Integration of heterogeneous datasets

Arthur Imbert, Inria, June 21, 2017

https://gitlab.inria.fr/parietal/arthur_imbert/

Introduction

Problems related to data integration

- Record linkage

Patient Name: Vivian Christensen
Visit ID: 837720
Date of Birth: 3/20/1953
SSN: 000-86-6628
MRN: 9968427
Dx: 414.00

Patient Name: Viv Christensen
Visit ID: 483005
Date of Birth: 3/20/1953
SSN: 000-68-6628
MRN: 0099523461
Dx: 493.01

Problems related to data integration

- Record linkage

Patient Name: Vivian Christensen
Visit ID: 837720
Date of Birth: 3/20/1953
SSN: 000-86-6628
MRN: 9968427
Dx: 414.00

Patient Name: Viv Christensen
Visit ID: 483005
Date of Birth: 3/20/1953
SSN: 000-68-6628
MRN: 0099523461
Dx: 493.01

- Foreign key discovery

Problems related to data integration

- Record linkage

Patient Name: Vivian Christensen
Visit ID: 837720
Date of Birth: 3/20/1953
SSN: 000-86-6628
MRN: 9968427
Dx: 414.00

Patient Name: Viv Christensen
Visit ID: 483005
Date of Birth: 3/20/1953
SSN: 000-68-6628
MRN: 0099523461
Dx: 493.01

- Foreign key discovery
- Semantic analysis

What is Open data?

"Open data is the idea that some data should be **freely available** to everyone to use and republish as they wish, **without restrictions** from copyright, patents or other mechanisms of control"

What is Open data?

"Open data is the idea that some data should be **freely available** to everyone to use and republish as they wish, **without restrictions** from copyright, patents or other mechanisms of control"

Volume and heterogeneity!

Reuses

1711 reuses using one or several open datasets are listed

- Map of public services
- Visualization of Presidential election results
- Statistical analysis crossing socio-economic data
- ...

Reuses

1711 reuses using one or several open datasets are listed

- Map of public services
- Visualization of Presidential election results
- Statistical analysis crossing socio-economic data
- ...

New task: Predict reuses

Closer datasets are in a file embedding, more relevant a reuse should be

Build a corpus of datasets

55138 downloaded files (348GB)

23112 cleaned files (72GB)

'Tablizing' everything

An example of file easy to clean

ods_adresse	Code_postal	Libelle_commune	Sy
23 RUE DOMREMY	75013	Paris 13e Arrondissement	Sa
SAINT VICTOR SUR LOIRE	42230	Saint-Étienne	Sa
LD LE GRAND HAMEAU	14520	Sainte-Honorine-des-Pertes	Tr
8 Rue de Vaugirard	75006	Paris 6e Arrondissement	Sa
Rue de la Mécanique	27400	Louviers	Tr

- a tabular form
- a first row as header
- consistency of the data below the header

