

# Stage Open Data Science

---

Florian LABORDE

Direction interministérielle du numérique  
Etalab



**PREMIER  
MINISTRE**

*Liberté  
Égalité  
Fraternité*

Direction  
interministérielle  
du numérique

À propos de moi

---

## Florian Laborde

- Elève-Ingénieur Telecom Paris en année de césure 3A d'école
- Master 2 Mathématiques Vision Apprentissage ENS Paris-Saclay
- Intérêt pour la fonction publique et le rôle de l'État dans le numérique

## Stage

- Pôle Open Data : Tam Kien Duong
- Pôle Lab IA : Pavel Soriano
- Travail sur le site [data.gouv.fr](https://data.gouv.fr)

# Open Data Science

---

**Montrer comment les data sciences peuvent contribuer à améliorer l'action publique en matière de circulation de la donnée publique**

- La donnée doit être active, vivante
- Qu'est-ce qui va la rendre prête à fournir des services ?
- Rendre la donnée disponible et accessible

## Idées directrices

- Faire un état des lieux des jeux de données ouverts
- Comment mieux comprendre les jeux de données
- Avoir une meilleure interaction utilisateur avec les jdd
- Enrichir les jdd

## Concrètement

- Trouver une représentation, **vectoriser** les jeux de données
- Trouver des **thèmes**, des manières de structurer automatiquement les données entre-elles
- Proposer des jdd similaires et améliorer la **recherche** de jdd
- Ajouter des informations structurées (**localisation**) à partir du texte

## Vectorisation

---

## Contexte

Le contexte est constitué des éléments sous forme de texte qui englobe un dataset. Sur data.gouv.fr cela s'articule autour des catégories suivantes :

- *Titre* du jeu de données
- *Description* qui peut être plus ou moins longue, précise et parfois très générale.
- *Nom du Producteur* de données qui peut produire des jeux de données variés.
- *Tags* plus ou moins précis et pertinents
- Date, Localisation et autres filtres

## Données

- *Fichiers* à différents formats : CSV, JSON, ZIP ... qui contiennent les données à proprement dites.

# Exemple d'un jeu de données

## Traitements de données personnelles déclarés à la CNIL depuis le 25 mai 2018

Titre

Ce jeu de données provient d'un service public certifié

Depuis l'entrée en application du règlement général sur la protection des données (RGPD), le 25 mai 2018, seuls les traitements numériques de données personnelles les plus sensibles doivent faire l'objet de formalités préalables auprès de la CNIL.

Ces formalités peuvent revêtir la forme de déclarations simplifiées (déclarations de conformité à un cadre de référence proposé par la CNIL), de demandes d'avis (pour les activités régaliennes de l'État) ou de demandes d'autorisation (dans le domaine de la santé). Pour en savoir plus : [cnil.fr](https://cnil.fr).

Conformément à la loi informatique et Libertés modifiée (article 36), la CNIL tient à la disposition du public la liste de ces formalités dans un format ouvert et aisément réutilisable, dite "Liste article 36".

### Avertissements :

1/ Les données publiées sont issues des formalités préalables accomplies, depuis le 25 mai 2018, par les responsables de traitements de données à caractère personnel auprès de la CNIL, via ses téléservices dédiés. La CNIL ne peut être tenue pour responsable de leur contenu.

2/ Des traitements mis en œuvre pour le compte de l'État peuvent ne pas apparaître dans le jeu de données, les formalités ayant été accomplies sous forme de demandes d'avis sur un projet d'acte réglementaire (décret ou arrêté) non soumis via les téléservices mentionnés. L'information relative à ces traitements est disponible sur Legifrance, l'avis de la CNIL étant publié avec l'acte autorisant le traitement (pour accéder aux délibérations de la CNIL : <https://www.legifrance.gouv.fr/initRechExpCnil.do>). Par ailleurs, certains traitements importants font l'objet de [fiches sur le site de la CNIL](#).

3/ Les traitements exceptionnellement dispensés de la publication de l'acte réglementaire qui les autorise (décret ou arrêté) ne figurent pas dans le jeu de données publié, conformément à l'article 36 de la loi informatique et Libertés modifiée. Les traitements mentionnés au I et au II de l'article 30 peuvent être dispensés, par décret en Conseil d'État, de la publication de l'acte réglementaire qui les autorise. Ces traitements sont mentionnés dans le [décret n°2007-914 du 15 mai 2007](#).

Description

### Ressources

#### Formalités préalables reçues par la CNIL depuis le 25 mai 2018

Séparateur " ; " Champs : Organisme raison sociale ; Organisme nom du service ; Organisme adresse ; Organisme code postal ; Organisme ville ; Organisme SIREN ; Service chargé du...

