
CLK hash Documentation

Release 0.12.1

N1 Analytics

Jun 26, 2019

Contents

1	Table of Contents	3
2	External Links	47
3	Indices and tables	49
	Bibliography	51
	Python Module Index	53
	Index	55

`clkhsh` is a python implementation of cryptographic linkage key hashing as described by Rainer Schnell, Tobias Bachteler, and Jörg Reiher in *A Novel Error-Tolerant Anonymous Linking Code* [[Schnell2011](#)].

Clkhash is Apache 2.0 licensed, supports Python versions 2.7+, 3.4+, and runs on Windows, OSX and Linux.

Install with pip:

```
pip install clkhsh
```

Hint: If you are interested in comparing CLK encodings (i.e carrying out record linkage) you might want to check out [anonlink](#) and [anonlink-entity-service](#) - our Python library and REST service for computing similarity scores, and matching between sets of cryptographic linkage keys.

1.1 Tutorials

The clkh hash library can be used via the Python API or the command line tool *clkutil*.

1.1.1 Tutorial for Python API

For this tutorial we are going to process a data set for private linkage with clkh hash using the Python API. Note you can also use the command line tool.

The Python package `recordlinkage` has a [tutorial](#) linking data sets in the clear, we will try duplicate that in a privacy preserving setting.

First install clkh hash, recordlinkage and a few data science tools (pandas and numpy):

```
$ pip install -U clkh hash anonlink recordlinkage numpy pandas
```

```
[1]: import io
import numpy as np
import pandas as pd
```

```
[2]: import clkh hash
from clkh hash import clk
from clkh hash.field_formats import *
from clkh hash.schema import Schema
```

```
[3]: import recordlinkage
from recordlinkage.datasets import load_febr14
```

Data Exploration

First we have a look at the dataset.

```
[4]: dfA, dfB = load_febrl4()

dfA.head()
```

	given_name	surname	street_number	address_1	\
rec_id					
rec-1070-org	michaela	neumann	8	stanley street	
rec-1016-org	courtney	painter	12	pinkerton circuit	
rec-4405-org	charles	green	38	salkauskas crescent	
rec-1288-org	vanessa	parr	905	macquoid place	
rec-3585-org	mikayla	malloney	37	randwick road	

	address_2	suburb	postcode	state	\
rec_id					
rec-1070-org	miami	winston hills	4223	nsw	
rec-1016-org	bega flats	richlands	4560	vic	
rec-4405-org	kela	dapto	4566	nsw	
rec-1288-org	broadbridge manor	south grafton	2135	sa	
rec-3585-org	avalind	hoppers crossing	4552	vic	

	date_of_birth	soc_sec_id
rec_id		
rec-1070-org	19151111	5304218
rec-1016-org	19161214	4066625
rec-4405-org	19480930	4365168
rec-1288-org	19951119	9239102
rec-3585-org	19860208	7207688

For this linkage we will **not** use the social security id column.

```
[5]: dfA.columns

[5]: Index(['given_name', 'surname', 'street_number', 'address_1', 'address_2',
         'suburb', 'postcode', 'state', 'date_of_birth', 'soc_sec_id'],
        dtype='object')
```

```
[6]: a_csv = io.StringIO()
dfA.to_csv(a_csv)
```

Hashing Schema Definition

A hashing schema instructs clkhash how to treat each column for generating CLKs. A detailed description of the hashing schema can be found in the [api docs](#). We will ignore the columns 'rec_id' and 'soc_sec_id' for CLK generation.

```
[7]: fields = [
    Ignore('rec_id'),
    StringSpec('given_name', FieldHashingProperties(ngram=2, num_bits=300)),
    StringSpec('surname', FieldHashingProperties(ngram=2, num_bits=300)),
    IntegerSpec('street_number', FieldHashingProperties(ngram=1, positional=True, num_
    ↳bits=300, missing_value=MissingValueSpec(sentinel=''))),
    StringSpec('address_1', FieldHashingProperties(ngram=2, num_bits=300)),
    StringSpec('address_2', FieldHashingProperties(ngram=2, num_bits=300)),
    StringSpec('suburb', FieldHashingProperties(ngram=2, num_bits=300)),
    IntegerSpec('postcode', FieldHashingProperties(ngram=1, positional=True, num_
    ↳bits=300)),
    StringSpec('state', FieldHashingProperties(ngram=2, num_bits=300)),
```

(continues on next page)

(continued from previous page)

```
IntegerSpec('date_of_birth', FieldHashingProperties(ngram=1, positional=True, num_
↳bits=300, missing_value=MissingValueSpec(sentinel=''))),
Ignore('soc_sec_id')
]

schema = Schema(fields, 1024)
```

Hash the data

We can now hash our PII data from the CSV file using our defined schema. We must provide a list of *secret keys* to this command - these keys have to be used by both parties hashing data. For this toy example we will use the keys 'key1' and 'key2', for real data, make sure that the keys contain enough entropy, as knowledge of these keys is sufficient to reconstruct the PII information from a CLK!

Also, **do not share these keys with anyone, except the other participating party.**

```
[8]: secret_keys = ('key1', 'key2')
```

```
[9]: a_csv.seek(0)
hashed_data_a = clk.generate_clk_from_csv(a_csv, secret_keys, schema)

generating CLKs: 100%|| 5.00k/5.00k [00:01<00:00, 786clk/s, mean=950, std=9.79]
```

Inspect the output

clkhsh has hashed the PII, creating a Cryptographic Longterm Key for each entity. The output of `generate_clk_from_csv` shows that the mean popcount is quite high (950 out of 1024) which can affect accuracy.

We can control the popcount by adjusting the hashing strategy. There are currently two different strategies implemented in the library. - *fixed k*: each n-gram of a feature's value is inserted into the CLK *k* times. Increasing *k* will give the corresponding feature more importance in comparisons, decreasing *k* will de-emphasise columns which are less suitable for linkage (e.g. information that changes frequently). The *fixed k* strategy is set with the 'k=30' argument for each feature's `FieldHashingProperties`. (for a total of `numberOfTokens * k` insertions) - *fixed number of bits*: In this strategy we always insert a fixed number of bits into the CLK for a feature, irrespective of the number of n-grams. This strategy is set with the 'numBits=100' argument for each feature's `FieldHashingProperties`.

In this example, we will reduce the value of `num_bits` for address related columns.

```
[10]: fields = [
    Ignore('rec_id'),
    StringSpec('given_name', FieldHashingProperties(ngram=2, num_bits=200)),
    StringSpec('surname', FieldHashingProperties(ngram=2, num_bits=200)),
    IntegerSpec('street_number', FieldHashingProperties(ngram=1, positional=True, num_
↳bits=100, missing_value=MissingValueSpec(sentinel=''))),
    StringSpec('address_1', FieldHashingProperties(ngram=2, num_bits=100)),
    StringSpec('address_2', FieldHashingProperties(ngram=2, num_bits=100)),
    StringSpec('suburb', FieldHashingProperties(ngram=2, num_bits=100)),
    IntegerSpec('postcode', FieldHashingProperties(ngram=1, positional=True, num_
↳bits=100)),
    StringSpec('state', FieldHashingProperties(ngram=2, num_bits=100)),
    IntegerSpec('date_of_birth', FieldHashingProperties(ngram=1, positional=True, num_
↳bits=200, missing_value=MissingValueSpec(sentinel=''))),
    Ignore('soc_sec_id')
```

(continues on next page)

(continued from previous page)

```
]

schema = Schema(fields, 1024)
a_csv.seek(0)
hashed_data_a = clk.generate_clk_from_csv(a_csv, secret_keys, schema)

generating CLKs: 100%|| 5.00k/5.00k [00:00<00:00, 1.41kclk/s, mean=705, std=15.5]
```

Each CLK is serialized in a JSON friendly base64 format:

```
[11]: hashed_data_a[0]
[11]: 'wTmf3/rPF3Pj/85fORXpee/9+v3/1o9714/7d/bW+G7+9N3Cij//a1//nr/9/cZn/BT9+kWn19203/
↪eOtvM4G4s3e8lX+7X+f0kXez7XbOfevz7/r6wvN99Mncp367yPeZW3uMYv9Evf9/sPuOq3+p79t6/qn/
↪v7O5e/Jurvr8='
```

Hash data set B

Now we hash the second dataset using the same keys and same schema.

```
[12]: b_csv = io.StringIO()
dfB.to_csv(b_csv)
b_csv.seek(0)
hashed_data_b = clkhask.clk.generate_clk_from_csv(b_csv, secret_keys, schema)

generating CLKs: 100%|| 5.00k/5.00k [00:00<00:00, 1.45kclk/s, mean=703, std=19.1]
```

```
[13]: len(hashed_data_b)
[13]: 5000
```

Find matches between the two sets of CLKs

We have generated two sets of CLKs which represent entity information in a privacy-preserving way. The more similar two CLKs are, the more likely it is that they represent the same entity.

For this task we will use [anonlink](#), a Python (and optimised C++) implementation of anonymous linkage using CLKs.

As the CLKs are in a string format we first deserialize to use the bitarray type:

```
[14]: from bitarray import bitarray
import base64

def deserialize_bitarray(bytes_data):
    ba = bitarray(endian='big')
    data_as_bytes = base64.decodebytes(bytes_data.encode())
    ba.frombytes(data_as_bytes)
    return ba

def deserialize_filters(filters):
    res = []
    for i, f in enumerate(filters):
        ba = deserialize_bitarray(f)
        res.append(ba)
    return res
```

(continues on next page)

(continued from previous page)

```

clks_a = deserialize_filters(hashed_data_a)
clks_b = deserialize_filters(hashed_data_b)

```

Using anonlink we find the candidate pairs - which is all possible pairs above the given threshold. Then we solve for the most likely mapping.

```

[15]: import anonlink

def mapping_from_clks(clks_a, clks_b, threshold):
    results_candidate_pairs = anonlink.candidate_generation.find_candidate_pairs(
        [clks_a, clks_b],
        anonlink.similarities.dice_coefficient,
        threshold
    )
    solution = anonlink.solving.greedy_solve(results_candidate_pairs)
    print('Found {} matches'.format(len(solution)))
    return {a:b for (_, a), (_, b) in solution}

```

```

[16]: mapping = mapping_from_clks(clks_a, clks_b, 0.9)

```

```

Found 4019 matches

```

Let's investigate some of those matches and the overall matching quality

```

[17]: a_csv.seek(0)
      b_csv.seek(0)
      a_raw = a_csv.readlines()
      b_raw = b_csv.readlines()

      num_entities = len(b_raw) - 1

      def describe_accuracy(mapping, show_examples=False):
          if show_examples:
              print('idx_a, idx_b,      rec_id_a,      rec_id_b')
              print('-----')
              for a_i in range(10):
                  if a_i in mapping:
                      a_data = a_raw[a_i + 1].split(',')
                      b_data = b_raw[mapping[a_i] + 1].split(',')
                      print('{:3}, {:6}, {:>15}, {:>15}'.format(a_i+1, mapping[a_i]+1, a_
→data[0], b_data[0]))
                      print('-----')

              TP = 0; FP = 0; TN = 0; FN = 0
              for a_i in range(num_entities):
                  if a_i in mapping:
                      if a_raw[a_i + 1].split(',')[0].split('-')[1] == b_raw[mapping[a_i] + 1].
→split(',')[0].split('-')[1]:
                          TP += 1
                      else:
                          FP += 1
                          # as we only report one mapping for each element in PII_a,
                          # then a wrong mapping is not only a false positive, but
                          # also a false negative, as we won't report the true mapping.
                          FN += 1
                  else:

```

(continues on next page)

(continued from previous page)

```

    FN += 1 # every element in PII_a has a partner in PII_b

    print()
    print("We've got {} true positives, {} false positives, and {} false negatives.".
→format(TP, FP, FN))
    print('Precision: {:.3f}, Recall: {:.3f}, Accuracy: {:.3f}'.format(
        TP / (TP+FP),
        TP / (TP+FN),
        (TP+TN) / (TP+TN+FP+FN)))

```

```
[18]: describe_accuracy(mapping, show_examples=True)
```

idx_a,	idx_b,	rec_id_a,	rec_id_b
2,	2751,	rec-1016-org,	rec-1016-dup-0
3,	4657,	rec-4405-org,	rec-4405-dup-0
4,	4120,	rec-1288-org,	rec-1288-dup-0
5,	3307,	rec-3585-org,	rec-3585-dup-0
6,	2306,	rec-298-org,	rec-298-dup-0
7,	3945,	rec-1985-org,	rec-1985-dup-0
8,	993,	rec-2404-org,	rec-2404-dup-0
9,	4613,	rec-1473-org,	rec-1473-dup-0
10,	3630,	rec-453-org,	rec-453-dup-0

```

We've got 4019 true positives, 0 false positives, and 981 false negatives.
Precision: 1.000, Recall: 0.804, Accuracy: 0.804

```

Precision tells us about how many of the found matches are actual matches. The score of 1.0 means that we did perfectly in this respect, however, recall, the measure of how many of the actual matches were correctly identified, is quite low with only 80%.

