

Crawling and structuring Open Data

from <http://www.data.gouv.fr/>

Arthur Imbert, Inria, June 21, 2017

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



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"

==> Technical restrictions!

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"

==> Technical restrictions!

There are different levels of quality:

- data available on the web under an open license
- data available in a structured format
- data available in a non-proprietary open format
- data with Uniform Ressource Identifier
- data linked to another data


A government platform



Accueil - Data.gouv.fr - Mozilla Firefox

Accueil - Data.gouv.fr x +

Gouvernement Premier Ministre (FR) | <https://www.data.gouv.fr/fr/> Search

 **data.gouv.fr** *Plateforme ouverte des données publiques françaises*

Découvrez l'OpenData Données Tableau de bord 🇫🇷 Connexion / Inscription


Recherche

- Agriculture et alimentation
- Culture
- Économie et Emploi
- Éducation et Recherche
- International et Europe
- Logement, Développement Durable et Énergie
- Santé et Social
- Société
- Territoires, Transports, Tourisme




Partagez, améliorez et réutilisez les données publiques

+ CONTRIBUEZ !

MEILLEURES RÉUTILISATIONS

 **DataFrance**

DERNIÈRES RÉUTILISATIONS

-  Khartis
-  Sciences Po
20 avril 2017
-  Obésité - Fastfoods, Alcool... : Où ?

<https://www.data.gouv.fr/fr/>

An example of dataset: food inspection reports

Résultats des contrôles officiels sanitaires : dispositif d'information « Alim'confiance » - Data.gouv.fr - Mozilla Firefox

Résultats des contrôle x Produits alimentaires x Découpage administr x Jeux de données - Dat x Carte des contrôles sa x +

Gouvernement Premier Ministre (FR) | <https://www.data.gouv.fr/fr/datasets/resultats-des-contrôles-officiels-s> Search

le 17 décembre 2016 généralise l'expérimentation menée à Paris et à Avignon de juillet à décembre 2015. La publication des résultats des contrôles réalisés à partir du 1er mars 2017 dans tous les établissements de la chaîne alimentaire sera effective à partir du 3 avril 2017, sur le site internet www.alim-confiance.gouv.fr.

Quels sont secteurs d'activité concernés ?

Il s'agit de rendre public le résultat des contrôles officiels en sécurité sanitaire des aliments réalisés dans tous les établissements de la chaîne alimentaire : abattoirs, commerces de détail (métiers de bouche, restaurants, supermarchés, marchés, vente à la ferme, etc.), restaurants collectifs et établissements agroalimentaires.

Quelles sont les modalités d'affichage ?

Les établissements de remise directe (restaurants, métiers de bouche, distributeurs) et de restauration collective auront la possibilité d'afficher sur leur devanture le niveau d'hygiène de l'établissement. Cette affichette leur sera transmise par les services départementaux de l'État. Elle sera également téléchargeable sur le site Internet.

Dans tous les pays, la mise en place de la mesure s'est toujours accompagnée d'une amélioration du niveau sanitaire des établissements

Plus d'information à l'adresse : https://dgal.opendatasoft.com/explore/dataset/export_alimconfiance/

Ressources

CSV	Données brutes CSV Dernière modification le lundi 3 avril 2017
XLS	Données brutes Excel Dernière modification le lundi 3 avril 2017
JSON	Données brutes JSON Dernière modification le lundi 3 avril 2017
NONE	API Dernière modification le lundi 13 mars 2017

ET DE LA FORÊT

Le ministre de l'agriculture, de l'agroalimentaire et de la forêt prépare et met en œuvre la politique du Gouvernement dans le domaine de l'agriculture, de la forêt et du bois. Il ... +

SUIVRE

Informations

- Licence Ouverte / Open Licence
- 03/2017 à 12/2020
- Hebdomadaire
- 1 juillet 2015
- 3 avril 2017
- 3 avril 2017
- POI

CONTROLE CONTROLE-SANIT...

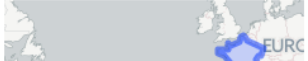
HYGIENE RESTAURANT RESTAURATION

RESULTATS SANITAIRE

SUGGERER UN MOT-CLÉ

DÉTAILS

Couverture spatiale



An example of reuse: Alim'confiance


Résultats des contrôles officiels sanitaires : dispositif d'information « Alim'confiance » - Data.gouv.fr - Mozilla Firefox

Résultats des contrôle x Produits alimentaires x Découpage administré x Jeux de données - Dat x Carte des contrôles sa x +


Gouvernement Premier Ministre (FR) | https://www.data.gouv.fr/fr/datasets/resultats-des-contrôles-officiels-s | Search


+ Nouvelle ressource communautaire

RÉUTILISATIONS





Carte des contrôles sanitaires dans les restaurants et métiers...

 **Chroniques Cartographiques**
5 avril 2017




Alim'confiance


 **Ministère de l'Agriculture, de l'Agroalimentaire et de la Fo**
13 mars 2017




Alim'Infos

 **Christophe Fernandes**
4 octobre 2015

Vous avez réutilisé ces données et publié un article, une infographie, ou une application ? C'est le moment de vous faire connaître ! Référez votre travail en quelques clics et augmentez votre visibilité.



Passage en revue des notations, par les services ministériels ou...


 **Romain Tales**
2 juillet 2015

+
Ajouter une réutilisation

L'OPEN DATA
En tant que citoyen

THÉMATIQUES
Agriculture et alimentation

RÉSEAU
Gouvernement.fr

CONTACT


etalab^{gouv.fr}

- 25715 datasets
- 1260 organisations
- 1690 reuses

Collecting reliable data

Collecting data from data.gouv

There are dataset, producer, reuse catalogs... and an API!

Collecting data from data.gouv

There are dataset, producer, reuse catalogs... and an API!

With an http request, we can collect:

- the title of the page
- the name of the producer
- the creation date
- the granularity
- the frequency (time information)
- the tags
- the number of downloads
- a description
- an url to download the files

Collecting data from data.gouv



Metadata collected for 24617 pages

39405 files downloaded

269 Go of data

'Tablizing' everything

An example of file easy to clean

D	E	F	
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
CELE	20107	Bellegarde	Tr

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

'Tablizing' everything

An example of file easy to clean

D	E	F	
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
CELE	20107	Bellegarde	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	2013	<u>Septembre</u>	1,844	1,741	2,237
3		<u>Octobre</u>	2,393	2,471	2,327
4		<u>Novembre</u>	1,949	2,264	1,889
5		<u>Décembre</u>	1,948	1,768	2,197
6		<u>Janvier</u>	1,745	2,187	2,151
7		<u>Février</u>	2,173	1,879	1,778
8		<u>Mars</u>	2,222	1,872	1,712
9		<u>Avril</u>	1,741	2,129	2,167
0		<u>Mai</u>	1,890	2,281	2,291
1		<u>Juin</u>	1,887	1,723	2,081
2		<u>Juillet</u>	2,298	2,337	2,370
3		<u>Août</u>	2,441	2,426	2,008
4		<u>Septembre</u>	1,818	2,138	2,327
5		<u>Octobre</u>	2,428	2,385	2,333
6		<u>Novembre</u>	1,830	1,834	1,868
7		<u>Décembre</u>	1,871	2,220	2,284
8					
9	Source : Insee, statistiques de l'état civil				
0					
1					