... still with some limitations

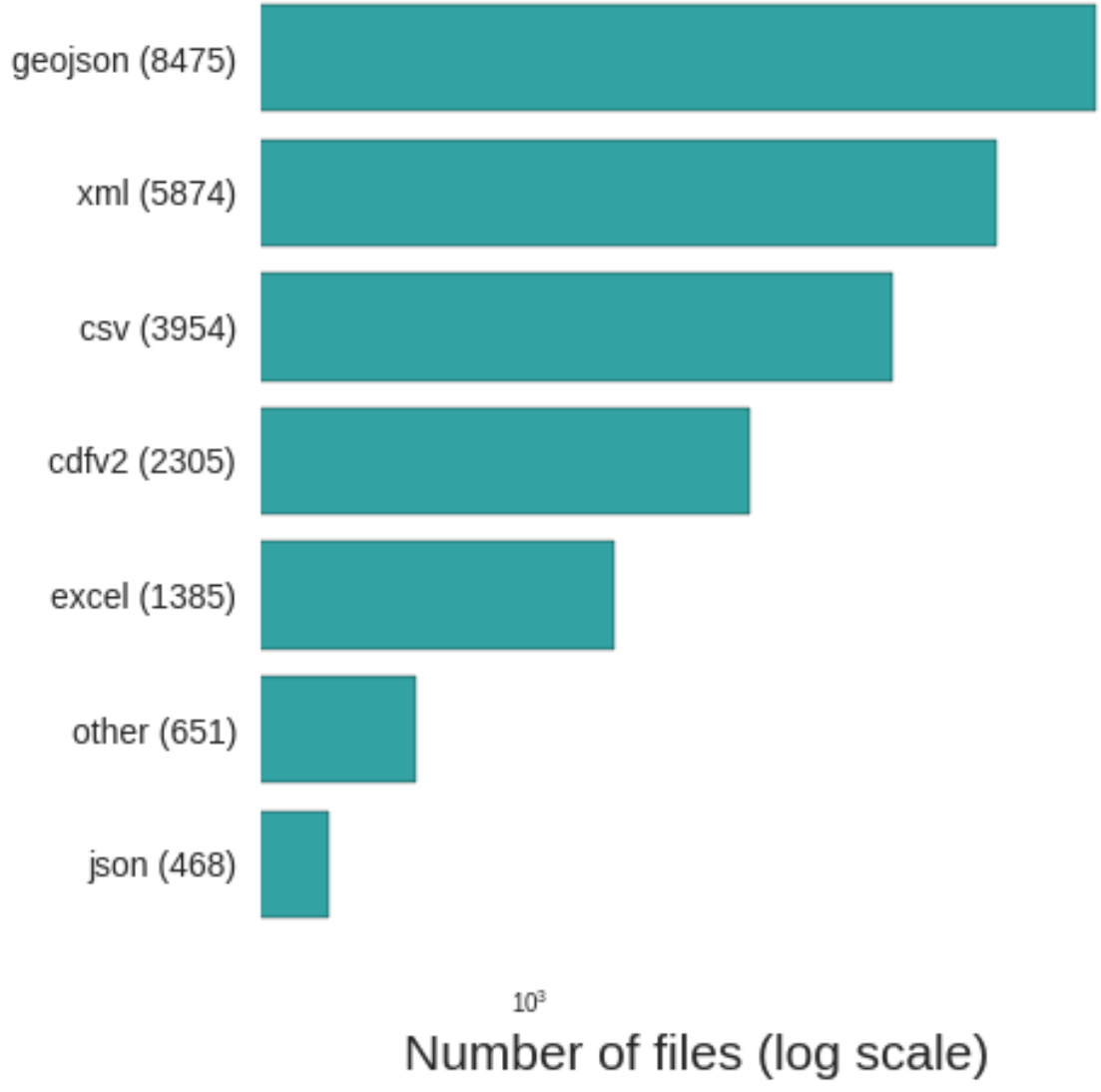
- no Uniforme Ressource Identifier for the adresses
- no Uniforme Ressource Identifier for the city names