Let's go back to the mapping calculation (`calculate_mapping_greedy`) and reduce the value for `threshold` to 0.8.

```
[19]: mapping = mapping_from_clks(clks_a, clks_b, 0.8)
describe_accuracy(mapping)
```

```
Found 4974 matches
```

```

We've got 4974 true positives, 0 false positives, and 26 false negatives.
Precision: 1.000, Recall: 0.995, Accuracy: 0.995

```

Great, for this threshold value we get a precision of 100% and a recall of 99.6%.

The explanation is that when the information about an entity differs slightly in the two datasets (e.g. spelling errors, abbreviations, missing values, ...) then the corresponding CLKs will differ in some number of bits as well. It is important to choose an appropriate threshold for the amount of perturbations present in the data (a threshold of 0.74 and below generates a mapping which misses just one true match).

This concludes the tutorial. Feel free to go back to the CLK generation and experiment on how different setting will affect the matching quality.

1.1.2 Tutorial for CLI tool clkhsh

For this tutorial we are going to process a data set for private linkage with clkhsh using the command line tool `clkutil` - equivalent to running `python -m clkhsh`.

Note you can also use the [Python API](#).

The Python package `recordlinkage` has a [tutorial](#) linking data sets in the clear, we will try duplicate that in a privacy preserving setting.

First install clkhsh, recordlinkage and a few data science tools (pandas and numpy).

```
$ pip install -U clkhsh recordlinkage numpy pandas
```

```
[1]: import json
import numpy as np
import pandas as pd
```

```
[2]: import recordlinkage
from recordlinkage.datasets import load_febrl4
```

Data Exploration

First we have a look at the dataset.

```
[3]: dfA, dfB = load_febrl4()
```

```
dfA.head()
```

```
[3]:
```

	given_name	surname	street_number	address_1	\
rec_id					
rec-1070-org	michaela	neumann	8	stanley street	
rec-1016-org	courtney	painter	12	pinkerton circuit	
rec-4405-org	charles	green	38	salkauskas crescent	
rec-1288-org	vanessa	parr	905	macquoid place	
rec-3585-org	mikayla	malloney	37	randwick road	

	address_2	suburb	postcode	state	\
rec_id					
rec-1070-org	miami	winston hills	4223	nsw	
rec-1016-org	bega flats	richlands	4560	vic	
rec-4405-org	kela	dapto	4566	nsw	
rec-1288-org	broadbridge manor	south grifton	2135	sa	
rec-3585-org	avalind	hoppers crossing	4552	vic	

	date_of_birth	soc_sec_id
rec_id		
rec-1070-org	19151111	5304218
rec-1016-org	19161214	4066625
rec-4405-org	19480930	4365168
rec-1288-org	19951119	9239102
rec-3585-org	19860208	7207688

Note that for computing this linkage we will **not** use the social security id column or the `rec_id` index.

```
[4]: dfA.columns
```

```
[4]: Index(['given_name', 'surname', 'street_number', 'address_1', 'address_2',
          'suburb', 'postcode', 'state', 'date_of_birth', 'soc_sec_id'],
          dtype='object')
```

```
[5]: dfA.to_csv('PII_a.csv')
```

Hashing Schema Definition

A hashing schema instructs clkhsh how to treat each column for generating CLKs. A detailed description of the hashing schema can be found in the [api docs](#). We will ignore the columns 'rec_id' and 'soc_sec_id' for CLK generation.

```
[6]: %cat _static/febrl_schema_v2_overweight.json
{
  "version": 2,
  "clkConfig": {
    "l": 1024,
    "kdf": {
      "type": "HKDF",
      "hash": "SHA256",
      "info": "c2NoZWlhX2V4YW1wbGU=",
      "salt": "SCbL2zHNnmsckfzchsNkZY9XoHk96P/
↪G5nUBrM7ybymLEFsMV6PAeDZCNp3rfNUPCtLDMOGQH4pCQpfhiHCyA==",
      "keySize": 64
    }
  },
  "features": [
    {
      "identifier": "rec_id",
      "ignored": true
    },
    {
      "identifier": "given_name",
      "format": { "type": "string", "encoding": "utf-8", "maxLength": 64 },
      "hashing": { "ngram": 2, "strategy": {"numBits": 300}, "hash": {"type": ↪
↪"doubleHash"} }
    },
    {
      "identifier": "surname",
      "format": { "type": "string", "encoding": "utf-8", "maxLength": 64 },
      "hashing": { "ngram": 2, "strategy": {"numBits": 300}, "hash": {"type": ↪
↪"doubleHash"} }
    },
    {
      "identifier": "street_number",
      "format": { "type": "integer" },
      "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 300}, ↪
↪"missingValue": {"sentinel": ""} }
    },
    {
      "identifier": "address_1",
      "format": { "type": "string", "encoding": "utf-8" },
      "hashing": { "ngram": 2, "strategy": {"numBits": 300} }
    },
    {
      "identifier": "address_2",
```

(continues on next page)

(continued from previous page)

```

    "format": { "type": "string", "encoding": "utf-8" },
    "hashing": { "ngram": 2, "strategy": {"numBits": 300} }
  },
  {
    "identifier": "suburb",
    "format": { "type": "string", "encoding": "utf-8" },
    "hashing": { "ngram": 2, "strategy": {"numBits": 300} }
  },
  {
    "identifier": "postcode",
    "format": { "type": "integer", "minimum": 100, "maximum": 9999 },
    "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 300} }
  },
  {
    "identifier": "state",
    "format": { "type": "string", "encoding": "utf-8", "maxLength": 3 },
    "hashing": { "ngram": 2, "strategy": {"numBits": 300} }
  },
  {
    "identifier": "date_of_birth",
    "format": { "type": "integer" },
    "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 300},
    ↪ "missingValue": {"sentinel": ""} }
  },
  {
    "identifier": "soc_sec_id",
    "ignored": true
  }
]
}

```

Validate the schema

The command line tool can check that the linkage schema is valid:

```
[7]: !clkutil validate-schema _static/febrl_schema_v2_overweight.json
schema is valid
```

Hash the data

We can now hash our Personally Identifiable Information (PII) data from the CSV file using our defined linkage schema. We must provide two *secret keys* to this command - these keys have to be used by both parties hashing data. For this toy example we will use the keys 'key1' and 'key2', for real data, make sure that the keys contain enough entropy, as knowledge of these keys is sufficient to reconstruct the PII information from a CLK! Also, **do not share these keys with anyone, except the other participating party.**

```
[8]: !clkutil hash PII_a.csv key1 key2 _static/febrl_schema_v2_overweight.json clks_a.json
generating CLKs: 100%| 5.00k/5.00k [00:01<00:00, 1.27kclk/s, mean=949, std=9.82]
CLK data written to clks_a.json
```

Inspect the output

clkhsh has hashed the PII, creating a Cryptographic Longterm Key for each entity. The stats output shows that the mean popcount (number of bits set) is quite high (949 out of 1024) which can effect accuracy.

To reduce the popcount you can modify the individual ‘*numBits*’ values for the different fields. It allows to tune the contribution of a column to the CLK. This can be used to de-emphasise columns which are less suitable for linkage (e.g. information that changes frequently).

First, we will reduce the value of *numBits* for each feature.

```
[9]: %cat _static/febrl_schema_v2_reduced.json
```

```
{
  "version": 2,
  "clkConfig": {
    "l": 1024,
    "kdf": {
      "type": "HKDF",
      "hash": "SHA256",
      "info": "c2NoZWlhX2V4YW1wbGU=",
      "salt": "SCbL2zHNnmsckfzchsNkZY9XoHk96P/
↪G5nUBrM7ybym1EFsMV6PAeDZCNp3rfNUPCtLDMOGQHg4pCQpfhiHCyA==",
      "keySize": 64
    }
  },
  "features": [
    {
      "identifier": "rec_id",
      "ignored": true
    },
    {
      "identifier": "given_name",
      "format": { "type": "string", "encoding": "utf-8", "maxLength": 64 },
      "hashing": { "ngram": 2, "strategy": {"numBits": 200}, "hash": {"type": ↪
↪"doubleHash"} }
    },
    {
      "identifier": "surname",
      "format": { "type": "string", "encoding": "utf-8", "maxLength": 64 },
      "hashing": { "ngram": 2, "strategy": {"numBits": 200}, "hash": {"type": ↪
↪"doubleHash"} }
    },
    {
      "identifier": "street_number",
      "format": { "type": "integer" },
      "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 200}, ↪
↪"missingValue": {"sentinel": ""} }
    },
    {
      "identifier": "address_1",
      "format": { "type": "string", "encoding": "utf-8" },
      "hashing": { "ngram": 2, "strategy": {"numBits": 200} }
    },
    {
      "identifier": "address_2",
      "format": { "type": "string", "encoding": "utf-8" },
      "hashing": { "ngram": 2, "strategy": {"numBits": 200} }
    }
  ],
}
```

(continues on next page)

(continued from previous page)

```

{
  "identifier": "suburb",
  "format": { "type": "string", "encoding": "utf-8" },
  "hashing": { "ngram": 2, "strategy": {"numBits": 200} }
},
{
  "identifier": "postcode",
  "format": { "type": "integer", "minimum": 100, "maximum": 9999 },
  "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 200} }
},
{
  "identifier": "state",
  "format": { "type": "string", "encoding": "utf-8", "maxLength": 3 },
  "hashing": { "ngram": 2, "strategy": {"numBits": 200} }
},
{
  "identifier": "date_of_birth",
  "format": { "type": "integer" },
  "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 200},
  ↪ "missingValue": {"sentinel": ""} }
},
{
  "identifier": "soc_sec_id",
  "ignored": true
}
]
}

```

```

[10]: !clkutil hash PII_a.csv key1 key2 _static/febrl_schema_v2_reduced.json clks_a.json
generating CLKs: 100%|| 5.00k/5.00k [00:00<00:00, 1.74kclk/s, mean=843, std=13.8]
CLK data written to clks_a.json

```

And now we will modify the numBits values again, this time de-emphasising the contribution of the address related columns.

```

[11]: %cat _static/febrl_schema_v2_final.json
{
  "version": 2,
  "clkConfig": {
    "l": 1024,
    "kdf": {
      "type": "HKDF",
      "hash": "SHA256",
      "info": "c2NoZWlhX2V4YW1wbGU=",
      "salt": "SCbL2zHnmsckfzchsNkZY9XoHk96P/
  ↪G5nUBrM7ybymlEFsMV6PAeDZCNp3rfNUPCtLDMOGQHg4pCQpfhiHCyA==",
      "keySize": 64
    }
  },
  "features": [
    {
      "identifier": "rec_id",
      "ignored": true
    },
    {