'Tablizing' everything

Cleaning steps

- Is it a zipfile?
- Encoding and extension detections (`chardet` and `magic` libraries)
- Is it a json? A geojson?
- Different extensions (csv, pdf, json, etc.), different strategies

CSV

- Detect the delimiter from a sample (`from csv import Sniffer`)
- Detect the header

CSV

- Detect the delimiter from a sample (`from csv import Sniffer`)
- Detect the header

```
# we test if the first row could be a header
def is_header(file):
    # we test the consistency of the types over the rows

# we test the first rows of the file
for row in file:
    is_header(row)
```

Excel

	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	2013	<u>Septembre</u>	1,844	1,741	2,237
3		<u>Octobre</u>	2,393	2,471	2,327
4		<u>Novembre</u>	1,949	2,264	1,889
5		<u>Décembre</u>	1,948	1,768	2,197
6		<u>Janvier</u>	1,745	2,187	2,151
7		<u>Février</u>	2,173	1,879	1,778
8		<u>Mars</u>	2,222	1,872	1,712
9		<u>Avril</u>	1,741	2,129	2,167
0		<u>Mai</u>	1,890	2,281	2,291
1		<u>Juin</u>	1,887	1,723	2,081
2		<u>Juillet</u>	2,298	2,337	2,370
3		<u>Août</u>	2,441	2,426	2,008
4		<u>Septembre</u>	1,818	2,138	2,327
5		<u>Octobre</u>	2,428	2,385	2,333
6		<u>Novembre</u>	1,830	1,834	1,868
7		<u>Décembre</u>	1,871	2,220	2,284
8					
9	Source : Insee, statistiques de l'état civil				
0					
1					

Excel

- Detect the number of columns
- Fill in the merged cells
- Detect a multiheader

Json

- Explore the json (recursive function)
- Flatten the json (`from pandas.io.json import json_normalize`)

Json

- Explore the json (recursive function)
- Flatten the json (from `pandas.io.json import json_normalize`)

```
# we explore the json until we find a good structure to flatten
def recursive(json):
    if ...
        # we test if the json is a list of dictionary
    else:
        # we keep searching that structure deeper in the json

# we flatten the json
right_structure = recursive(json)
df = json_normalize(right_structure)
```

Json

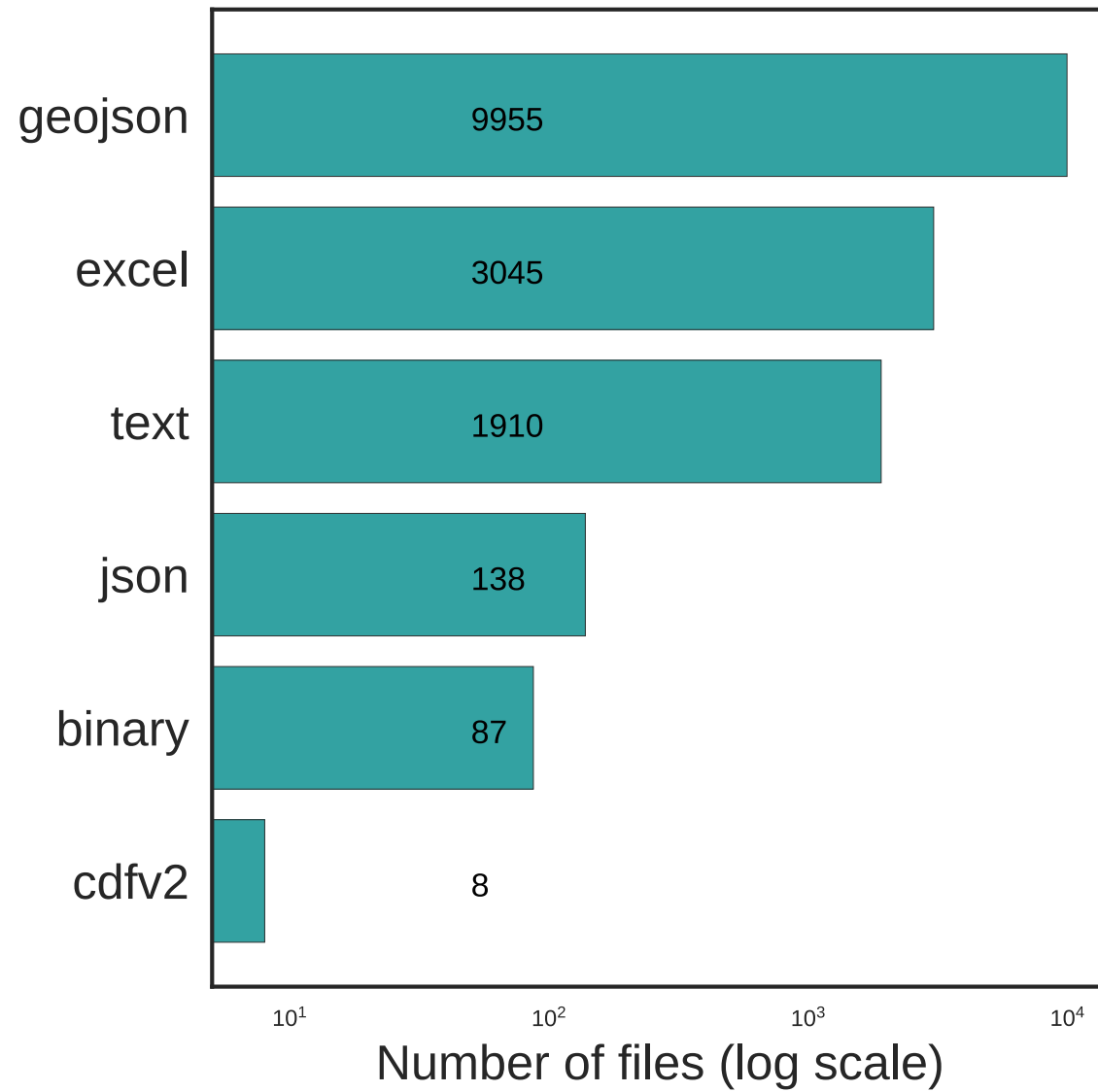
- Explore the json (recursive function)
- Flatten the json (from `pandas.io.json import json_normalize`)

```
# we explore the json until we find a good structure to flatten
def recursive(json):
    if ...
        # we test if the json is a list of dictionary
    else:
        # we keep searching that structure deeper in the json

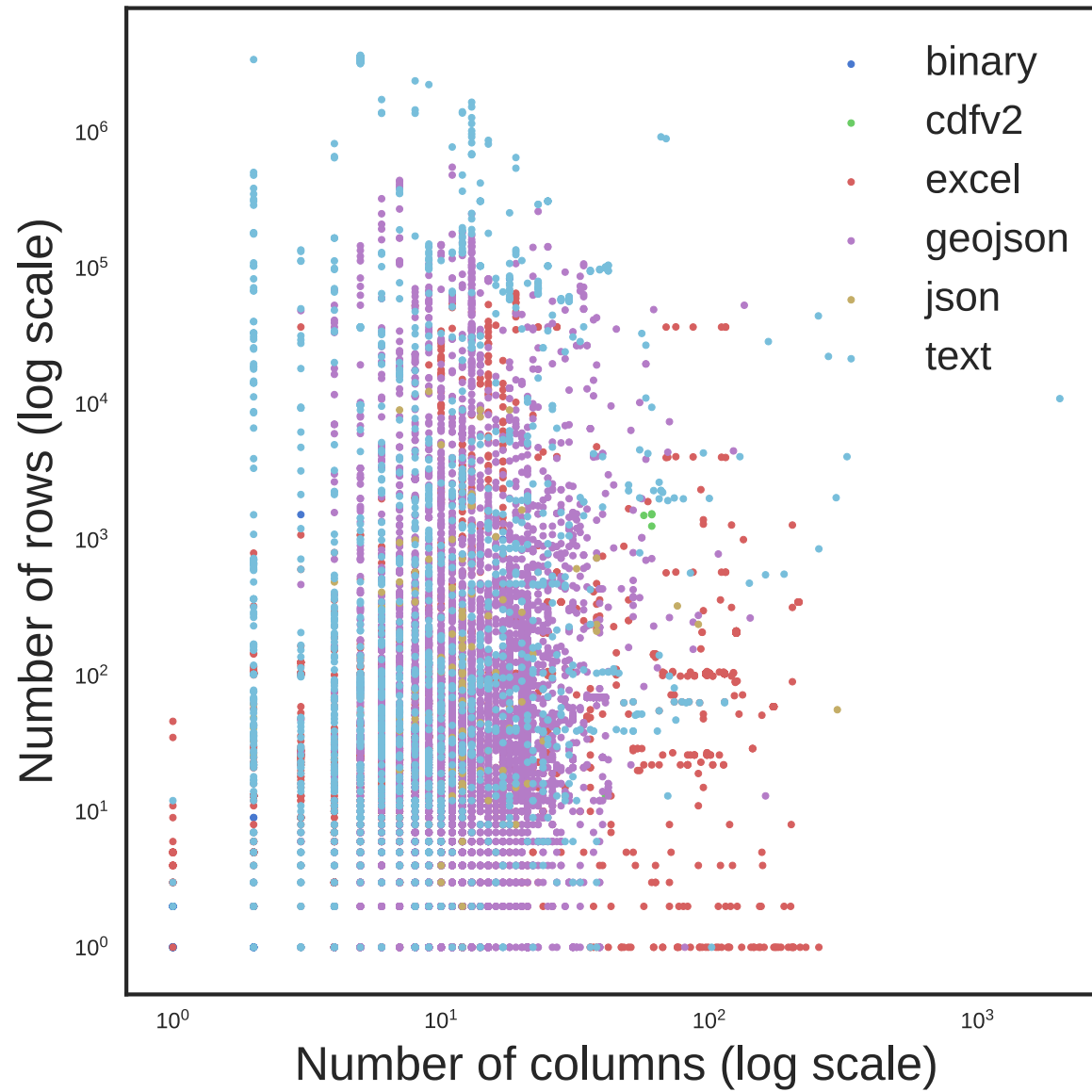
# we flatten the json
right_structure = recursive(json)
df = json_normalize(right_structure)
```

