Onigiri

A collaborative and open-source NE resolution application for the Humanities and Social Sciences

While humanities scholars engage more and more into studies that involve computational approaches (ie. distant reading), researchers face new difficulties when preparing their datasets. While data collection tools and techniques such as crawling and scraping tend to make their way into scholars practices, it becomes harder to use a unique ensemble of data coming from different sources. Data integration is the task of joining such data based on an entity's attributes (ie. a person's name). While common data integration applications are usually used for commercial purposes and use machine learning to instantly match a large number entities across datasets, Onigiri has been developed to avoid false positives by letting the expert do the match via an easy-to-use interface. It has been designed so scholars can choose the matching method to apply across datasets stored as CSV files and resolve these matches collaboratively.

Motivations

What are data integration and NE resolution ?

Data integration is the task of merging data comming from different sources in single unified dataset.
Named-Entity (NE) resolution is the task of identifying different manifestations of a real world object.
In the field of Digital Humanities, data integration becomes meaningfull as we tend to perform more and more complex analysis. For example, tracing actors through networks analysis usually requires multiple informations about the actors themselves as well as informations about the relationships they have with other kind of actors.

What tools already exists ?

Data integration:
- Talend (complex framework)
- Oracle Data Integrator (complex framework)
- Actian,
- IMB
- …

NE Resolution:
- SERF (machine learning),
- Duke (machine learning),
- Dedoop (complex framework)
- …

Then, why develop Onigiri ?

Common NE resolution softwares rely on machine learning and entities' metadata to quickly find matches across datasets. These approaches can be useful when false positives are tolerated. Unfortunately no applications exists to assist researchers in performing manual entity resolution.

Onigiri is a **straight forward** application that rely on the expert knowledge during all the matching phase to **avoid false positives**.

Methods

String normalization

Reduces the number of possible keys by two (ultra) simple steps :

- Lower all characters in the string,
- Alphasort the different elemtents composing the string

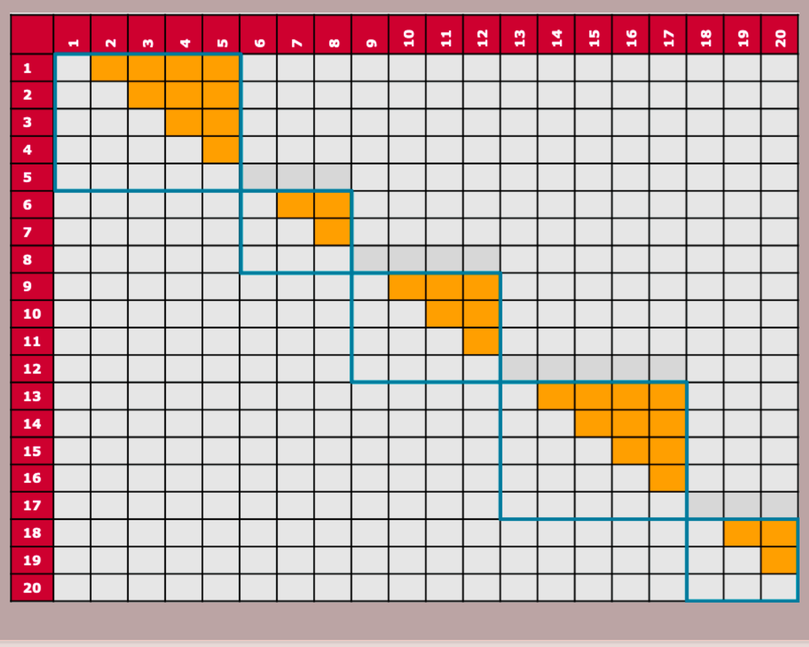
François du Chesnay

↓ normalize()

chesnay du françois

Blocking technique (SNM)