```

(continues on next page)

(continued from previous page)

```

        "identifier": "given_name",
        "format": { "type": "string", "encoding": "utf-8", "maxLength": 64 },
        "hashing": { "ngram": 2, "strategy": {"numBits": 200}, "hash": {"type": ↵
↵ "doubleHash"} }
    },
    {
        "identifier": "surname",
        "format": { "type": "string", "encoding": "utf-8", "maxLength": 64 },
        "hashing": { "ngram": 2, "strategy": {"numBits": 200}, "hash": {"type": ↵
↵ "doubleHash"} }
    },
    {
        "identifier": "street_number",
        "format": { "type": "integer" },
        "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 100}, ↵
↵ "missingValue": {"sentinel": ""} }
    },
    {
        "identifier": "address_1",
        "format": { "type": "string", "encoding": "utf-8" },
        "hashing": { "ngram": 2, "strategy": {"numBits": 100} }
    },
    {
        "identifier": "address_2",
        "format": { "type": "string", "encoding": "utf-8" },
        "hashing": { "ngram": 2, "strategy": {"numBits": 100} }
    },
    {
        "identifier": "suburb",
        "format": { "type": "string", "encoding": "utf-8" },
        "hashing": { "ngram": 2, "strategy": {"numBits": 100} }
    },
    {
        "identifier": "postcode",
        "format": { "type": "integer", "minimum": 50, "maximum": 9999 },
        "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 100} }
    },
    {
        "identifier": "state",
        "format": { "type": "string", "encoding": "utf-8"},
        "hashing": {"ngram": 2, "positional": true, "strategy": {"numBits": 100}, ↵
↵ "missingValue": {"sentinel": ""}
    }
    },
    {
        "identifier": "date_of_birth",
        "format": { "type": "integer" },
        "hashing": { "ngram": 1, "positional": true, "strategy": {"numBits": 200}, ↵
↵ "missingValue": {"sentinel": ""} }
    },
    {
        "identifier": "soc_sec_id",
        "ignored": true
    }
]
}

```

```
[12]: !clkutil hash PII_a.csv key1 key2 _static/febrl_schema_v2_final.json clks_a.json
generating CLKs: 100%|| 5.00k/5.00k [00:00<00:00, 7.08kclk/s, mean=705, std=16]
CLK data written to clks_a.json
```

Great, now approximately half the bits are set in each CLK.

Each CLK is serialized in a JSON friendly base64 format:

```
[13]: # If you have jq tool installed:
#!jq .clks[0] clks_a.json

import json
json.load(open('clks_a.json'))['clks'][0]

[13]: 'unsZ/W7D35s8q759bf77155ean+p8fq96fzf9u9bnXf3rX2gGfntPvR2/tOd314aOvuv/
↪97z+lrY8st+fP8PYVd9/KjZN6rMx+T/O6r/v/
↪Hdvt1f1at2+f+Xe53iX94f9988b3mhTsIQbf+7Xr3Sff71fuze9k3sX++db4d73v0='
```

Hash data set B

Now we hash the second dataset using the same keys and same schema.

```
[14]: dfB.to_csv('PII_b.csv')

!clkutil hash PII_b.csv key1 key2 _static/febrl_schema_v2_final.json clks_b.json

generating CLKs: 100%|| 5.00k/5.00k [00:00<00:00, 7.11kclk/s, mean=703, std=19.4]
CLK data written to clks_b.json
```

Find matches between the two sets of CLKs

We have generated two sets of CLKs which represent entity information in a privacy-preserving way. The more similar two CLKs are, the more likely it is that they represent the same entity.

For this task we will use the entity service, which is provided by Data61. The necessary steps are as follows: - The analyst creates a new project with the output type 'mapping'. They will receive a set of credentials from the server. - The analyst then distributes the `update_tokens` to the participating data providers. - The data providers then individually upload their respective CLKs. - The analyst can create *runs* with various thresholds (and other settings) - After the entity service successfully computed the mapping, it can be accessed by providing the `result_token`

First we check the status of an entity service:

```
[15]: SERVER = 'https://testing.es.data61.xyz'

!clkutil status --server={SERVER}

{"project_count": 772, "rate": 2083409, "status": "ok"}
```

The analyst creates a new project on the entity service by providing the hashing schema and result type. The server returns a set of credentials which provide access to the further steps for project.

```
[16]: !clkutil create-project --server={SERVER} --schema _static/febrl_schema_v2_final.json_
↪--output credentials.json --type "mapping" --name "tutorial"

Project created
```

The returned credentials contain a - `project_id`, which identifies the project - `result_token`, which gives access to the mapping result, once computed - `upload_tokens`, one for each provider, allows uploading CLKs.

```
[17]: credentials = json.load(open('credentials.json', 'rt'))
print(json.dumps(credentials, indent=4))

{
  "project_id": "90fbe7e6de6e085afffdd8403722d44ffc18756b9e204c8f",
  "result_token": "c2b0b8e133dae678eaacalcdb32f28493cb7a7ff5728ed79",
  "update_tokens": [
    "a451f0b5a3cc4829701f236deaa036955bed9aceb34f7242",
    "710b1e1f7485e184eb1ffac1d97d41e3fa2a845b2d4ea70d"
  ]
}
```

Uploading the CLKs to the entity service

Each party individually uploads its respective CLKs to the entity service. They need to provide the `resource_id`, which identifies the correct mapping, and an `update_token`.

```
[18]: !clkutil upload \
      --project="{credentials['project_id']}" \
      --apikey="{credentials['update_tokens'][0]}" \
      --output "upload_a.json" \
      --server="{SERVER}" \
      "clks_a.json"

!clkutil upload \
      --project="{credentials['project_id']}" \
      --apikey="{credentials['update_tokens'][1]}" \
      --output "upload_b.json" \
      --server="{SERVER}" \
      "clks_b.json"
```

Now that the CLK data has been uploaded the analyst can create one or more *runs*. Here we will start by calculating a mapping with a threshold of 0.9:

```
[19]: !clkutil create --verbose \
      --server="{SERVER}" \
      --output "run_info.json" \
      --threshold=0.9 \
      --project="{credentials['project_id']}" \
      --apikey="{credentials['result_token']}" \
      --name="CLI tutorial run A"

Entity Matching Server: https://testing.es.data61.xyz
```

```
[20]: run_info = json.load(open('run_info.json', 'rt'))
run_info

[20]: {'name': 'CLI tutorial run A',
      'notes': 'Run created by clckhash 0.13.0',
      'run_id': '1dd059e252d2f31c7e9aeacf2c5e69a59a0a5592add1dbd3',
      'threshold': 0.9}
```

Results

Now after some delay (depending on the size) we can fetch the results. This can be done with clkutil:

```
[21]: !clkutil results --watch \
      --project="{credentials['project_id']}" \
      --apikey="{credentials['result_token']}" \
      --run="{run_info['run_id']}" \
      --server="{SERVER}" \
      --output results.txt
```

```
State: running
Stage (2/3): compute similarity scores
State: completed
Stage (3/3): compute output
State: completed
Stage (3/3): compute output
Downloading result
Received result
```

```
[22]: with open('results.txt') as f:
      str_mapping = json.load(f)['mapping']

mapping = {int(k): int(v) for k,v in str_mapping.items()}
print('The service linked {} entities.'.format(len(mapping)))

The service linked 4001 entities.
```

Let's investigate some of those matches and the overall matching quality. In this case we have the ground truth so we can compute the precision, recall, and accuracy.

```
[23]: with open('PII_a.csv', 'rt') as f:
      a_raw = f.readlines()
      with open('PII_b.csv', 'rt') as f:
          b_raw = f.readlines()

num_entities = len(b_raw) - 1

def describe_accuracy(mapping, show_examples=False):
    if show_examples:
        print('idx_a, idx_b,      rec_id_a,      rec_id_b')
        print('-----')
        for a_i in range(10):
            if a_i in mapping:
                a_data = a_raw[a_i + 1].split(',')
                b_data = b_raw[mapping[a_i] + 1].split(',')
                print('{:3}, {:6}, {:>15}, {:>15}'.format(a_i+1, mapping[a_i]+1, a_
→data[0], b_data[0]))
            print('-----')

    TP = 0; FP = 0; TN = 0; FN = 0
    for a_i in range(num_entities):
        if a_i in mapping:
            if a_raw[a_i + 1].split(',')[0].split('-')[1] == b_raw[mapping[a_i] + 1].
→split(',')[0].split('-')[1]:
                TP += 1
            else:
                FP += 1
```

(continues on next page)

(continued from previous page)

```

        # as we only report one mapping for each element in PII_a,
        # then a wrong mapping is not only a false positive, but
        # also a false negative, as we won't report the true mapping.
        FN += 1

    else:
        FN += 1 # every element in PII_a has a partner in PII_b

print('Precision: {:.2f}, Recall: {:.2f}, Accuracy: {:.2f}'.format(
    TP / (TP + FP),
    TP / (TP + FN),
    (TP + TN) / (TP + TN + FP + FN)))

```

```
[24]: describe_accuracy(mapping, True)
```

```

idx_a, idx_b,      rec_id_a,      rec_id_b
-----
 2,    2751,    rec-1016-org,    rec-1016-dup-0
 3,    4657,    rec-4405-org,    rec-4405-dup-0
 4,    4120,    rec-1288-org,    rec-1288-dup-0
 5,    3307,    rec-3585-org,    rec-3585-dup-0
 6,    2306,    rec-298-org,     rec-298-dup-0
 7,    3945,    rec-1985-org,    rec-1985-dup-0
 8,     993,    rec-2404-org,    rec-2404-dup-0
 9,    4613,    rec-1473-org,    rec-1473-dup-0
10,    3630,    rec-453-org,     rec-453-dup-0
-----
Precision: 1.00, Recall: 0.80, Accuracy: 0.80

```

Precision tells us about how many of the found matches are actual matches. The score of 1.0 means that we did perfectly in this respect, however, **recall**, the measure of how many of the actual matches were correctly identified, is quite low with only 80%.

Let's go back and create another mapping with a threshold value of 0.8.

```

[25]: !clkutil create --verbose \
      --server="{SERVER}" \
      --output "run_info.json" \
      --threshold=0.8 \
      --project="{credentials['project_id']}" \
      --apikey="{credentials['result_token']}" \
      --name="CLI tutorial run B"

run_info = json.load(open('run_info.json', 'rt'))

Entity Matching Server: https://testing.es.data61.xyz

```

```

[26]: !clkutil results --watch \
      --project="{credentials['project_id']}" \
      --apikey="{credentials['result_token']}" \
      --run="{run_info['run_id']}" \
      --server="{SERVER}" \
      --output results.txt

State: running
Stage (2/3): compute similarity scores
State: running
Stage (2/3): compute similarity scores

```

(continues on next page)

(continued from previous page)

```
State: completed
Stage (3/3): compute output
Downloading result
Received result
```

```
[27]: with open('results.txt') as f:
        str_mapping = json.load(f)['mapping']

mapping = {int(k): int(v) for k,v in str_mapping.items()}

print('The service linked {} entities.'.format(len(mapping)))
describe_accuracy(mapping)
```

```
The service linked 4975 entities.
Precision: 1.00, Recall: 0.99, Accuracy: 0.99
```

Great, for this threshold value we get a precision of 100% and a recall of 99%.