- What about XML?

Results



Results



Learning semantic structure

Preprocessing

- Tokenization

Count words which contain characters only

=> Count_matrix [n_files, n_words]

Preprocessing

- Tokenization
- Normalization

Weight each file to compensate for varying file sizes

Preprocessing

- Tokenization
- Normalization
- Content, header and metadata

$$\text{total} = 0.5 \text{ content} + 0.25 \text{ header} + 0.25 \text{ metadata}$$

Preprocessing

- Tokenization
- Normalization
- Content, header and metadata
- Stemming and unstemming

Reduce each word to its root form:

continuer -> continu

continuant -> continu

Preprocessing

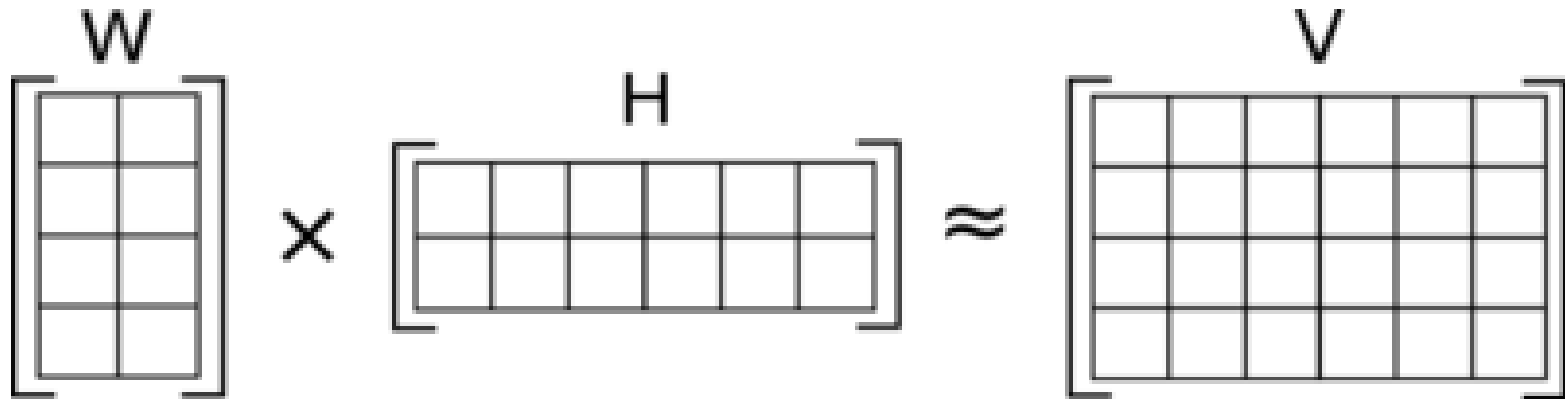
- Tokenization
- Normalization
- Content, header and metadata
- Stemming and unstemming
- TFIDF

