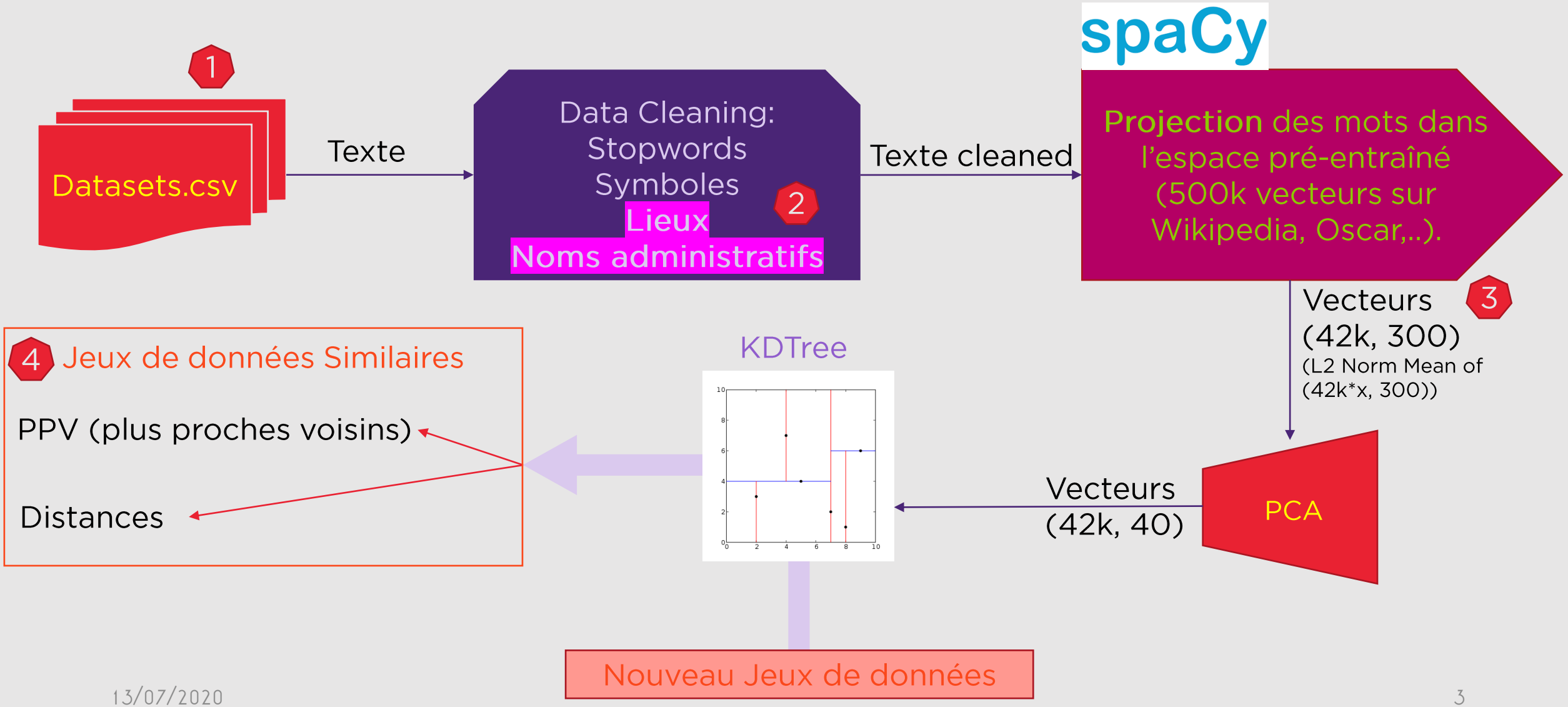
Topic Modelling sur les datasets de data.gouv.fr

Vectorisation et traitement automatique du langage des données contextuelles des jeux de données.