The explanation is that when the information about an entity differs slightly in the two datasets (e.g. spelling errors, abbreviations, missing values, ...) then the corresponding CLKs will differ in some number of bits as well. For the datasets in this tutorial the perturbations are such that only 80% of the derived CLK pairs overlap more than 90% (the first threshold). Whereas 99% of all matching pairs overlap more than 80%.

If we keep reducing the threshold value, then we will start to observe mistakes in the found matches – the precision decreases (if an entry in dataset A has no match in dataset B, but we keep reducing the threshold, eventually a comparison with an entry in B will be above the threshold leading to a false match). But at the same time the recall value will keep increasing for a while, as a lower threshold allows for more of the actual matches to be found. However, as our example dataset only contains matches (every entry in A has a match in B), this phenomenon cannot be observed. With the threshold 0.75 we identify all matches correctly (at the cost of a longer execution time).

```
[28]: !clkutil create --verbose \
        --server="{SERVER}" \
        --output "run_info.json" \
        --threshold=0.75 \
        --project="{credentials['project_id']}" \
        --apikey="{credentials['result_token']}" \
        --name="CLI tutorial run B"

run_info = json.load(open('run_info.json', 'rt'))

Entity Matching Server: https://testing.es.data61.xyz
```

```
[29]: !clkutil results --watch \
        --project="{credentials['project_id']}" \
        --apikey="{credentials['result_token']}" \
        --run="{run_info['run_id']}" \
        --server="{SERVER}" \
        --output results.txt
```

```
State: running
Stage (2/3): compute similarity scores
State: running
Stage (2/3): compute similarity scores
State: running
Stage (3/3): compute output
State: completed
Stage (3/3): compute output
```

(continues on next page)

(continued from previous page)

```
Downloading result
Received result
```

```
[30]: with open('results.txt') as f:
        str_mapping = json.load(f)['mapping']

mapping = {int(k): int(v) for k,v in str_mapping.items()}

print('The service linked {} entities.'.format(len(mapping)))
describe_accuracy(mapping)
```

```
The service linked 4995 entities.
Precision: 1.00, Recall: 1.00, Accuracy: 1.00
```

It is important to choose an appropriate threshold for the amount of perturbations present in the data.

Feel free to go back to the CLK generation and experiment on how different setting will affect the matching quality.

Finally to remove the uploaded CLK data from the service delete the project:

```
[31]: # TODO

#!clkutil delete-project --project="{credentials['project_id']}" \
#      --apikey="{credentials['result_token']}" \
#      --run="{run_info['run_id']}" \
#      --server="{SERVER}"
```

1.2 Command Line Tool

clkh hash includes a command line tool which can be used to interact without writing Python code. The primary use case is to encode personally identifiable data from a csv into Cryptographic Longterm Keys.

The command line tool can be accessed in two equivalent ways:

- Using the `clkutil` script which gets added to your path during installation.
- directly running the python module with `python -m clkh hash`.

A list of valid commands can be listed with the `--help` argument:

```
$ clkutil --help
Traceback (most recent call last):
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkh hash/envs/v0.13.0/bin/
↳ clkutil", line 11, in <module>
    load_entry_point('clkh hash==0.13.0', 'console_scripts', 'clkutil')()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkh hash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 489, in load_entry_point
    return get_distribution(dist).load_entry_point(group, name)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkh hash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 2843, in load_entry_point
    return ep.load()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkh hash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 2434, in load
    return self.resolve()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkh hash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 2440, in resolve
```

(continues on next page)

(continued from previous page)

```

    module = __import__(self.module_name, fromlist=['__name__'], level=0)
    File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/clkhsh-0.13.0-py3.7.egg/clkhsh/__init__.py", line 3, in
↳<module>
        from . import bloomfilter, field_formats, key_derivation, schema, randomnames,
↳describe
    File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/clkhsh-0.13.0-py3.7.egg/clkhsh/bloomfilter.py", line 11,
↳in <module>
        from typing import Callable, Iterable, List, Optional, Sequence, Text, Tuple
    File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/typing.py", line 1357, in <module>
        class Callable(extra=collections_abc.Callable, metaclass=CallableMeta):
    File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/typing.py", line 1005, in __new__
        self._abc_registry = extra._abc_registry
AttributeError: type object 'Callable' has no attribute '_abc_registry'

```

1.2.1 Command specific help

The `clkutil` tool has help pages for all commands built in - simply append `--help` to the command.

1.2.2 Hashing

The command line tool `clkutil` can be used to hash a csv file of personally identifiable information. The tool needs to be provided with keys and a [Linkage Schema](#); it will output a file containing json serialized hashes.

```

$ clkutil hash --help
Traceback (most recent call last):
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/bin/
↳clkutil", line 11, in <module>
    load_entry_point('clkhsh==0.13.0', 'console_scripts', 'clkutil')()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/pkg_resources/__init__.py", line 489, in load_entry_point
    return get_distribution(dist).load_entry_point(group, name)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/pkg_resources/__init__.py", line 2843, in load_entry_point
    return ep.load()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/pkg_resources/__init__.py", line 2434, in load
    return self.resolve()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/pkg_resources/__init__.py", line 2440, in resolve
    module = __import__(self.module_name, fromlist=['__name__'], level=0)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/clkhsh-0.13.0-py3.7.egg/clkhsh/__init__.py", line 3, in
↳<module>
        from . import bloomfilter, field_formats, key_derivation, schema, randomnames,
↳describe
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/clkhsh-0.13.0-py3.7.egg/clkhsh/bloomfilter.py", line 11,
↳in <module>
        from typing import Callable, Iterable, List, Optional, Sequence, Text, Tuple
  File "/home/docs/checkouts/readthedocs.org/user_builds/clkhsh/envs/v0.13.0/lib/
↳python3.7/site-packages/typing.py", line 1357, in <module>

```

(continues on next page)

(continued from previous page)

```
class Callable(extra=collections_abc.Callable, metaclass=CallableMeta):
    File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
    ↳python3.7/site-packages/typing.py", line 1005, in __new__
        self._abc_registry = extra._abc_registry
AttributeError: type object 'Callable' has no attribute '_abc_registry'
```

Example

Assume a csv (fake-pii.csv) contains rows like the following:

```
0,Libby Slemmer,1933/09/13,F
1,Garold Staten,1928/11/23,M
2,Yaritza Edman,1972/11/30,F
```

It can be hashed using clkutil with:

```
$ clkutil hash --schema simple-schema.json fake-pii.csv horse staple clk.json
```

Where:

- horse staple is the two part secret key that both participants will use to hash their data.
- simple-schema.json is a *Linkage Schema* describing how to hash the csv. E.g, ignore the first column, use bigram tokens of the name, use positional unigrams of the date of birth etc.
- clk.json is the output file.

1.2.3 Describing

Users can inspect the distribution of the number of bits set in CLKs by using the describe command.

```
$ clkutil describe --help
Traceback (most recent call last):
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/bin/
  ↳clkutil", line 11, in <module>
    load_entry_point('clckhash==0.13.0', 'console_scripts', 'clkutil')()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
  ↳python3.7/site-packages/pkg_resources/__init__.py", line 489, in load_entry_point
    return get_distribution(dist).load_entry_point(group, name)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
  ↳python3.7/site-packages/pkg_resources/__init__.py", line 2843, in load_entry_point
    return ep.load()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
  ↳python3.7/site-packages/pkg_resources/__init__.py", line 2434, in load
    return self.resolve()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
  ↳python3.7/site-packages/pkg_resources/__init__.py", line 2440, in resolve
    module = __import__(self.module_name, fromlist=['__name__'], level=0)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
  ↳python3.7/site-packages/clckhash-0.13.0-py3.7.egg/clckhash/__init__.py", line 3, in
  ↳<module>
    from . import bloomfilter, field_formats, key_derivation, schema, randomnames,
  ↳describe
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
  ↳python3.7/site-packages/clckhash-0.13.0-py3.7.egg/clckhash/bloomfilter.py", line 11,
  ↳in <module>
```

(continues on next page)

(continued from previous page)

```

from typing import Callable, Iterable, List, Optional, Sequence, Text, Tuple
File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
python3.7/site-packages/typing.py", line 1357, in <module>
    class Callable(extra=collections_abc.Callable, metaclass=CallableMeta):
File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
python3.7/site-packages/typing.py", line 1005, in __new__
    self._abc_registry = extra._abc_registry
AttributeError: type object 'Callable' has no attribute '_abc_registry'

```

Example

```

$ clckutil describe example_clks_a.json

339|                                     oo
321|                                     ooo
303|                                     ooo
285|                                     ooo o
268|                                     oooooo
250|                                     oooooooooo
232|                                     oooooooooo
214|                                     oooooooooo
196|                                     o oooooooooo o
179|                                     o ooooooooooooo
161|                                     ooooooooooooooooo
143|                                     ooooooooooooooooo
125|                                     ooooooooooooooooo
107|                                     ooooooooooooooooo
 90|                                     ooooooooooooooooo
 72|                                     ooooooooooooooooo
 54|                                     ooooooooooooooooo
 36|                                     ooooooooooooooooo
 18|                                     ooooooooooooooooooooo
  1| o  o  oooooooooooooooooooooooooooooooooooooooooooooo oo
-----
 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 7 7 7 7
 1 2 3 4 5 6 7 9 0 1 2 3 4 5 7 8 9 0 1 2 3 5 6 7 8 9 0 1 3 4
 0 1 2 4 5 7 8 0 1 2 4 5 7 8 0 1 2 4 5 7 8 0 1 2 4 5 7 8 0 1
  . . . . . . . . . . . . . . . . . . . . . . . . . . .
 4 8 3 7 1 6 0 4 9 3 7 2 6 0 5 9 3 8 2 6 1 5 9 4 8 2 7 1 5

-----
|           Summary           |
-----
| observations: 5000 |
| min value: 410.000000 |
| mean : 601.571600 |
| max value: 753.000000 |
-----

```

Note: It is an indication of problems in the hashing if the distribution is skewed towards no bits set or all bits set. Consult the [Tutorial for CLI tool clckhash](#) for further details.

1.2.4 Data Generation

The command line tool has a generate command for generating fake pii data.

```
$ clkutil generate --help
Traceback (most recent call last):
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/bin/
↳ clkutil", line 11, in <module>
    load_entry_point('clckhash==0.13.0', 'console_scripts', 'clkutil')()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 489, in load_entry_point
    return get_distribution(dist).load_entry_point(group, name)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 2843, in load_entry_point
    return ep.load()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 2434, in load
    return self.resolve()
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/pkg_resources/__init__.py", line 2440, in resolve
    module = __import__(self.module_name, fromlist=['__name__'], level=0)
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/clckhash-0.13.0-py3.7.egg/clckhash/__init__.py", line 3, in
↳ <module>
    from . import bloomfilter, field_formats, key_derivation, schema, randomnames,
↳ describe
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/clckhash-0.13.0-py3.7.egg/clckhash/bloomfilter.py", line 11,
↳ in <module>
    from typing import Callable, Iterable, List, Optional, Sequence, Text, Tuple
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/typing.py", line 1357, in <module>
    class Callable(extra=collections_abc.Callable, metaclass=CallableMeta):
  File "/home/docs/checkouts/readthedocs.org/user_builds/clckhash/envs/v0.13.0/lib/
↳ python3.7/site-packages/typing.py", line 1005, in __new__
    self.__abc_registry = extra.__abc_registry
AttributeError: type object 'Callable' has no attribute '__abc_registry'
```

```
$ clkutil generate 1000 fake-pii-out.csv
$ head -n 4 fake-pii-out.csv
INDEX,NAME freetext,DOB YYYY/MM/DD,GENDER M or F
0,Libby Slemmer,1933/09/13,F
1,Garold Staten,1928/11/23,M
2,Yaritza Edman,1972/11/30,F
```