Weight each word frequency by its inverse document frequency

=> A word relatively frequent in a specific file will be discriminant

Topic modeling

NMF



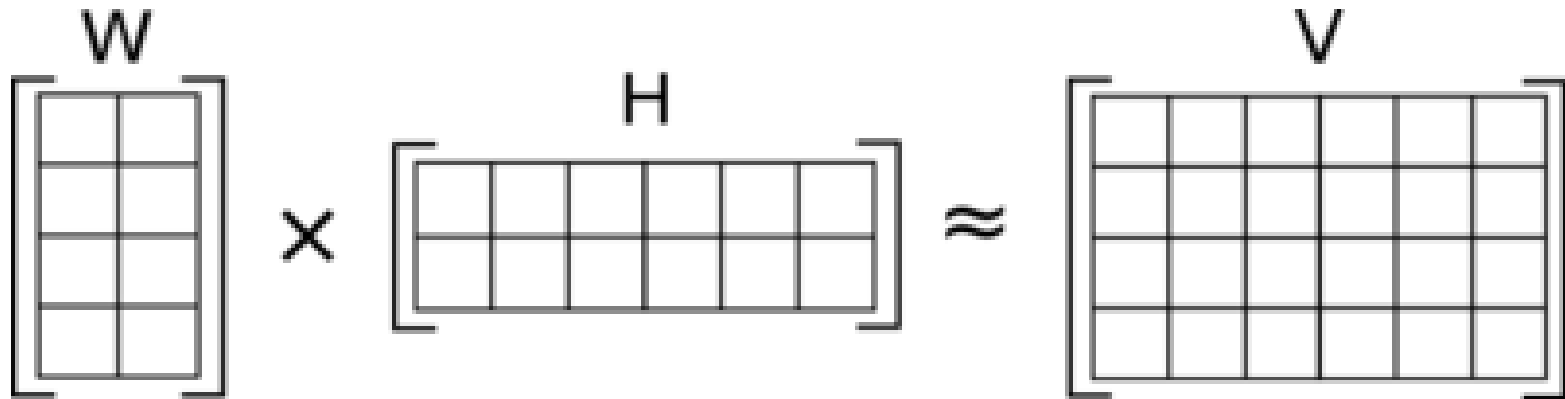
V is a TFIDF matrix [n_files, n_words]

W is a matrix [n_files, n_topics] => files embedding

H is a matrix [n_topics, n_words] => wordcloud per topic

Topic modeling

NMF



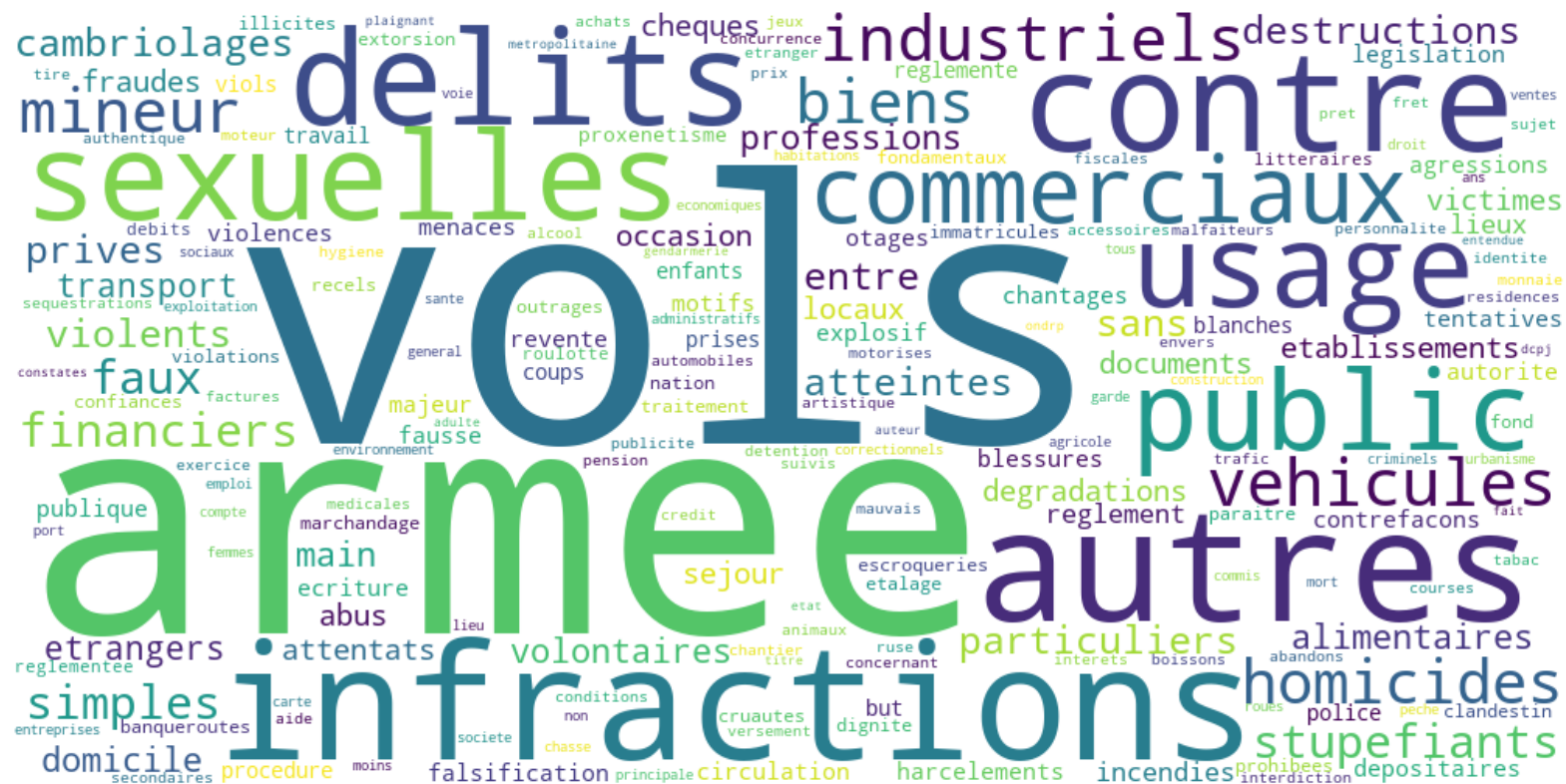
V is a TFIDF matrix [n_files, n_words]