'Tablizing' everything

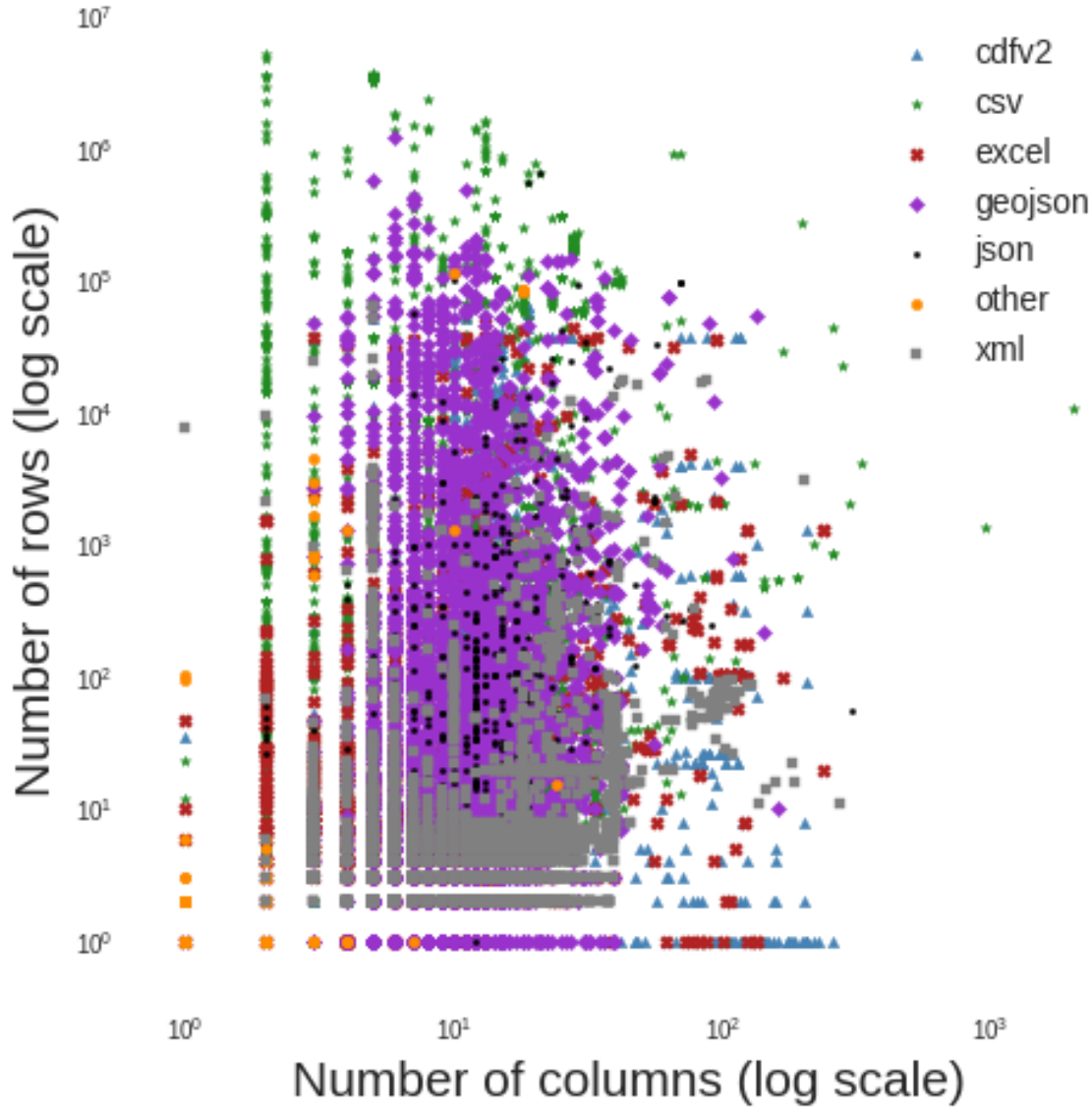
An example of 'dirty data'

	A	B	C	D	E
	T79JNAIS : Répartition quotidienne des naissances vivantes				
	CHAMP : France métropolitaine, territoire au 31 décembre 2013				
		JOUR	01	02	03
	ANNÉE DE 1968 À 2013	*MOIS*			
2		Septembre	1,844	1,741	2,237
3		Octobre	2,393	2,471	2,327
4		Novembre	1,949	2,264	1,889
5		Décembre	1,948	1,768	2,197
6		Janvier	1,745	2,187	2,151
7		Février	2,173	1,879	1,778
8		Mars	2,222	1,872	1,712
9		Avril	1,741	2,129	2,167
0		Mai	1,890	2,281	2,291
1		Juin	1,887	1,723	2,081
2	2013	Juillet	2,298	2,337	2,370
3		Août	2,441	2,426	2,008
4		Septembre	1,818	2,138	2,327
5		Octobre	2,428	2,385	2,333
6		Novembre	1,830	1,834	1,868
7		Décembre	1,871	2,220	2,284
8					
9	Source : Insee, statistiques de l'état civil				
0					

