19022952	1062	closing the key lo	nenad juk	si
11	3593	0.983090937	548	model-based info	bertram l	vl
12	3434	0.246651679	1118	index nesting - ar	jiawei har	vl
13	5984	0.234713286	2494	exact : an extensi	arturo jair	vl

** In general, ‘confidence’ above 0.5 is considered a match*

Done. Enjoy using !

SMART
MATCHER

Appendix

This software can be used to match different types of entities.

Here we show some example datasets that can be analyzed.

Appendix - Example datasets - DBLP-GoogleScholar

title	auther	venue	year
a performance study of workfile disk management for concurrent mergesorts in a multiprocessor database system	k wu , p yu , j chung , j teng	vldb	1995
fastmap : a fast algorithm for indexing , data-mining and visualization of traditional and multimedia datasets	c faloutsos , k lin	sigmod conference	1995
semantic integration of environmental models for application to global information systems and decision-making	d mackay	sigmod record	1999
deadlock detection in distributed database systems : a new algorithm and a comparative performance analysis	n krivokapic , a kemper , e gudes	vldb j.	1999

Accuracy (F1 core)	
Structured	94.7 - 95.1
Dirty (with missing information)	92.7 - 93.8

Appendix - Example datasets - iTunes-Amazon

Song	Artist Name	Album Name	Genre	Price	CopyRight	Time	Released
Ca n't Stop Now (feat . Jovi Rockwell and Mr. Vegas)	Major Lazer	Guns Do n't Kill People ... Lazers Do	Electronic,Music,Hip-Hop / Rap , Rap , Alternative , Reggae , Dance , Modern Dancehall , Rock	S$ 1.29	?€? ???? 2009 Downtown Music , LLC .	4.03	2009
I 'm a Machine (feat . Crystal Nicole and Tyrese Gibson)	David Guetta	Nothing But the Beat	Dance , Music , House , Electronic , Rock	S$ 1.29	2011 What A Music Ltd , Licence exclusive Parlophone Music France	3.34	8/26/2011

Accuracy (F1 core)	
Structured	88.0 - 90.9
Dirty (with missing information)	69.2 - 74.5