W is a matrix [n_files, n_topics] => files embedding

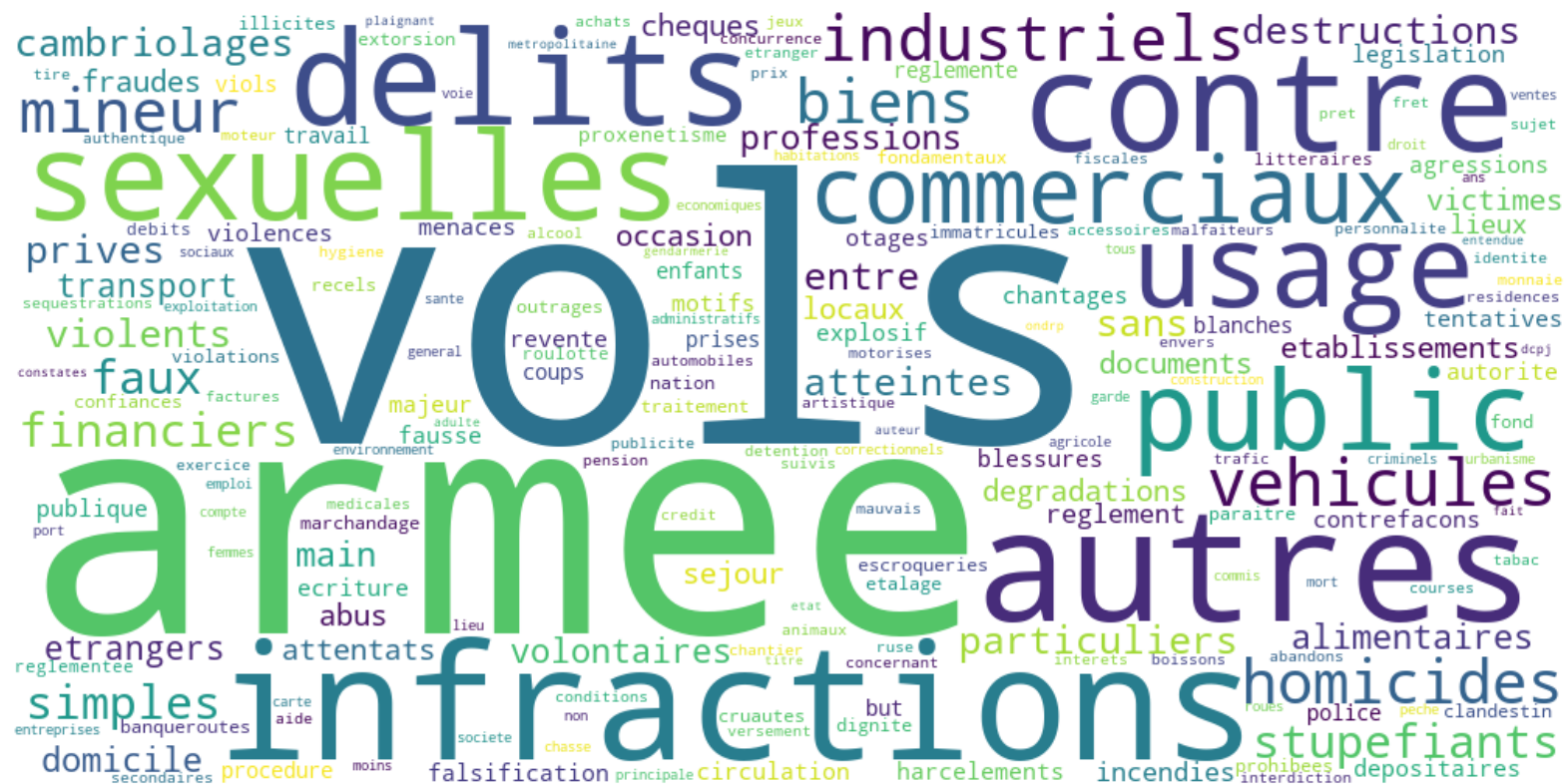
H is a matrix [n_topics, n_words] => wordcloud per topic

n_topics = 20

A reliable embedding ?



A reliable embedding ?