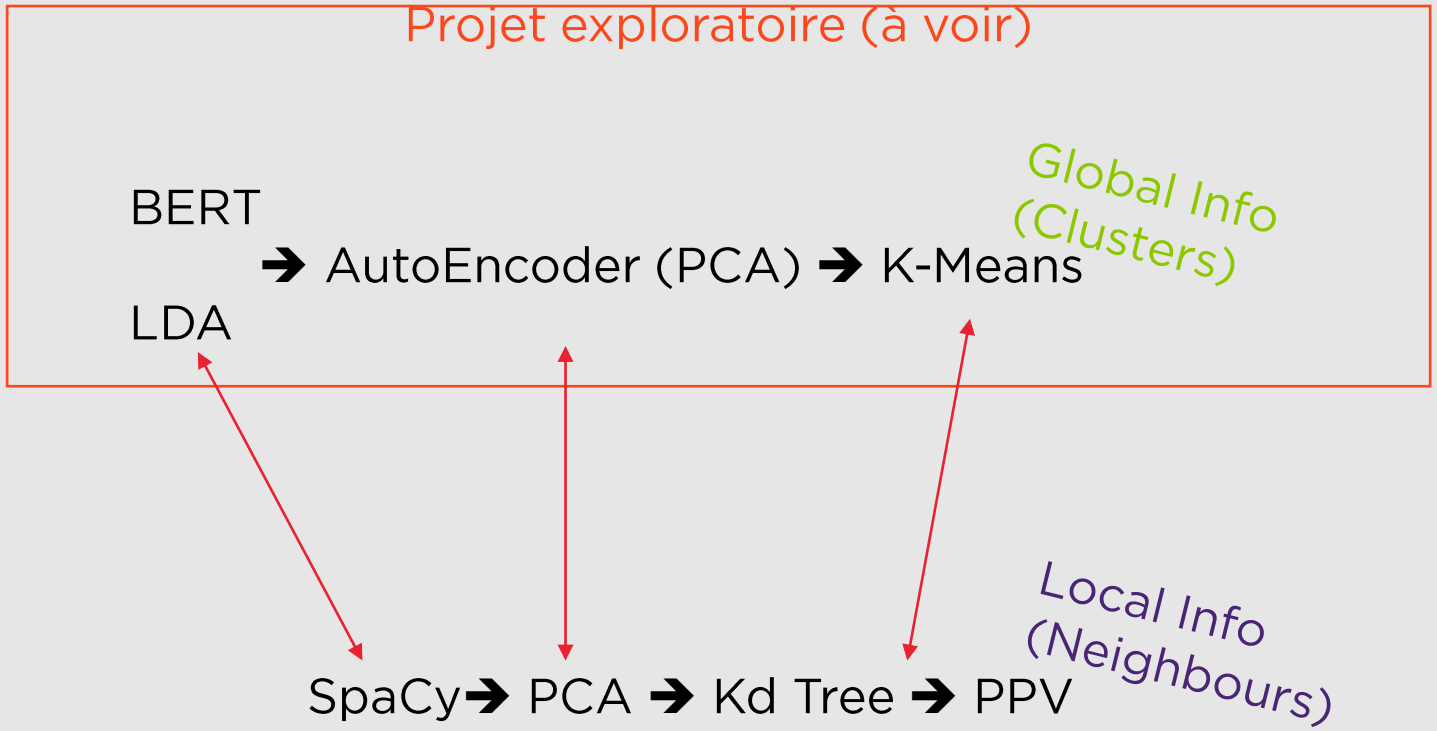
Objectifs de la représentation

- Représentation vectorielle pertinente et réutilisable pour d'autres applications
- Recherche de similarités thématiques
- Adaptable et applicable aux nouvelles données
- Améliorer le référencement des jeux de données
  - Extraction automatique des localisations
  - Propositions automatique de tags basés sur la description

Workflow (1)


1
Datasets.csv
Texte
Data Cleaning:
Stopwords
Symboles
Lieux
Noms administratifs
2
Texte cleaned
spaCy
Projection des mots dans l'espace pré-entraîné (500k vecteurs sur Wikipedia, Oscar,..).
3
Vecteurs
(42k, 300)
(L2 Norm Mean of (42k*x, 300))
PCA
Vecteurs
(42k, 40)
KDTree
4 Jeux de données Similaires
PPV (plus proches voisins)
Distances
Nouveau Jeux de données

