dedupe

An active learning library

How to use: https://github.com/dedupeio/dedupe-examples
Documentation: https://docs.dedupe.io/en/latest/API-documentation.html

What is dedupe?

- A library that uses machine learning (active learning) to perform de-duplication and entity resolution quickly on structured data.
- It finds clusters (same data records) in the given data, by training passively and actively. How does active learning look like:

```
Phone :  2850617
Address :  3801 s. wabash
Zip :
Site name :  ada s. mckinley st. thomas cdc

Phone :  2850617
Address :  3801 s wabash ave
Zip :
Site name :  ada s. mckinley community services - mckinley - st. thomas

Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished
```

dedupe's workflow

- Train using labeled examples
- Predict label for unlabeled data
- If confidence on prediction is low, ask user (active learning phase).
- Once the user has labeled some amount of data (at least 10 true and 10 false), train again.
- Cluster using the trained model.
- Save trained model and labels (in case the user plans on re-running everything, dedupe can continue training from where left).

An Example

- De-duplicate addresses
- Takes around 45 seconds to run

Setup

- Clone the repo:

```
git clone https://github.com/dedupeio/dedupe-examples.git
cd dedupe-examples
```

- Install the required packages:

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

- Run csv_example:

```
cd csv_example
pip install unidecode
python csv_example.py
```

Some tips for setup

- Ignore their virtualenv and virtualenvwrapper commands:
  - There are just 4 packages anyways.
  - Creating a separate virtual environment for just this task isn't necessary.
  - Plus, it complicates things a bit.
- If you get an error along the lines of 'unidecode missing', just do pip install unidecode.

Major Functions

- Create deduper object: `dedupe.Dedupe(fields)`
- Prepare training (include previously saved labels): `deduper.prepare_training(data)`
- Ask user to label: `dedupe.console_label(deduper)`
- Train: `deduper.train()`
- Save newly trained labels: `deduper.write_training(training_fname)`
- Save trained settings: `deduper.write_settings(settings_fname)`

Reference: https://github.com/dedupeio/dedupe-examples/blob/master/csv_example/csv_example.py

A PostgreSQL (Bigger) Example

- De-duplicate addresses.
- Passive clustering takes around **50 minutes** of time and **12 GB** of RAM.
- Demo video is located at Files/Hands-on Session/dedupe within Canvas

How to run the PostgreSQL example locally?

- Make sure PostgreSQL is running
- `cd pgsql_big_dedupe_example`
- Download the below 3 files into this directory (located at Files/Hands-on Session/dedupe within Canvas):
  - `pgsql_big_dedupe_example_init_db_mod.py`
  - `pgsql_big_dedupe_example_mod.py`
  - `Illinois-campaign-contributions.csv`
- These 3 files are required because the original ones have some errors.
- The csv file (11 MB) is a subset of the original csv file (113 MB) that this example tries to download.
- The python files are modified versions of their original files (ones without the _mod suffix). They help in interfacing with pgsql much more easily.
- `python pgsql_big_dedupe_example_init_db_mod.py`
- `python pgsql_big_dedupe_example_mod.py`

END.