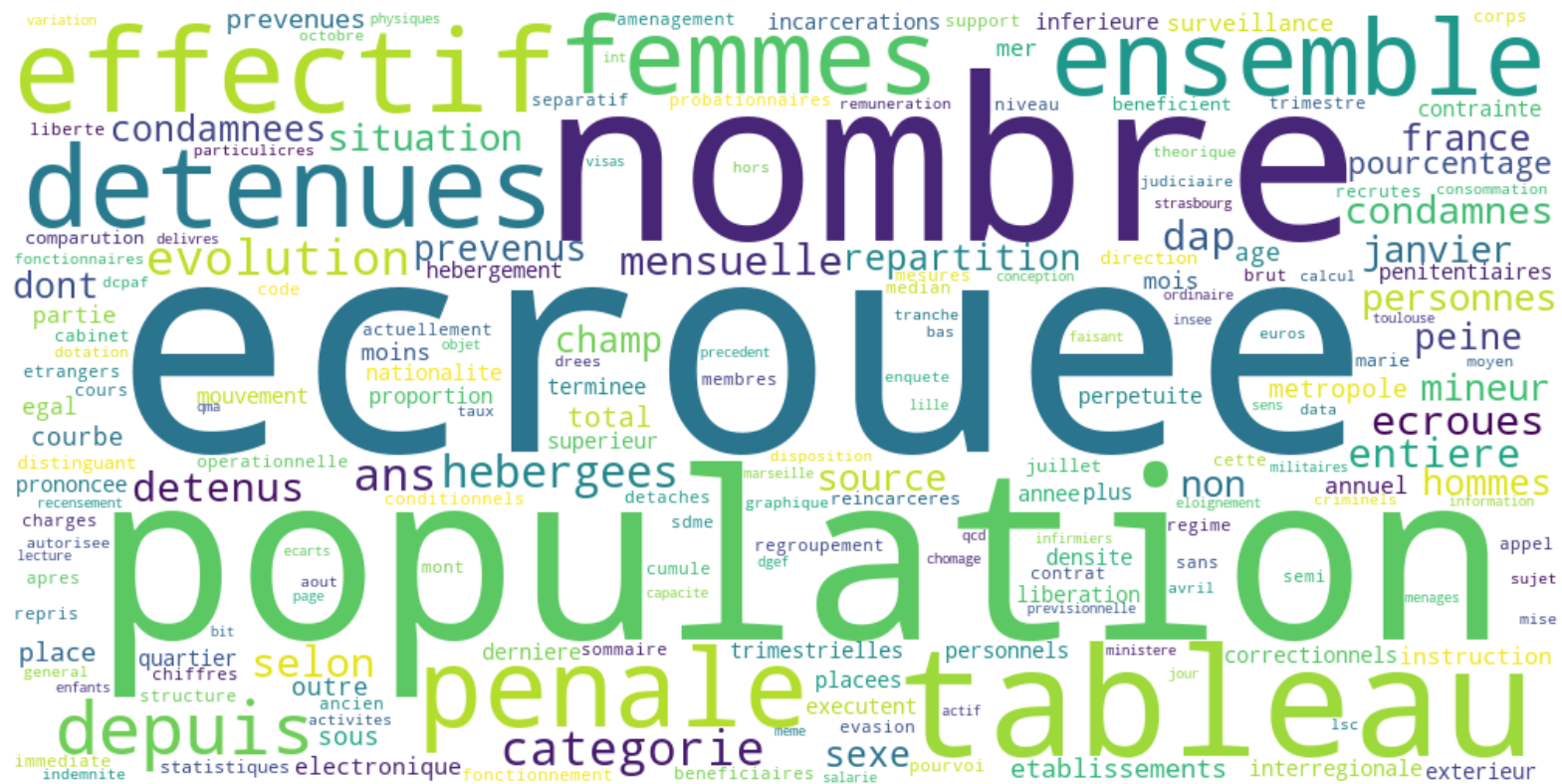


Tags : faits_constates, police, criminalite

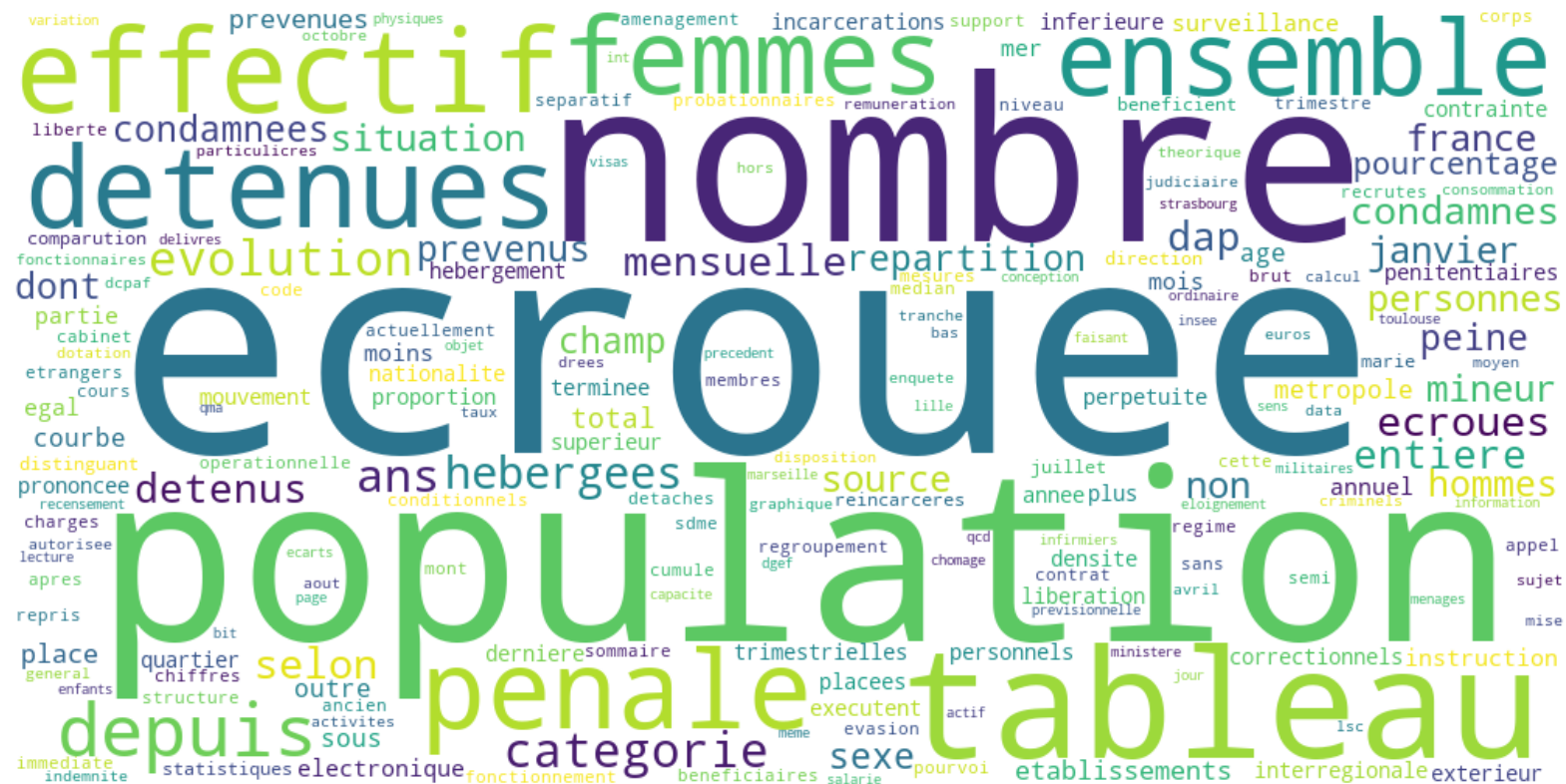
Producer : Observatoire national de la délinquance et des réponses pénales,
Ministère de la Justice, Ministère de l'Intérieur

Extension : excel, geojson, text

A reliable embedding ?



A reliable embedding ?