Helps reducing the space of search



String comparison

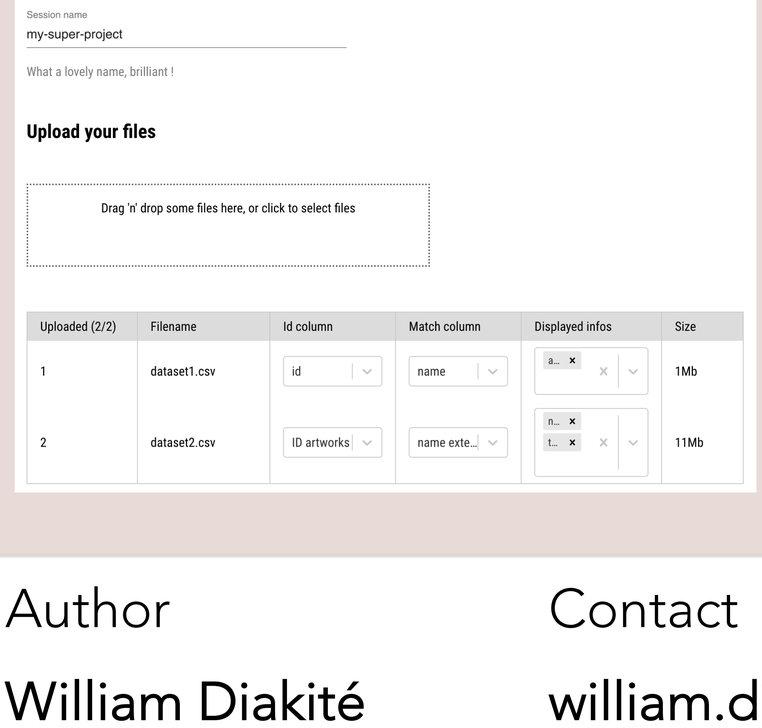
Compute the edit distance between two strings

		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

Application

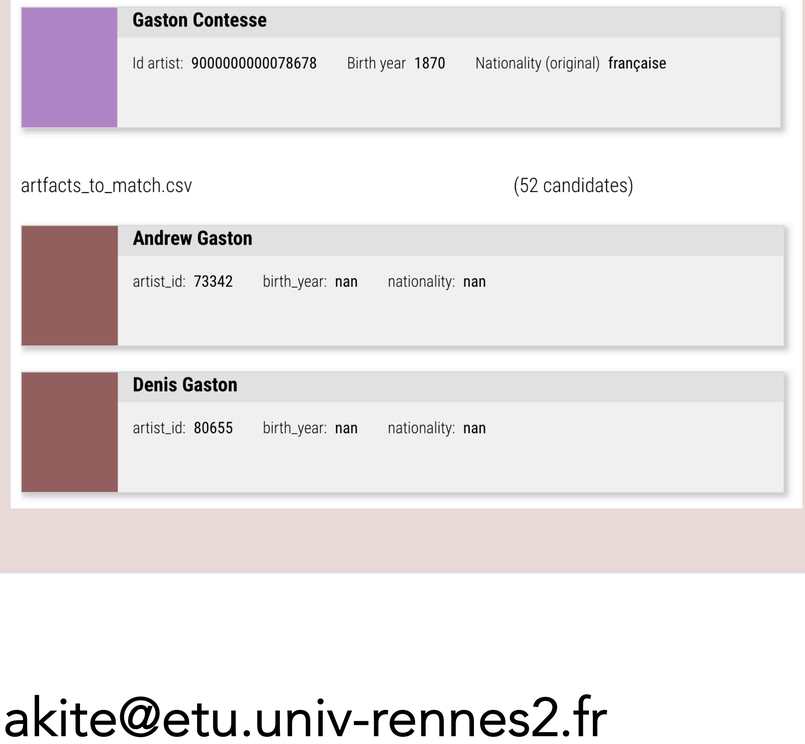
Upload

- Upload your csv files
- Configure your project



Match

Resolve the matches in few clicks



Future features

- NE linking: link the named entities to a reference knowledge base (ie. Wikidata, Getty authorities, Dbpedia)
- More string comparison algorithms (ie Metaphone colision)

Author

William Diakité

Contact

william.diakite@etu.univ-rennes2.fr

Website

www.williamdiakite.gitlab.io/onigiri