Extensions of the cleaned files



Size of the cleaned files



Build a file embedding

TF-IDF

Non-negative Matrix Factorization

Metric learning

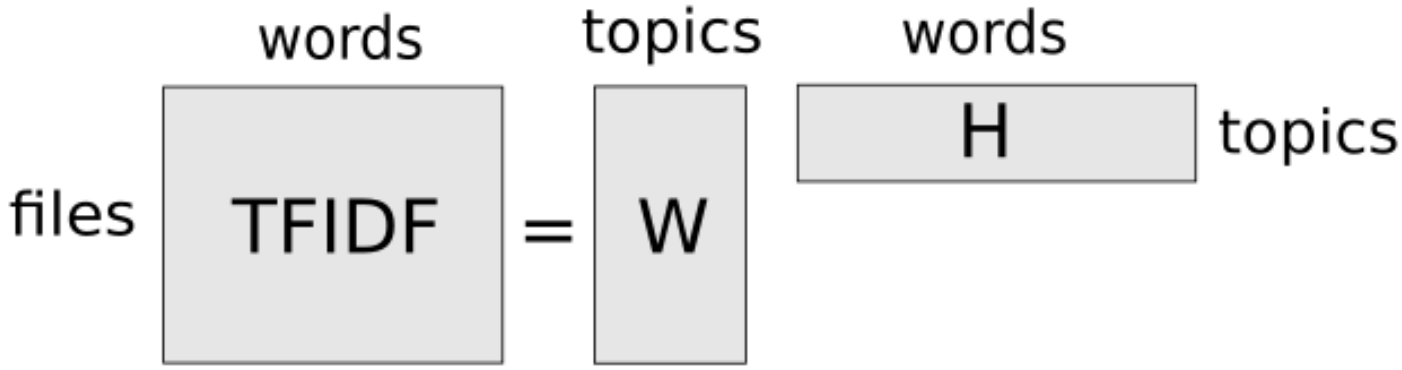
Cross-validation

TF-IDF and NMF

Weight each word frequency by its inverse document frequency

TF-IDF and NMF

Weight each word frequency by its inverse document frequency