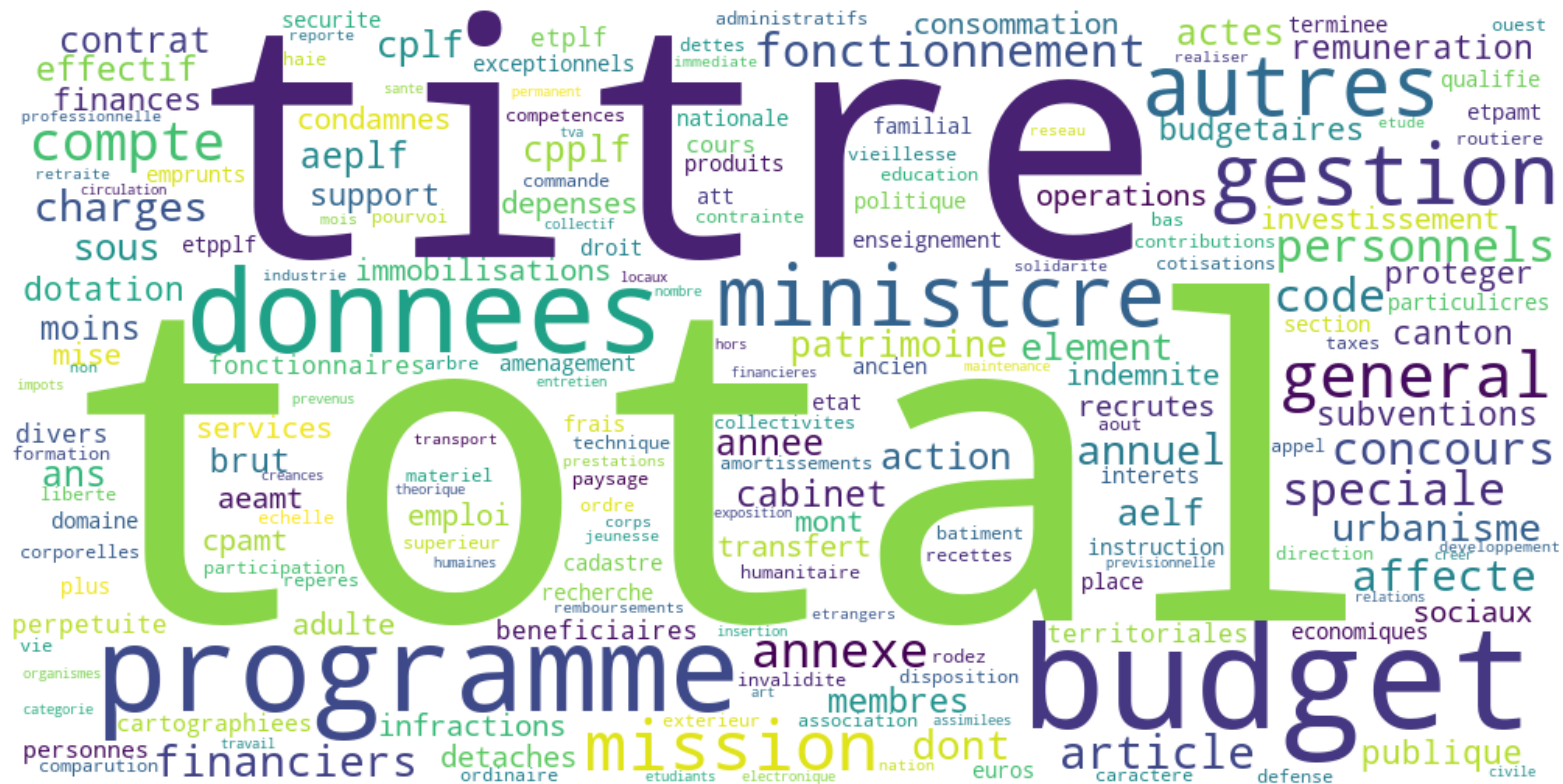


Tags : aménagements de peine, ecroues, immigration

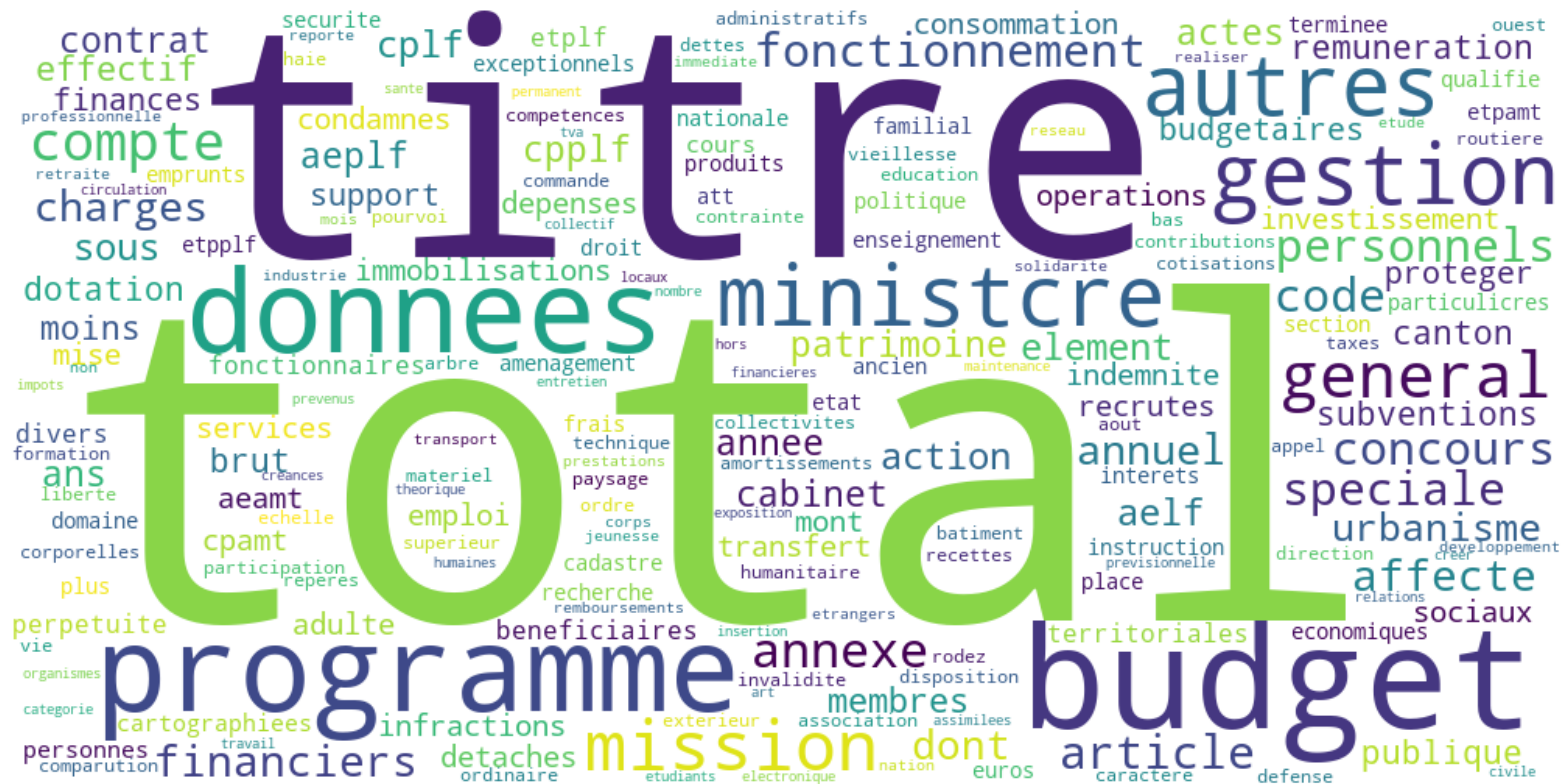
Producer : Ministère de la Justice, Ministère des finances et des comptes publics, Ministère de l'Intérieur

Extension : excel, text, geojson

A reliable embedding ?



A reliable embedding ?