A corresponding hashing schema can be generated as well:

```
$ clkutil generate-default-schema schema.json
$ cat schema.json
{
  "version": 1,
  "clkConfig": {
    "l": 1024,
    "k": 30,
    "hash": {
      "type": "doubleHash"
    }
  },
```

(continues on next page)

(continued from previous page)

```

    "kdf": {
        "type": "HKDF",
        "hash": "SHA256",
        "salt": "SCbL2zHNnmsckfzchsNkZY9XoHk96P/
→G5nUBrM7ybymlEFsMV6PAeDZCNp3rfNUPCtLDMOGQHg4pCQpfhiHCyA==",
        "info": "c2NoZWlhX2V4YW1wbGU=",
        "keySize": 64
    }
},
"features": [
    {
        "identifier": "INDEX",
        "format": {
            "type": "integer"
        },
        "hashing": {
            "ngram": 1,
            "weight": 0
        }
    },
    {
        "identifier": "NAME freetext",
        "format": {
            "type": "string",
            "encoding": "utf-8",
            "case": "mixed",
            "minLength": 3
        },
        "hashing": {
            "ngram": 2,
            "weight": 0.5
        }
    },
    {
        "identifier": "DOB YYYY/MM/DD",
        "format": {
            "type": "string",
            "encoding": "ascii",
            "description": "Numbers separated by slashes, in the year, month, day order",
            "pattern": "(?:\\d\\d\\d\\d/\\d\\d/\\d\\d)\\Z"
        },
        "hashing": {
            "ngram": 1,
            "positional": true
        }
    },
    {
        "identifier": "GENDER M or F",
        "format": {
            "type": "enum",
            "values": ["M", "F"]
        },
        "hashing": {
            "ngram": 1,
            "weight": 2
        }
    }
]

```

(continues on next page)

(continued from previous page)

```
]
}
```

1.2.5 Benchmark

A quick hashing benchmark can be carried out to determine the rate at which the current machine can generate 10000 clks from a simple schema (data as generated [above](#)):

```
python -m clkhash.cli benchmark
generating CLKs: 100%          10.0K/10.0K [00:01<00:00, 7.72Kclk/s, mean=521,
↪std=34.7]
10000 hashes in 1.350489 seconds. 7.40 KH/s
```

As a rule of thumb a single modern core will hash around 1M entities in about 20 minutes.

Note: Hashing speed is effected by the number of features and the corresponding schema. Thus these numbers will, in general, not be a good predictor for the performance of a specific use-case.

The output shows a running mean and std deviation of the generated clks' popcounts. This can be used as a basic sanity check - ensure the CLK's popcount is not around 0 or 1024.

1.2.6 Interaction with Entity Service

There are several commands that interact with a REST api for carrying out privacy preserving linking. These commands are:

- status
- create-project
- create
- upload
- results

See also the [Tutorial for CLI](#).

1.3 Linkage Schema

As CLKs are usually used for privacy preserving linkage, it is important that participating organisations agree on how raw personally identifiable information is encoded to create the CLKs. The linkage schema allows putting more emphasis on particular features and provides a basic level of data validation.

We call the configuration of how to create CLKs a *linkage schema*. The organisations agree on a linkage schema to ensure that their respective CLKs have been created in the same way.

This aims to be an open standard such that different client implementations could take the schema and create identical CLKs given the same data (and secret keys).

The linkage schema is a detailed description of exactly how to carry out the encoding operation, along with any configuration for the low level hashing itself.

The format of the linkage schema is defined in a separate [JSON Schema](#) specification document - [schemas/v2.json](#).

Earlier versions of the linkage schema will continue to work, internally they are converted to the latest version (currently v2).

1.3.1 Basic Structure

A linkage schema consists of three parts:

- *version*, contains the version number of the hashing schema.
- *clkConfig*, CLK wide configuration, independent of features.
- *features*, an array of configuration specific to individual features.

1.3.2 Example Schema

```
{
  "version": 2,
  "clkConfig": {
    "l": 1024,
    "kdf": {
      "type": "HKDF",
      "hash": "SHA256",
      "salt": "SCbL2zHNmsckfzchsNkZY9XoHk96P/
↪G5nUBrM7ybymlEFsMV6PAeDZCNp3rfNUPCtLDMOGQH4pCQpfhiHCyA==",
      "info": "",
      "keySize": 64
    }
  },
  "features": [
    {
      "identifier": "INDEX",
      "ignored": true
    },
    {
      "identifier": "NAME freetext",
      "format": {
        "type": "string",
        "encoding": "utf-8",
        "case": "mixed",
        "minLength": 3
      },
      "hashing": {
        "ngram": 2,
        "numBits": 100,
        "hash": {"type": "doubleHash"}
      }
    },
    {
      "identifier": "DOB YYYY/MM/DD",
      "format": {
        "type": "date",
        "description": "Numbers separated by slashes, in the year, month, day order",
        "format": "%Y/%m/%d"
      }
    }
  ]
}
```

(continues on next page)

(continued from previous page)

```
    "hashing": {
      "ngram": 1,
      "positional": true,
      "numBits": 200,
      "hash": {"type": "doubleHash"}
    },
  },
  {
    "identifier": "GENDER M or F",
    "format": {
      "type": "enum",
      "values": ["M", "F"]
    },
    "hashing": {
      "ngram": 1,
      "numBits": 400,
      "hash": {"type": "doubleHash"}
    }
  }
]
}
```

A more advanced example can be found [here](#).

1.3.3 Schema Components

Version

Integer value which describes the version of the hashing schema.

clkConfig

Describes the general construction of the CLK.

name	type	op- tional	description
l	inte- ger	no	the length of the CLK in bits
kdf	<i>KDF</i>	no	defines the key derivation function used to generate individual secrets for each feature derived from the master secret
xor- Folds	inte- ger	yes	number of XOR folds (as proposed in [Schnell2016]).

KDF

We currently only support HKDF (for a basic description, see <https://en.wikipedia.org/wiki/HKDF>).

name	type	optional	description
type	string	no	must be set to “HKDF”
hash	enum	yes	hash function used by HKDF, either “SHA256” or “SHA512”
salt	string	yes	base64 encoded bytes
info	string	yes	base64 encoded bytes
keySize	integer	yes	size of the generated keys in bytes

features

A feature is either described by a *featureConfig*, or alternatively, it can be ignored by the clkh hash library by defining a *ignoreFeature* section.

ignoreFeature

If defined, then clkh hash will ignore this feature.

name	type	optional	description
identifier	string	no	the name of the feature
ignored	boolean	no	has to be set to “True”
description	string	yes	free text, ignored by clkh hash

featureConfig

Each feature is configured by:

- identifier, the human readable name. E.g. "First Name".
- description, a human readable description of this feature.
- format, describes the expected format of the values of this feature
- *hashing*, configures the hashing

name	type	optional	description
identifier	string	no	the name of the feature
description	string	yes	free text, ignored by clkh hash
hashing	<i>hashingConfig</i>	no	configures feature specific hashing parameters
format	one of: <i>textFormat</i> , <i>textPatternFormat</i> , <i>numberFormat</i> , <i>dateFormat</i> , <i>enumFormat</i>	no	describes the expected format of the feature values

hashingConfig

name	type	optional	description
ngram	integer	no	specifies the n in n-gram (the tokenization of the input values).
strategy	<i>strategy</i>	no	the strategy for assigning bits to the encoding.
positional	boolean	yes	adds the position to the n-grams. String “222” would be tokenized (as uni-grams) to “1 2”, “2 2”, “3 2”
missing-Value	<i>missing-Value</i>	yes	allows to define how missing values are handled

strategy

An object where either `numBits` or `k` is defined.

name	type	optional	description
k	integer	yes	max number of indices per n-gram
numBits	integer	yes	max number of indices per feature

Hash

Describes and configures the hash that is used to encode the n-grams.

Choose one of:

- *double hash*, as described in [Schnell2011].

name	type	optional	description
type	string	no	must be set to “doubleHash”
prevent_singularity	boolean	yes	see discussion in https://github.com/data61/clkhhash/issues/33

- *blake hash* (default)

name	type	optional	description
type	string	no	must be set to “blakeHash”

missingValue

Data sets are not always complete – they can contain missing values. If specified, then `clkhhash` will not check the format for these missing values, and will optionally replace the `sentinel` with the `replaceWith` value.

name	type	optional	description
sentinel	string	no	the sentinel value indicates missing data, e.g. ‘Null’, ‘N/A’, ‘’, ...
replaceWith	string	yes	specifies the value <code>clkhhash</code> should use instead of the sentinel value.

textFormat

name	type	optional	description
type	string	no	has to be “string”
encoding	enum	yes	one of “ascii”, “utf-8”, “utf-16”, “utf-32”. Default is “utf-8”.
case	enum	yes	one of “upper”, “lower”, “mixed”.
minLength	integer	yes	positive integer describing the minimum length of the input string.
maxLength	integer	yes	positive integer describing the maximum length of the input string.
description	string	yes	free text, ignored by clkhsh.

textPatternFormat

name	type	optional	description
type	string	no	has to be “string”
encoding	enum	yes	one of “ascii”, “utf-8”, “utf-16”, “utf-32”. Default is “utf-8”.
pattern	string	no	a regular expression describing the input format.
description	string	yes	free text, ignored by clkhsh.

numberFormat

name	type	optional	description
type	string	no	has to be “integer”
minimum	integer	yes	integer describing the lower bound of the input values.
maximum	integer	yes	integer describing the upper bound of the input values.
description	string	yes	free text, ignored by clkhsh.

dateFormat

A date is described by an ISO C89 compatible `strptime()` format string. For example, the format string for the internet date format as described in rfc3339, would be ‘%Y-%m-%d’. The clkhsh library will convert the given date to the ‘%Y%m%d’ representation for hashing, as any fill character like ‘-’ or ‘/’ do not add to the uniqueness of an entity.

name	type	optional	description
type	string	no	has to be “date”
format	string	no	ISO C89 compatible format string, eg: for 1989-11-09 the format is ‘%Y-%m-%d’
description	string	yes	free text, ignored by clkhsh.

The following subset contains the most useful format codes:

directive	meaning	example
%Y	Year with century as a decimal number	1984, 3210, 0001
%y	Year without century, zero-padded	00, 09, 99
%m	Month as a zero-padded decimal number	01, 12
%d	Day of the month, zero-padded	01, 25, 31

enumFormat

name	type	optional	description
type	string	no	has to be “enum”
values	array	no	an array of items of type “string”
description	string	yes	free text, ignored by clkhash.

1.4 Development

1.4.1 API Documentation

Bloom filter

Generate a Bloom filter

`clkhash.bloomfilter.blake_encode_ngrams` (*ngrams*, *keys*, *ks*, *l*, *encoding*)
Computes the encoding of the ngrams using the BLAKE2 hash function.

We deliberately do not use the double hashing scheme as proposed in [Schnell2011], because this would introduce an exploitable structure into the Bloom filter. For more details on the weakness, see [Kroll2015].

In short, the double hashing scheme only allows for l^2 different encodings for any possible n-gram, whereas the use of k different independent hash functions gives you $\sum_{j=1}^k \binom{l}{j}$ combinations.

Our construction

It is advantageous to construct Bloom filters using a family of hash functions with the property of *k-independence* to compute the indices for an entry. This approach minimises the change of collisions.

An informal definition of *k-independence* of a family of hash functions is, that if selecting a function at random from the family, it guarantees that the hash codes of any designated k keys are independent random variables.

Our construction utilises the fact that the output bits of a cryptographic hash function are uniformly distributed, independent, binary random variables (well, at least as close to as possible. See [Kaminsky2011] for an analysis). Thus, slicing the output of a cryptographic hash function into k different slices gives you k independent random variables.