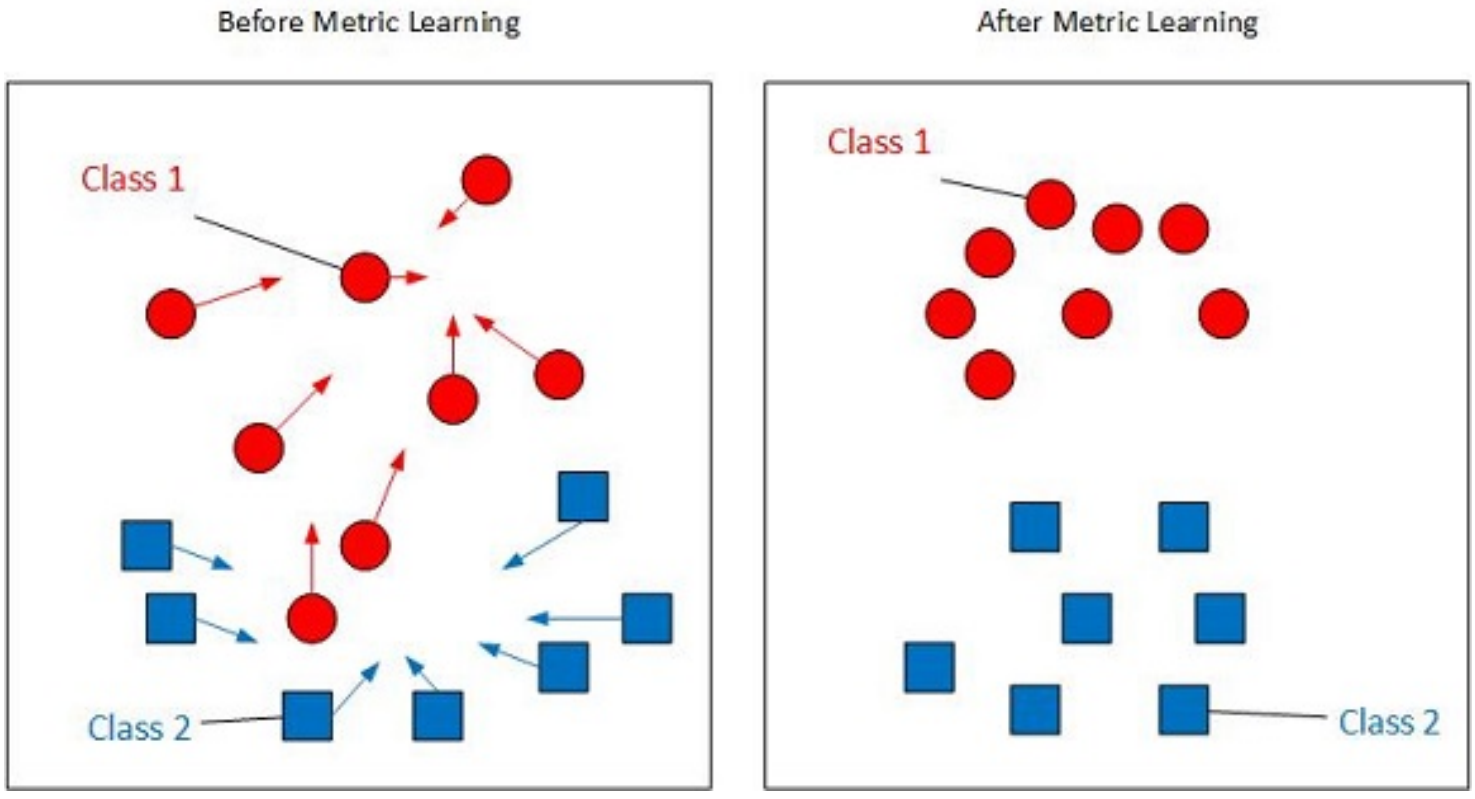


W = the topic space

H = the dictionary

Metric learning

X and **y** built from **W**

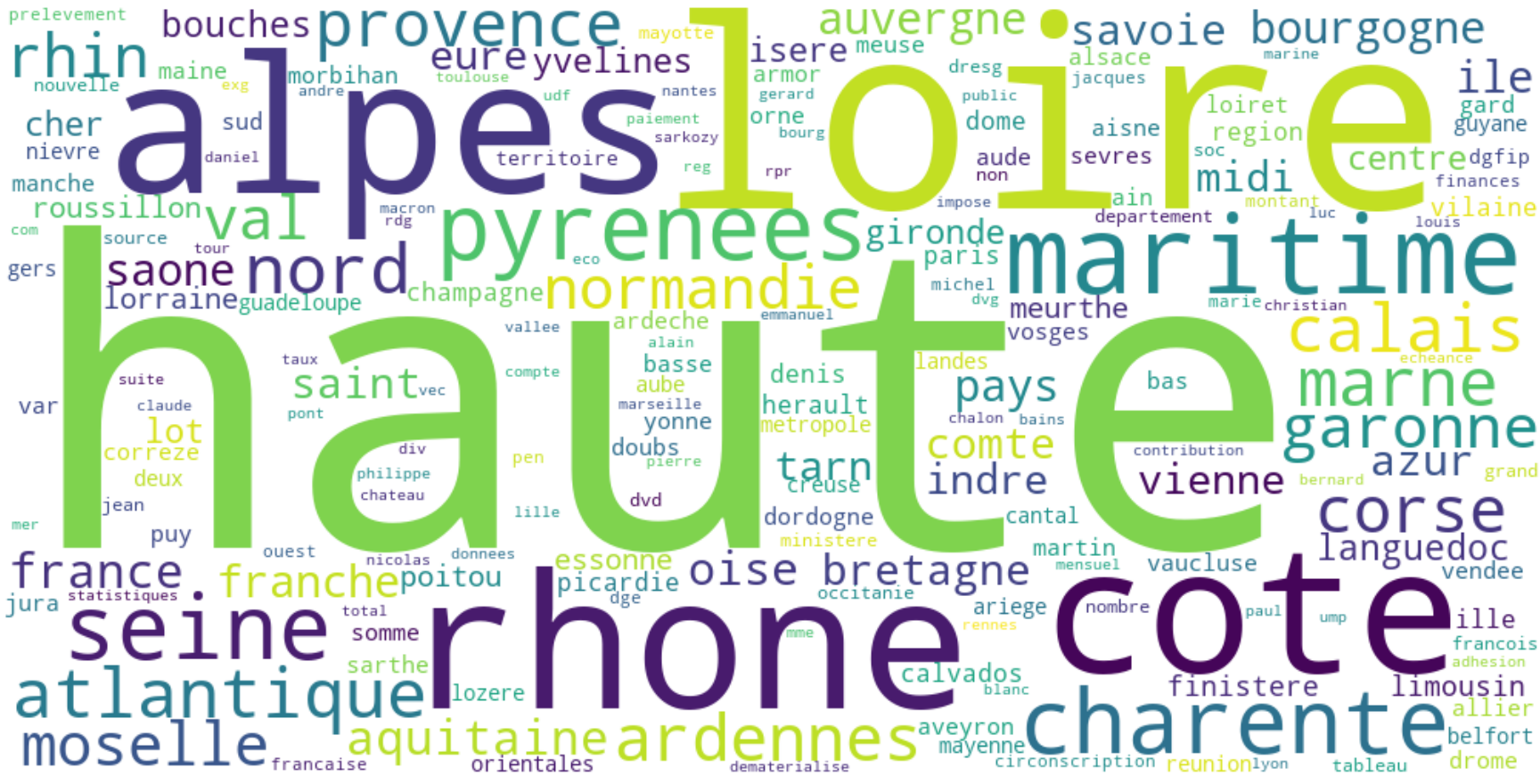


Cross-validation results

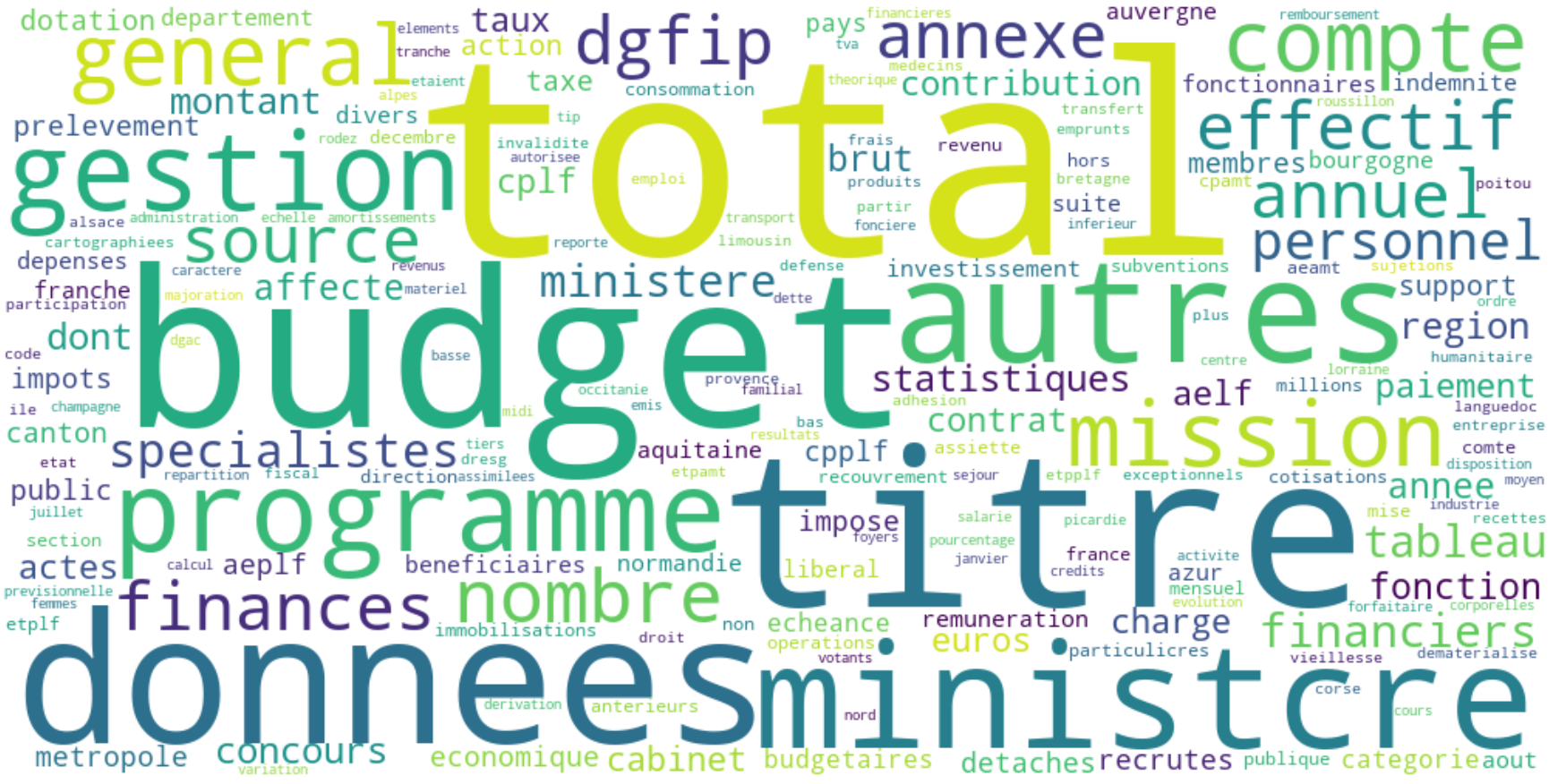
Results

Cosine distance in the embedding

Region topic



Budget topic



Criminality topic

Limits

- Heterogeneous data and cleaning failures
- Not enough reuses
- Undersampling of non reused pairs
- Reuse = similarity?

Future work

- Use `geopandas` to infer geographical information from GEOJSON
- Infer temporal information
- Perform foreign key discovery and record linkage over the closest files and merge them
- Share the results with https://www.data.gouv.fr/fr/


KEEP
CALM
AND
OPEN
DATA