Workflow (2)


Projet exploratoire (à voir)
BERT
➔ AutoEncoder (PCA) ➔ K-Means
LDA
Global Info
(Clusters)
SpaCy➔ PCA ➔ Kd Tree ➔ PPV
Local Info
(Neighbours)

Points de détails

1. Texte = Titre + tags
2. Suppression des Lieux et Définitions d’administrations
   - Ex: 'données', 'open', 'data', 'commune’ ,'region', 'departement', 'communes', 'canton', 'publique’
4. Correspondance Index – Id unique

3 Projection (spacy)

- Globalement distribution uniforme, manque de signification
  - Incapacité à clustériser l'ensemble des données par thèmes
- Localement, les textes proches sont effectivement proches
  - Contenu similaires à ~40 voisins significatif
  - Capacité limité pour la recherche (moins performant que le moteur actuel)
  - Indication de la proximité de la similarité

Idées pour la suite

- Predict & Confirm méthode lors de l'upload des fichiers
  - Enrichi la qualité des nouvelles données
  - A moindre coût pour l'utilisateur et à moindre exigence pour l'algorithme
- Extraire les localisation et procéder selon le code INSEE
  - Permet de créer un arbre de dépendance de la situation géographique des datasets.
  - Recherche interactive type leboncoin ?
- Prédire des tags lors de l'upload du fichier

- Métrique plus globale pour inclure géographie, type de donnée, producteur
  - Essayer de dépasser le concepts de filtrage vers une métrique de recherche continue
  - Inclure ou non des dimensions dans l'espace de recherche

- Etude globale au niveau des producteurs de données
  - Résultats négatifs pour l'instant

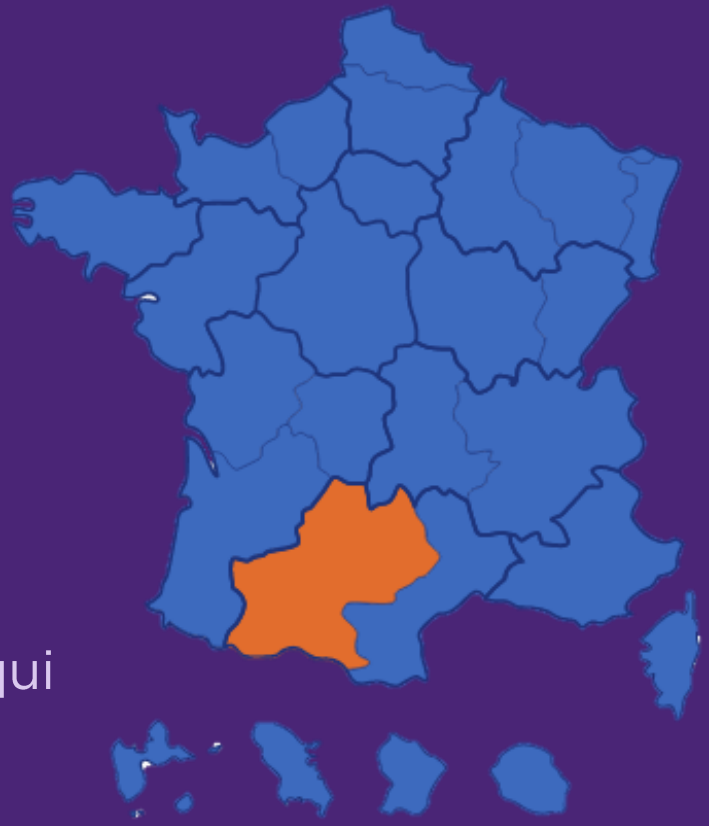
Extraction des localisation

ɔecified			
ʿte / Open Licence version 2.0			
14/05/2014	25/05/2014	other	World
09/07/2018	03/07/2020	fr:commune	CC Mad et Moselle
01/01/2019	31/12/2019	fr:region	
ɔecified			
02/09/2019	04/01/2020		
ɔecified			
10/10/2019	10/10/2020		
ɔecified			
ɔecified			
01/01/2018	31/12/2018	fr:epci	
08/10/2019	08/10/2020		
ɔecified			
ɔecified			
ʿte / Open Licence version 2.0			
01/10/2019	31/12/2029	fr:commune	Ainhoa
ɔecified			
01/01/2019	31/12/2020	other	Francescas
07/10/2019	08/10/2020	other	Metropolitan France
ɔecified			