We chose Blake2 as the cryptographic hash function mainly for two reasons:

- it is fast.
- in keyed hashing mode, Blake2 provides MACs with just one hash function call instead of the two calls in the HMAC construction used in the double hashing scheme.

Warning: Please be aware that, although this construction makes the attack of [Kroll2015] infeasible, it is most likely not enough to ensure security. Or in their own words:

However, we think that using independent hash functions alone will not be sufficient to ensure security, since in this case other approaches (maybe related to or at least inspired through work from the area of Frequent Itemset Mining) are promising to detect at least the most frequent atoms automatically.

Parameters

- **ngrams** – list of n-grams to be encoded
- **keys** – secret key for blake2 as bytes
- **ks** – ks[i] is k value to use for ngram[i]
- **l** – length of the output bitarray (has to be a power of 2)
- **encoding** – the encoding to use when turning the ngrams to bytes

Returns bitarray of length l with the bits set which correspond to the encoding of the ngrams

`clkhash.bloomfilter.crypto_bloom_filter(record, tokenizers, schema, keys)`

Computes the composite Bloom filter encoding of a record.

Using the method from <http://www.record-linkage.de/-download=wp-grlc-2011-02.pdf>

Parameters

- **record** – plaintext record tuple. E.g. (index, name, dob, gender)
- **tokenizers** – A list of tokenizers. A tokenizer is a function that returns tokens from a string.
- **schema** – Schema
- **keys** – Keys for the hash functions as a tuple of lists of bytes.

Returns 3-tuple: - bloom filter for record as a bitarray - first element of record (usually an index) - number of bits set in the bloomfilter

`clkhash.bloomfilter.double_hash_encode_ngrams(ngrams, keys, ks, l, encoding)`

Computes the double hash encoding of the ngrams with the given keys.

Using the method from: Schnell, R., Bachteler, T., & Reiher, J. (2011). A Novel Error-Tolerant Anonymous Linking Code. <http://grlc.german-microsimulation.de/wp-content/uploads/2017/05/downloadwp-grlc-2011-02.pdf>

Parameters

- **ngrams** – list of n-grams to be encoded
- **keys** – hmac secret keys for md5 and sha1 as bytes
- **ks** – ks[i] is k value to use for ngram[i]
- **l** – length of the output bitarray
- **encoding** – the encoding to use when turning the ngrams to bytes

Returns bitarray of length l with the bits set which correspond to the encoding of the ngrams

`clkhash.bloomfilter.double_hash_encode_ngrams_non_singular(ngrams, keys, ks, l, encoding)`
computes the double hash encoding of the n-grams with the given keys.

The original construction of [Schnell2011] displays an abnormality for certain inputs:

An n-gram can be encoded into just one bit irrespective of the number of k.

Their construction goes as follows: the k different indices g_i of the Bloom filter for an n-gram x are defined as:

$$g_i(x) = (h_1(x) + ih_2(x)) \mod l$$

with $0 \leq i < k$ and l is the length of the Bloom filter. If the value of the hash of x of the second hash function is a multiple of l , then

$$h_2(x) = 0 \mod l$$

and thus

$$g_i(x) = h_1(x) \mod l,$$

irrespective of the value i . A discussion of this potential flaw can be found [here](#).

Parameters

- **ngrams** – list of n-grams to be encoded
- **keys** – tuple with (key_sha1, key_md5). That is, (hmac secret keys for sha1 as bytes, hmac secret keys for md5 as bytes)
- **ks** – ks[i] is k value to use for ngram[i]
- **l** – length of the output bitarray
- **encoding** – the encoding to use when turning the ngrams to bytes

Returns bitarray of length l with the bits set which correspond to the encoding of the ngrams

`clkhash.bloomfilter.fold_xor(bloomfilter, folds)`

Performs XOR folding on a Bloom filter.

If the length of the original Bloom filter is n and we perform r folds, then the length of the resulting filter is $n / 2^{**r}$.

Parameters

- **bloomfilter** – Bloom filter to fold
- **folds** – number of folds

Returns folded bloom filter

`clkhash.bloomfilter.hashing_function_from_properties(fhp)`

Get the hashing function for this field :param fhp: hashing properties for this field :return: the hashing function

`clkhash.bloomfilter.int_from_bytes()`

Return the integer represented by the given array of bytes.

bytes Holds the array of bytes to convert. The argument must either support the buffer protocol or be an iterable object producing bytes. Bytes and bytearray are examples of built-in objects that support the buffer protocol.

byteorder The byte order used to represent the integer. If byteorder is ‘big’, the most significant byte is at the beginning of the byte array. If byteorder is ‘little’, the most significant byte is at the end of the byte array. To request the native byte order of the host system, use ‘sys.byteorder’ as the byte order value.

signed Indicates whether two’s complement is used to represent the integer.

`clkhash.bloomfilter.stream_bloom_filters(dataset, keys, schema)`

Compute composite Bloom filters (CLKs) for every record in an iterable dataset.

Parameters

- **dataset** – An iterable of indexable records.
- **schema** – An instantiated Schema instance
- **keys** – A tuple of two lists of secret keys used in the HMAC.

Returns Generator yielding bloom filters as 3-tuples

CLK

Generate CLK from data.

`clckhash.clk.chunks(seq, chunk_size)`
Split seq into chunk_size-sized chunks.

Parameters

- **seq** – A sequence to chunk.
- **chunk_size** – The size of chunk.

`clckhash.clk.generate_clk_from_csv(input_f, keys, schema, validate=True, header=True, progress_bar=True)`

Generate Bloom filters from CSV file, then serialise them.

This function also computes and outputs the Hamming weight (a.k.a popcount – the number of bits set to high) of the generated Bloom filters.

Parameters

- **input_f** – A file-like object of csv data to hash.
- **keys** – A tuple of two lists of secret keys.
- **schema** – Schema specifying the record formats and hashing settings.
- **validate** – Set to *False* to disable validation of data against the schema. Note that this will silence warnings whose aim is to keep the hashes consistent between data sources; this may affect linkage accuracy.
- **header** – Set to *False* if the CSV file does not have a header. Set to *'ignore'* if the CSV file does have a header but it should not be checked against the schema.
- **progress_bar** (*bool*) – Set to *False* to disable the progress bar.

Returns A list of serialized Bloom filters and a list of corresponding popcounts.

`clckhash.clk.generate_clks(pii_data, schema, keys, validate=True, callback=None)`

`clckhash.clk.hash_and_serialize_chunk(chunk_pii_data, keys, schema)`

Generate Bloom filters (ie hash) from chunks of PII then serialize the generated Bloom filters. It also computes and outputs the Hamming weight (or popcount) – the number of bits set to one – of the generated Bloom filters.

Parameters

- **chunk_pii_data** – An iterable of indexable records.
- **keys** – A tuple of two lists of secret keys used in the HMAC.
- **schema** (*Schema*) – Schema specifying the entry formats and hashing settings.

Returns A list of serialized Bloom filters and a list of corresponding popcounts

key derivation

```
clckhash.key_derivation.generate_key_lists(master_secrets, num_identifier, key_size=64,
                                           salt=None, info=None, kdf='HKDF',
                                           hash_algo='SHA256')
```

Generates a derived key for each identifier for each master secret using a key derivation function (KDF).

The only supported key derivation function for now is 'HKDF'.

The previous key usage can be reproduced by setting kdf to 'legacy'. This is highly discouraged, as this strategy will map the same n-grams in different identifier to the same bits in the Bloom filter and thus does not lead to good results.

Parameters

- **master_secrets** – a list of master secrets (either as bytes or strings)
- **num_identifier** – the number of identifiers
- **key_size** – the size of the derived keys
- **salt** – salt for the KDF as bytes
- **info** – optional context and application specific information as bytes
- **kdf** – the key derivation function algorithm to use
- **hash_algo** – the hashing algorithm to use (ignored if kdf is not 'HKDF')

Returns The derived keys. First dimension is of size num_identifier, second dimension is the same as master_secrets. A key is represented as bytes.

```
clckhash.key_derivation.hkdf(master_secret, num_keys, hash_algo='SHA256', salt=None,
                             info=None, key_size=64)
```

Executes the HKDF key derivation function as described in rfc5869 to derive num_keys keys of size key_size from the master_secret.

Parameters

- **master_secret** – input keying material
- **num_keys** – the number of keys the kdf should produce
- **hash_algo** – The hash function used by HKDF for the internal HMAC calls. The choice of hash function defines the maximum length of the output key material. Output bytes $\leq 255 * \text{hash digest size}$ (in bytes).
- **salt** – HKDF is defined to operate with and without random salt. This is done to accommodate applications where a salt value is not available. We stress, however, that the use of salt adds significantly to the strength of HKDF, ensuring independence between different uses of the hash function, supporting “source-independent” extraction, and strengthening the analytical results that back the HKDF design.

Random salt differs fundamentally from the initial keying

material in two ways: it is non-secret and can be re-used. Ideally, the salt value is a random (or pseudorandom) string

of the length HashLen. Yet, even a salt value of less quality (shorter in size or with limited entropy) may still make a significant contribution to the security of the output keying material.

- **info** – While the ‘info’ value is optional in the definition of HKDF, it is often of great importance in applications. Its main objective is to bind the derived key material to application- and context-specific information. For example, ‘info’ may contain a protocol number, algorithm identifiers, user identities, etc. In particular, it may prevent the derivation of the same keying material for different contexts (when the same input key material (IKM) is used in such different contexts). It may also accommodate additional inputs to the key expansion part, if so desired (e.g., an application may want to bind the key material to its length L, thus making L part of the ‘info’ field). There is one technical requirement from ‘info’: it should be independent of the input key material value IKM.
- **key_size** – the size of the produced keys

Returns Derived keys

random names

Module to produce a dataset of names, genders and dates of birth and manipulate that list

Currently very simple and not realistic. Additional functions for manipulating the list of names - producing reordered and subset lists with a specific overlap

ClassList class - generate a list of length n of [id, name, dob, gender] lists

TODO: Get age distribution right by using a mortality table TODO: Get first name distributions right by using distributions TODO: Generate realistic errors TODO: Add RESTfull api to generate reasonable name data as requested

```
class clkhash.randomnames.NameList(n)
```

Bases: `object`

Randomly generated PII records.

```
SCHEMA = <Schema (v2): 4 fields>
```

```
generate_random_person(n)
```

Generator that yields details on a person with plausible name, sex and age.

Yields Generated data for one person tuple - (id: int, name: str(‘First Last’), birthdate: str(‘DD/MM/YYYY’), sex: str(‘M’ | ‘F’))

```
generate_subsets(sz, overlap=0.8, subsets=2)
```

Return random subsets with nonempty intersection.

The random subsets are of specified size. If an element is common to two subsets, then it is common to all subsets. This overlap is controlled by a parameter.

Parameters

- **sz** – size of subsets to generate
- **overlap** – size of the intersection, as fraction of the subset length
- **subsets** – number of subsets to generate

Raises `ValueError` – if there aren’t sufficiently many names in the list to satisfy the request; more precisely, raises if $(1 - \text{subsets}) * \text{floor}(\text{overlap} * \text{sz})$

- $\text{subsets} * \text{sz} > \text{len}(\text{self.names})$.

Returns tuple of subsets

load_names()

Loads a name database from package data

Uses data files sourced from <http://www.quietaffiliate.com/free-first-name-and-last-name-databases-csv-and-sql/>

```
randomname_schema = {'clkConfig': {'hash': {'type': 'doubleHash'}, 'k': 30, 'kdf':
```

```
randomname_schema_bytes = b'{\n "version": 1,\n "clkConfig": {\n "l": 1024,\n "k":
```

schema_types

`clkhask.randomnames.load_csv_data(resource_name)`

Loads first column of specified CSV file from package data.

`clkhask.randomnames.random_date(start, end)`

Generate a random datetime between two datetime objects.