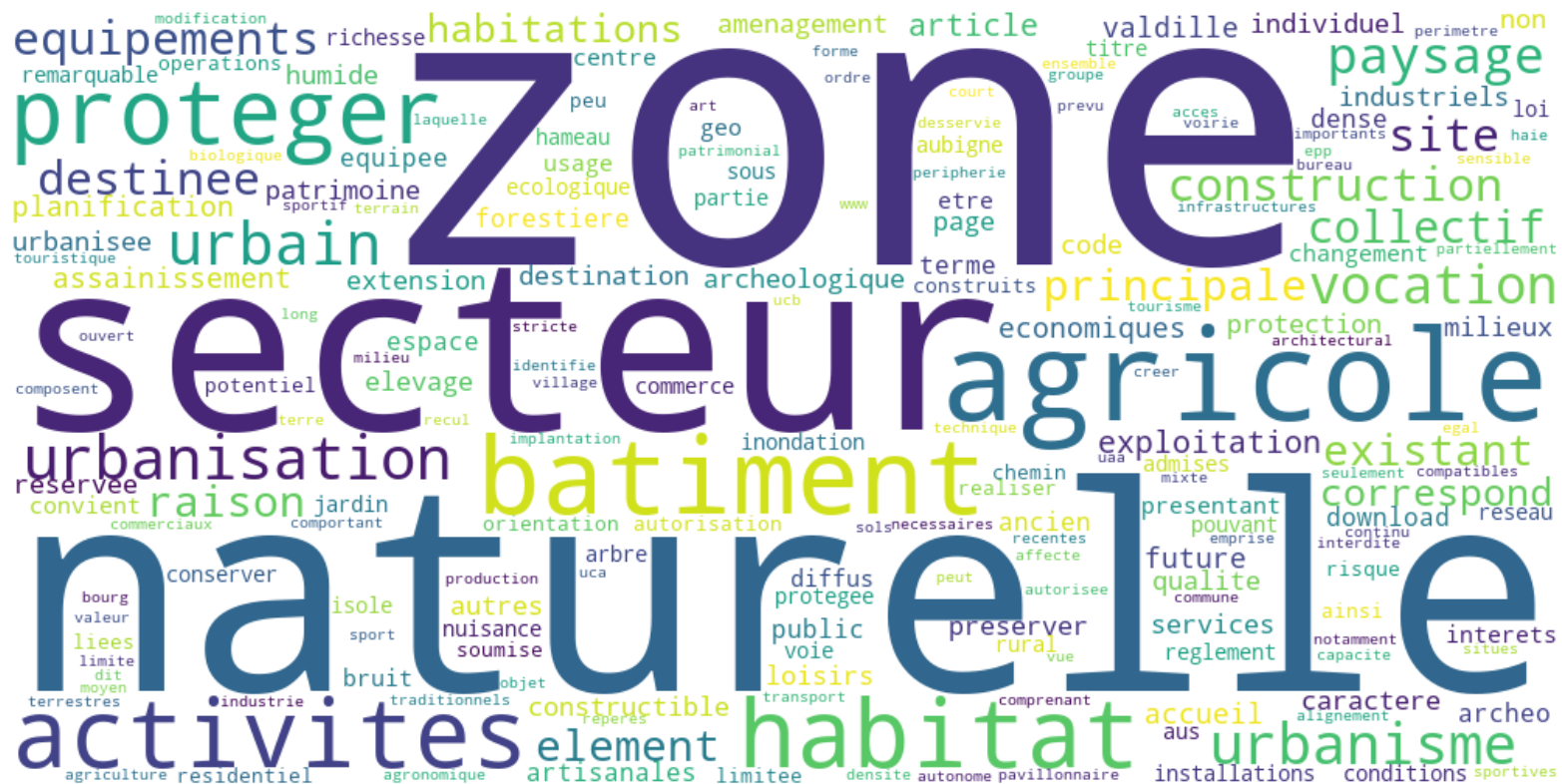


Tags : finances publiques, planning cadastre, usage des sols

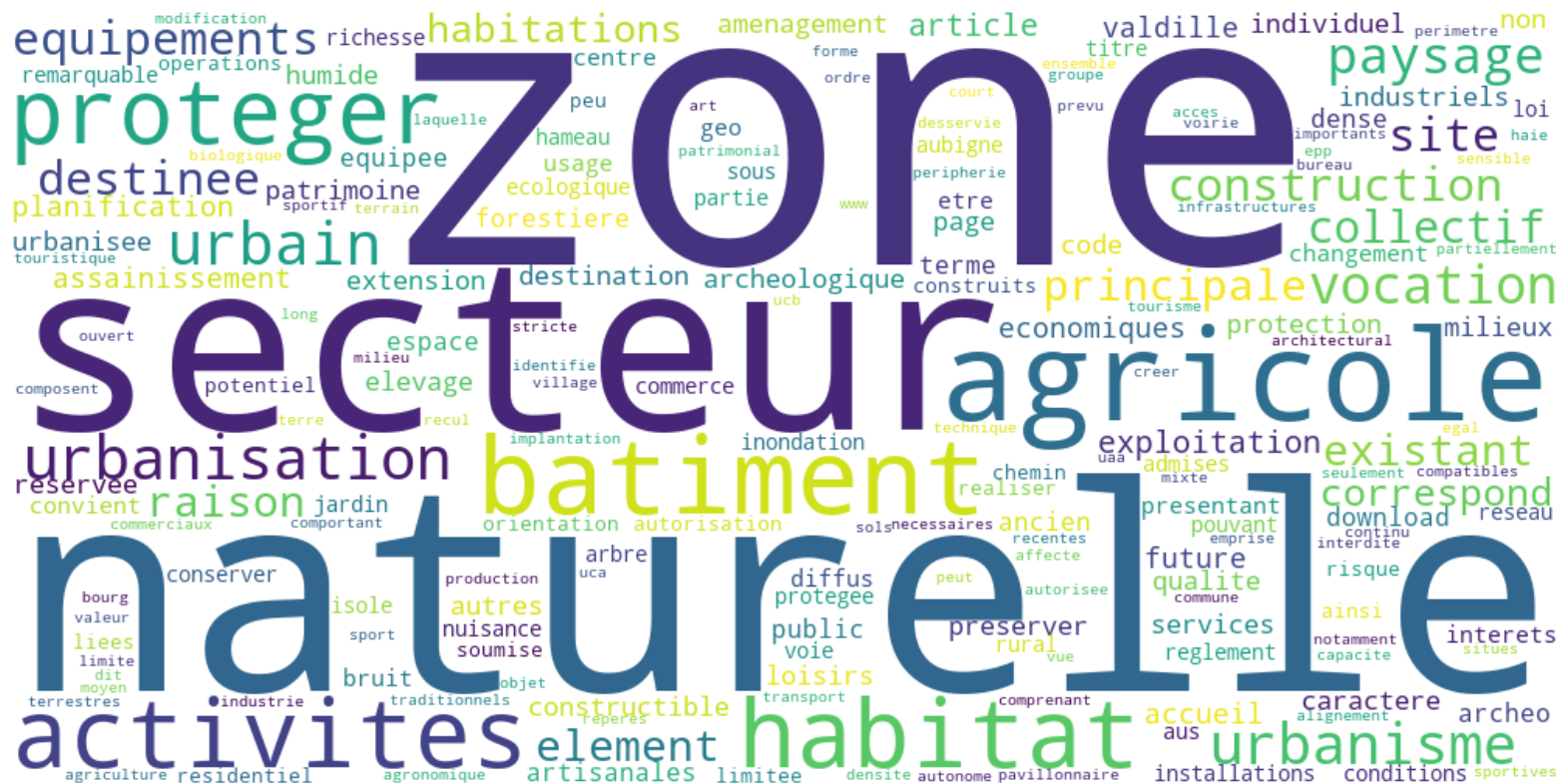
Producer : Ministère des finances et des comptes publics, Ministère de l'Intérieur, Ministère de la Justice

Extension : text, excel, geojson

A reliable embedding ?



A reliable embedding ?



Tags : planning_cadastre, usage_des_sols, plu

Producer : Direction Départementale des Territoires de *

Extension : geojson, excel, text

What are the next steps ?

- Keep working on the embedding
- Find new ways to visualize the relevance of our embedding
- Explore new kind of relations between the files (mutual geographic space, datetime)



**KEEP
CALM
AND
OPEN
DATA**