- 6777 Localisations sur 42858 datasets
- Représente uniquement 15% des datasets de data.gouv.fr
- Format non standardisé

Format code INSEE - user

- Répertorier par communes
  - Sinon, par départements
    - Sinon par régions
      - Sinon par Pays …
- Une commune est incluse dans un département/ région définie
- Permet de faire une recherche interactive par situation géographique
- Ludique, interactif: “Quelles sont les données publique qui concerne mon territoire ?”



Format code INSEE - producter

- Predicted: Hautes Pyrénées.
- Confirmer | Rejeter et corriger: « Pyrénées Atlantiques»

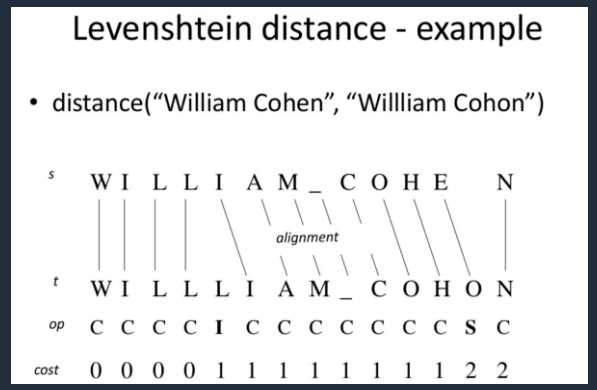
En projet:

- Détection d'entités nommés « Localisation » (package SpaCy)

par débordement de la rivière marne de châlons-en-champagne LOC – secteur amont. Direction Départementale des Territoires de la Marne LOC .

catégories de zones -1 : les zones exposées aux risques et les zones qui ne sont pas directement exposées aux risques mais sur lesquelles des mesures

- Comparaison avec les tables INSEE



	typecom	com	reg	dep	arr	tncc	ncc	nccenr	libelle	can
0	COM	01001	84.0	01	012	5	ABERGEMENT CLEMENCIAT	Abergement-Clémenciat	L'Abergement-Clémenciat	0108
1	COM	01002	84.0	01	011	5	ABERGEMENT DE VAREY	Abergement-de-Varey	L'Abergement-de-Varey	0101
2	COM	01004	84.0	01	011	1	AMBERIEU EN BUGEY	Ambérieu-en-Bugey	Ambérieu-en-Bugey	0101
3	COM	01005	84.0	01	012	1	AMBERIEUX EN DOMBES	Ambérieux-en-Dombes	Ambérieux-en-Dombes	0122

'rivière marne de châlons-en-champagne Direction Départementale des Territoires de la Marne'

- Rule based + robuste (token_set ration & ratio) pour les cas comme :
  - « Bordeaux » et « Artigue-près-bordeaux »
  - « Fort-de-France » et « France »
  - « Val de drôme » et « Drôme » (département)

En projet:

- Exemples

```
TEXTE INITIAL: listing descriptif des producteurs de données ouvertes en aquitaine. Ressourcerie datalocale.
fférents producteurs aquitain. liste opendata

Localisation INSEE:
    reg cheflieu  tncc                 ncc              nccenr  \
13   75    33063     3  NOUVELLE AQUITAINE  Nouvelle-Aquitaine

               libelle
13  Nouvelle-Aquitaine
```

```
TEXTE INITIAL: cantons du doubs. Département du Doubs. les cantons du département du doubs conformément la loi du 17 mai .
ues gouvernement

Localisation INSEE:
   dep  reg cheflieu  tncc    ncc nccenr libelle
23  25   27    25056     2  DOUBS  Doubs   Doubs
```

- A rajouter 3 colonnes dans la description des fichiers avant de pouvoir interfacer une carte de France graphiquement
- A ajouter au processing des nouveau fichiers uploadés sur data.gouv.fr