Parameters

- **start** – datetime of start
- **end** – datetime of end

Returns random datetime between start and end

`clkhask.randomnames.save_csv(data, headers, file)`

Output generated data to file as CSV with header.

Parameters

- **data** – An iterable of tuples containing raw data.
- **headers** – Iterable of feature names
- **file** – A writeable stream in which to write the CSV

schema

Schema loading and validation.

exception `clkhask.schema.MasterSchemaError`

Bases: `Exception`

Master schema missing? Corrupted? Otherwise surprising? This is the exception for you!

```
class clkhask.schema.Schema(fields, l, xor_folds=0, kdf_type='HKDF', kdf_hash='SHA256',  
                             kdf_info=None, kdf_salt=None, kdf_key_size=64)
```

Bases: `object`

Linkage Schema which describes how to encode plaintext identifiers.

exception `clkhask.schema.SchemaError(msg, errors=None)`

Bases: `Exception`

The user-defined schema is invalid.

`clkhask.schema.convert_v1_to_v2(dict)`

Convert v1 schema dict to v2 schema dict. :param dict: v1 schema dict :return: v2 schema dict

`clkhask.schema.from_json_dict(dct, validate=True)`

Create a Schema for v1 or v2 according to dct

Parameters

- **dct** – This dictionary must have a *‘features’* key specifying the columns of the dataset. It must have a *‘version’* key containing the master schema version that this schema conforms to. It must have a *‘hash’* key with all the globals.
- **validate** – (default True) Raise an exception if the schema does not conform to the master schema.

Raises *SchemaError* – An exception containing details about why the schema is not valid.

Returns the Schema

`clckhash.schema.from_json_file(schema_file, validate=True)`

Load a Schema object from a json file. :param schema_file: A JSON file containing the schema. :param validate: (default True) Raise an exception if the

schema does not conform to the master schema.

Raises *SchemaError* – When the schema is invalid.

Returns the Schema

`clckhash.schema.validate_schema_dict(schema)`

Validate the schema.

This raises iff either the schema or the master schema are invalid. If it’s successful, it returns nothing.

Parameters *schema* – The schema to validate, as parsed by *json*.

Raises

- *SchemaError* – When the schema is invalid.
- *MasterSchemaError* – When the master schema is invalid.

field_formats

Classes that specify the requirements for each column in a dataset. They take care of validation, and produce the settings required to perform the hashing.

class `clckhash.field_formats.DateSpec(identifier, hashing_properties, format, description=None)`

Bases: `clckhash.field_formats.FieldSpec`

Represents a field that holds dates.

Dates are specified as full-dates in a format that can be described as a *strptime()* (C89 standard) compatible format string. E.g.: the format for the standard internet format [RFC3339](#) (e.g. 1996-12-19) is *‘%Y-%m-%d’*.

ivar str format The format of the date.

OUTPUT_FORMAT = *‘%Y%m%d’*

classmethod from_json_dict (*json_dict*)

Make a DateSpec object from a dictionary containing its properties.

Parameters

- **json_dict** (*dict*) – This dictionary must contain a *‘format’* key. In addition, it must contain a *‘hashing’* key, whose contents are passed to *FieldHashingProperties*.
- **json_dict** – The properties dictionary.

validate (*str_in*)

Validates an entry in the field.

Raises *InvalidEntryError* iff the entry is invalid.

An entry is invalid iff (1) the string does not represent a date in the correct format; or (2) the date it represents is invalid (such as 30 February).

Parameters *str_in* (*str*) – String to validate.

Raises

- *InvalidEntryError* – Iff entry is invalid.
- *ValueError* – When self.format is unrecognised.

class `clckhash.field_formats.EnumSpec` (*identifier*, *hashing_properties*, *values*, *description=None*)
Bases: `clckhash.field_formats.FieldSpec`

Represents a field that holds an enum.

The finite collection of permitted values must be specified.

Variables *values* – The set of permitted values.

classmethod `from_json_dict` (*json_dict*)

Make a EnumSpec object from a dictionary containing its properties.

Parameters *json_dict* (*dict*) – This dictionary must contain an ‘enum’ key specifying the permitted values. In addition, it must contain a ‘hashing’ key, whose contents are passed to *FieldHashingProperties*.

validate (*str_in*)

Validates an entry in the field.

Raises *InvalidEntryError* iff the entry is invalid.

An entry is invalid iff it is not one of the permitted values.

Parameters *str_in* (*str*) – String to validate.

Raises *InvalidEntryError* – When entry is invalid.

class `clckhash.field_formats.FieldHashingProperties` (*ngram*, *encoding*=‘utf-8’, *positional*=False, *hash_type*=‘blakeHash’, *prevent_singularity*=None, *num_bits*=None, *k*=None, *missing_value*=None)

Bases: `object`

Stores the settings used to hash a field.

This includes the encoding and tokenisation parameters.

ivar *str encoding* The encoding to use when converting the string to bytes. Refer to *Python’s documentation* <<https://docs.python.org/3/library/codecs.html#standard-encodings>> for possible values.

ivar *int ngram* The n in n-gram. Possible values are 0, 1, and 2.

ivar *bool positional* Controls whether the n-grams are positional.

ivar *int num_bits* dynamic *k* = num_bits / number of n-grams

ivar *int k* max number of bits per n-gram

ks (*num_ngrams*)

Provide a k for each ngram in the field value. :param num_ngrams: number of ngrams in the field value
:return: [k, ...] a k value for each of num_ngrams such that the sum is exactly num_bits

replace_missing_value (*str_in*)

returns 'str_in' if it is not equals to the 'sentinel' as defined in the missingValue section of the schema. Else it will return the 'replaceWith' value.

Parameters *str_in* –

Returns *str_in* or the missingValue replacement value

class `clckhash.field_formats.FieldSpec` (*identifier, hashing_properties, description=None*)

Bases: `object`

Abstract base class representing the specification of a column in the dataset. Subclasses validate entries, and modify the *hashing_properties* ivar to customise hashing procedures.

Variables

- **identifier** (*str*) – The name of the field.
- **description** (*str*) – Description of the field format.
- **hashing_properties** (`FieldHashingProperties`) – The properties for hashing. None if field ignored.

format_value (*str_in*)

formats the value 'str_in' for hashing according to this field's spec.

There are several reasons why this might be necessary:

1. This field contains missing values which have to be replaced by some other string
2. There are several different ways to describe a specific value for this field, e.g.: all of '+65', '65', '65' are valid representations of the integer 65.
3. Entries of this field might contain elements with no entropy, e.g. dates might be formatted as yyyy-mm-dd, thus all dates will have '-' at the same place. These artifacts have no value for entity resolution and should be removed.

param str str_in the string to format

return a string representation of 'str_in' which is ready to

be hashed

classmethod `from_json_dict` (*field_dict*)

Initialise a FieldSpec object from a dictionary of properties.

Parameters **field_dict** (*dict*) – The properties dictionary to use. Must contain a 'hashing' key that meets the requirements of `FieldHashingProperties`. Subclasses may require

Raises `InvalidSchemaError` – When the *properties* dictionary contains invalid values. Exactly what that means is decided by the subclasses.

is_missing_value (*str_in*)

tests if 'str_in' is the sentinel value for this field

Parameters **str_in** (*str*) – String to test if it stands for missing value

Returns True if a missing value is defined for this field and

str_in matches this value

validate (*str_in*)

Validates an entry in the field.

Raises *InvalidEntryError* iff the entry is invalid.

Subclasses must override this method with their own validation. They should call the parent's *validate* method via *super*.

Parameters *str_in* (*str*) – String to validate.

Raises *InvalidEntryError* – When entry is invalid.

class *clckhash.field_formats.Ignore* (*identifier=None*)

Bases: *clckhash.field_formats.FieldSpec*

represent a field which will be ignored throughout the clk processing.

validate (*str_in*)

Validates an entry in the field.

Raises *InvalidEntryError* iff the entry is invalid.

Subclasses must override this method with their own validation. They should call the parent's *validate* method via *super*.

Parameters *str_in* (*str*) – String to validate.

Raises *InvalidEntryError* – When entry is invalid.

class *clckhash.field_formats.IntegerSpec* (*identifier, hashing_properties, description=None, minimum=None, maximum=None, **kwargs*)

Bases: *clckhash.field_formats.FieldSpec*

Represents a field that holds integers.

Minimum and maximum values may be specified.

Variables

- **minimum** (*int*) – The minimum permitted value.
- **maximum** (*int*) – The maximum permitted value or None.

classmethod *from_json_dict* (*json_dict*)

Make a IntegerSpec object from a dictionary containing its properties.

Parameters

- **json_dict** (*dict*) – This dictionary may contain '*minimum*' and '*maximum*' keys. In addition, it must contain a '*hashing*' key, whose contents are passed to *FieldHashingProperties*.
- **json_dict** – The properties dictionary.

validate (*str_in*)

Validates an entry in the field.

Raises *InvalidEntryError* iff the entry is invalid.

An entry is invalid iff (1) the string does not represent a base-10 integer; (2) the integer is not between *self.minimum* and *self.maximum*, if those exist; or (3) the integer is negative.

Parameters *str_in* (*str*) – String to validate.

Raises *InvalidEntryError* – When entry is invalid.

exception `clckhash.field_formats.InvalidEntryError`

Bases: `ValueError`

An entry in the data file does not conform to the schema.

field_spec = `None`

exception `clckhash.field_formats.InvalidSchemaError`

Bases: `ValueError`

Raised if the schema of a field specification is invalid.

For example, a regular expression included in the schema is not syntactically correct.

field_spec_index = `None`

json_field_spec = `None`

class `clckhash.field_formats.MissingValueSpec` (*sentinel, replace_with=None*)

Bases: `object`

Stores the information about how to find and treat missing values.

Variables **sentinel** (*str*) – sentinel is the string that identifies a missing value e.g.: ‘N/A’, ‘’. The sentinel will not be validated against the feature format definition :ivar str replaceWith: defines the string which replaces the sentinel whenever present, can be ‘None’, then sentinel will not be replaced.

classmethod **from_json_dict** (*json_dict*)

class `clckhash.field_formats.StringSpec` (*identifier, hashing_properties, description=None, regex=None, case='mixed', min_length=0, max_length=None*)

Bases: `clckhash.field_formats.FieldSpec`

Represents a field that holds strings.

One way to specify the format of the entries is to provide a regular expression that they must conform to. Another is to provide zero or more of: minimum length, maximum length, casing (lower, upper, mixed).

Each string field also specifies an encoding used when turning characters into bytes. This is stored in *hashing_properties* since it is needed for hashing.

Variables

- **encoding** (*str*) – The encoding to use when converting the string to bytes. Refer to [Python’s documentation](#) for possible values.
- **regex** – Compiled regular expression that entries must conform to. Present only if the specification is regex-based.
- **case** (*str*) – The casing of the entries. One of ‘lower’, ‘upper’, or ‘mixed’. Default is ‘mixed’. Present only if the specification is not regex-based.
- **min_length** (*int*) – The minimum length of the string. *None* if there is no minimum length. Present only if the specification is not regex-based.
- **max_length** (*int*) – The maximum length of the string. *None* if there is no maximum length. Present only if the specification is not regex-based.

classmethod **from_json_dict** (*json_dict*)

Make a StringSpec object from a dictionary containing its properties.

Parameters `json_dict` (*dict*) – This dictionary must contain an *‘encoding’* key associated with a Python-conformant encoding. It must also contain a *‘hashing’* key, whose contents are passed to *FieldHashingProperties*. Permitted keys also include *‘pattern’*, *‘case’*, *‘minLength’*, and *‘maxLength’*.

Raises *InvalidSchemaError* – When a regular expression is provided but is not a valid pattern.

validate (*str_in*)

Validates an entry in the field.

Raises *InvalidEntryError* iff the entry is invalid.

An entry is invalid iff (1) a pattern is part of the specification of the field and the string does not match it; (2) the string does not match the provided casing, minimum length, or maximum length; or (3) the specified encoding cannot represent the string.

Parameters `str_in` (*str*) – String to validate.

Raises

- *InvalidEntryError* – When entry is invalid.
- *ValueError* – When self.case is not one of the permitted values (*‘lower’*, *‘upper’*, or *‘mixed’*).

`clkh.hash.field_formats.fhp_from_json_dict` (*json_dict*)

Make a *FieldHashingProperties* object from a dictionary.

Parameters `json_dict` (*dict*) – Conforming to the *hashingConfig* definition in the v2 linkage schema.

Returns A *FieldHashingProperties* instance.

`clkh.hash.field_formats.spec_from_json_dict` (*json_dict*)

Turns a dictionary into the appropriate *FieldSpec* object.

Parameters `json_dict` (*dict*) – A dictionary with properties.

Raises *InvalidSchemaError* –

Returns An initialised instance of the appropriate *FieldSpec* subclass.

tokenizer

Functions to tokenize words (PII)

`clkh.tokenizer.get_tokenizer` (*fhp*)

Get tokenizer function from the hash settings.

This function takes a *FieldHashingProperties* object. It returns a function that takes a string and tokenises based on those properties.

1.4.2 Testing

Make sure you have all the required modules before running the tests (modules that are only needed for tests are not included during installation):

```
$ pip install -r requirements.txt
```

Now run the unit tests and print out code coverage with *py.test*:


```
$ python -m pytest --cov=clkhsh
```

Note several tests will be skipped by default. To enable the command line tests set the `INCLUDE_CLI` environment variable. To enable the tests which interact with an entity service set the `TEST_ENTITY_SERVICE` environment variable to the target service's address:

```
$ TEST_ENTITY_SERVICE= INCLUDE_CLI= python -m pytest --cov=clkhsh
```

1.4.3 Type Checking

clkhsh uses static typechecking with `mypy`. To run the type checker (in Python 3.5 or later):

```
$ pip install mypy
$ mypy clkhsh --ignore-missing-imports --strict-optional --no-implicit-optional --
  ↳ disallow-untyped-calls
```

1.4.4 Packaging

The `clkutil` command line tool can be frozen into an exe using [PyInstaller](#):

```
pyinstaller cli.spec
```

Look for `clkutil.exe` in the `dist` directory.

1.5 Rest Client API Documentation

clkhsh includes a module for interacting with the anonlink-entity-service.

exception `clkhsh.rest_client.RateLimitedClient` (*msg, response*)

Bases: `clkhsh.rest_client.ServiceError`

Exception indicating client is asking for updates too frequently.

exception `clkhsh.rest_client.ServiceError` (*msg, response*)

Bases: `Exception`

Problem with the upstream API

`clkhsh.rest_client.format_run_status` (*status*)

`clkhsh.rest_client.project_create` (*server, schema, result_type, name, notes=None, parties=2*)

`clkhsh.rest_client.project_delete` (*server, project, apikey*)

`clkhsh.rest_client.project_get_description` (*server, project, apikey*)

`clkhsh.rest_client.project_upload_clks` (*server, project, apikey, clk_data*)

`clkhsh.rest_client.run_create` (*server, project_id, apikey, threshold, name, notes=None*)

`clkhsh.rest_client.run_delete` (*server, project, run, apikey*)

`clkhsh.rest_client.run_get_result_text` (*server, project, run, apikey*)

`clkhsh.rest_client.run_get_status` (*server, project, run, apikey*)

```
clkhask.rest_client.server_get_status(server)
```

```
clkhask.rest_client.wait_for_run(server, project, run, apikey, timeout=None, up-  
                                date_period=1)
```

Monitor a linkage run and return the final status updates. If a timeout is provided and the run hasn't entered a terminal state (error or completed) when the timeout is reached a `TimeoutError` will be raised.

Parameters

- **server** – Base url of the upstream server.
- **project** –
- **run** –
- **apikey** –
- **timeout** – Stop waiting after this many seconds. The default (`None`) is to never give you up.
- **update_period** – Time in seconds between queries to the run's status.

:raises `TimeoutError`

```
clkhask.rest_client.watch_run_status(server, project, run, apikey, timeout=None, up-  
                                   date_period=1)
```

Monitor a linkage run and yield status updates. Will immediately yield an update and then only yield further updates when the status object changes. If a timeout is provided and the run hasn't entered a terminal state (error or completed) when the timeout is reached, updates will cease and a `TimeoutError` will be raised.

Parameters

- **server** – Base url of the upstream server.
- **project** –
- **run** –
- **apikey** –
- **timeout** – Stop waiting after this many seconds. The default (`None`) is to never give you up.
- **update_period** – Time in seconds between queries to the run's status.

:raises `TimeoutError`

1.6 References

CHAPTER 2

External Links

- [clckhash on Github](#)
- [clckhash on PyPi](#)

CHAPTER 3

Indices and tables

- `genindex`
- `modindex`

Bibliography

- [Schnell2011] Schnell, R., Bachteler, T., & Reiher, J. (2011). [A Novel Error-Tolerant Anonymous Linking Code](#).
- [Schnell2016] Schnell, R., & Borgs, C. (2016). XOR-Folding for hardening Bloom Filter-based Encryptions for Privacy-preserving Record Linkage.
- [Kroll2015] Kroll, M., & Steinmetzer, S. (2015). Who is 1011011111...1110110010? automated cryptanalysis of bloom filter encryptions of databases with several personal identifiers. In *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-319-27707-3_21
- [Kaminsky2011] Kaminsky, A. (2011). [GPU Parallel Statistical and Cube Test Analysis of the SHA-3 Finalist Candidate Hash Functions](#).

C

- `clckhash.bloomfilter`, 32
- `clckhash.clk`, 35
- `clckhash.field_formats`, 39
- `clckhash.key_derivation`, 36
- `clckhash.randomnames`, 37
- `clckhash.rest_client`, 45
- `clckhash.schema`, 38
- `clckhash.tokenizer`, 44

B

`blake_encode_ngrams()` (in module `clkhash.bloomfilter`), 32

C

`chunks()` (in module `clkhash.clk`), 35
`clkhash.bloomfilter` (module), 32
`clkhash.clk` (module), 35
`clkhash.field_formats` (module), 39
`clkhash.key_derivation` (module), 36
`clkhash.randomnames` (module), 37
`clkhash.rest_client` (module), 45
`clkhash.schema` (module), 38
`clkhash.tokenizer` (module), 44
`convert_v1_to_v2()` (in module `clkhash.schema`), 38
`crypto_bloom_filter()` (in module `clkhash.bloomfilter`), 33

D

`DateSpec` (class in `clkhash.field_formats`), 39
`double_hash_encode_ngrams()` (in module `clkhash.bloomfilter`), 33
`double_hash_encode_ngrams_non_singular()` (in module `clkhash.bloomfilter`), 33

E

`EnumSpec` (class in `clkhash.field_formats`), 40

F

`fhp_from_json_dict()` (in module `clkhash.field_formats`), 44
`field_spec` (`clkhash.field_formats.InvalidEntryError` attribute), 43
`field_spec_index` (`clkhash.field_formats.InvalidSchemaError` attribute), 43
`FieldHashingProperties` (class in `clkhash.field_formats`), 40
`FieldSpec` (class in `clkhash.field_formats`), 41

`fold_xor()` (in module `clkhash.bloomfilter`), 34
`format_run_status()` (in module `clkhash.rest_client`), 45
`format_value()` (`clkhash.field_formats.FieldSpec` method), 41
`from_json_dict()` (`clkhash.field_formats.DateSpec` class method), 39
`from_json_dict()` (`clkhash.field_formats.EnumSpec` class method), 40
`from_json_dict()` (`clkhash.field_formats.FieldSpec` class method), 41
`from_json_dict()` (`clkhash.field_formats.IntegerSpec` class method), 42
`from_json_dict()` (`clkhash.field_formats.MissingValueSpec` class method), 43
`from_json_dict()` (`clkhash.field_formats.StringSpec` class method), 43
`from_json_dict()` (in module `clkhash.schema`), 38
`from_json_file()` (in module `clkhash.schema`), 39

G

`generate_clk_from_csv()` (in module `clkhash.clk`), 35
`generate_clks()` (in module `clkhash.clk`), 35
`generate_key_lists()` (in module `clkhash.key_derivation`), 36
`generate_random_person()` (`clkhash.randomnames.NameList` method), 37
`generate_subsets()` (`clkhash.randomnames.NameList` method), 37
`get_tokenizer()` (in module `clkhash.tokenizer`), 44

H

`hash_and_serialize_chunk()` (in module `clkhash.clk`), 35
`hashing_function_from_properties()` (in module `clkhash.bloomfilter`), 34
`hkdf()` (in module `clkhash.key_derivation`), 36

I

Ignore (class in *clkhash.field_formats*), 42
int_from_bytes() (in module *clkhash.bloomfilter*), 34
IntegerSpec (class in *clkhash.field_formats*), 42
InvalidEntryError, 42
InvalidSchemaError, 43
is_missing_value() (in module *clkhash.field_formats.FieldSpec* method), 41

J

json_field_spec(*clkhash.field_formats.InvalidSchemaError* attribute), 43

K

ks() (*clkhash.field_formats.FieldHashingProperties* method), 40

L

load_csv_data() (in module *clkhash.randomnames*), 38
load_names() (*clkhash.randomnames.NameList* method), 37

M

MasterSchemaError, 38
MissingValueSpec (class in *clkhash.field_formats*), 43

N

NameList (class in *clkhash.randomnames*), 37

O

OUTPUT_FORMAT (*clkhash.field_formats.DateSpec* attribute), 39

P

project_create() (in module *clkhash.rest_client*), 45
project_delete() (in module *clkhash.rest_client*), 45
project_get_description() (in module *clkhash.rest_client*), 45
project_upload_clks() (in module *clkhash.rest_client*), 45

R

random_date() (in module *clkhash.randomnames*), 38
randomname_schema (*clkhash.randomnames.NameList* attribute), 38

randomname_schema_bytes (*clkhash.randomnames.NameList* attribute), 38
RateLimitedClient, 45
replace_missing_value() (*clkhash.field_formats.FieldHashingProperties* method), 41
run_create() (in module *clkhash.rest_client*), 45
run_delete() (in module *clkhash.rest_client*), 45
run_get_result_text() (in module *clkhash.rest_client*), 45
run_get_status() (in module *clkhash.rest_client*), 45

S

save_csv() (in module *clkhash.randomnames*), 38
Schema (class in *clkhash.schema*), 38
SCHEMA (*clkhash.randomnames.NameList* attribute), 37
schema_types (*clkhash.randomnames.NameList* attribute), 38
SchemaError, 38
server_get_status() (in module *clkhash.rest_client*), 45
ServiceError, 45
spec_from_json_dict() (in module *clkhash.field_formats*), 44
stream_bloom_filters() (in module *clkhash.bloomfilter*), 34
StringSpec (class in *clkhash.field_formats*), 43

V

validate() (*clkhash.field_formats.DateSpec* method), 39
validate() (*clkhash.field_formats.EnumSpec* method), 40
validate() (*clkhash.field_formats.FieldSpec* method), 41
validate() (*clkhash.field_formats.Ignore* method), 42
validate() (*clkhash.field_formats.IntegerSpec* method), 42
validate() (*clkhash.field_formats.StringSpec* method), 44
validate_schema_dict() (in module *clkhash.schema*), 39

W

wait_for_run() (in module *clkhash.rest_client*), 46
watch_run_status() (in module *clkhash.rest_client*), 46