CSV 1 Téléchargeable

PRÉVISUALISER

TÉLÉCHARGER

Producteur



**CNIL**  
COMMISSION NATIONALE  
INFORMATIQUE & LIBERTÉS

Nom de l'organisation / Producteur de données **CNIL**

La Commission nationale de l'informatique et des libertés est l'autorité administrative indépendante en charge de la protection des données personnelles en France. En tant que...

VOIR LE PROFIL

CONTACTER

SUIVRE

### Informations

© Licence Ouverte / Open Licence version 2.0

25/05/2018 à 25/05/2018

Hebdomadaire

25 juin 2020

6 juillet 2020

25 juin 2020

Pays Localisation

cnil déclarations formalités informatique...  
liste article 36 rgpd télécharger un mot-clé

DÉTAILS

Tags

Figure 1: Extrait d'une page d'un jeu de données sur data.gouv.fr



## Tâches auxquelles répondent les techniques de NLP

- Repérer les textes semblables
- Détecter des types de mots (Entités Nommées : Lieux, Noms propres)
- Extraire les thèmes principaux d'un paragraphe
- Comprendre sémantiquement des phrases

## Différentes Méthodes

- Méthodes statistiques
- A l'aide des réseaux de neurones

### **Nettoyer et extraire les mots principaux**

- Eliminer les mots de liaison
- Supprimer les terminaisons, 'lemmatizer'
- Réduction aux mots-clés principaux, dans leur forme de base, issues du texte.

## Vectoriser

- Reperer les fréquences de certains mots, compter (TF-IDF)
- Traduire tous les mots en un seul vecteur rapporté à un nombre fini d'idées simples

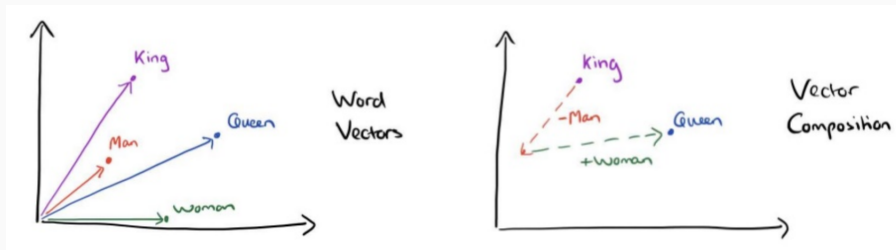
## 'Embedding' d'un mot dans un vecteur de dimension finie

- L'esprit humain associe une infinité de sens et de liens à un mot.
- Un vecteur est de dimension finie
- Chaque dimension d'un vecteur est une idée pure élémentaire, qu'on ne sait pas nommée
- On essaie de rendre à chaque mot une représentation dans un espace d'un nombre fini d'idées simples
- Placer les mots à un seul endroit, logique les uns par rapport aux autres

## Une nécessité matérielle

- Traitement Informatique
- Limitations de mémoire et de capacité de calcul des ordinateurs

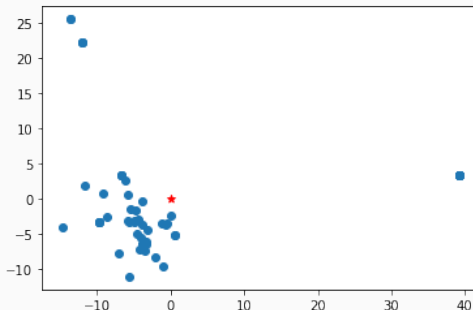
Toute la logique et tous les outils mathématiques sont applicables



**Figure 2:** Intérêt et puissance de la représentation vectorielle dans un embedding

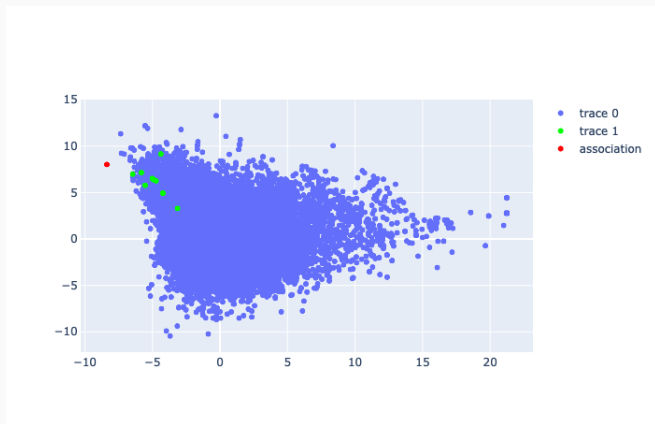
Chaque mot est représenté par un vecteur ([0.1, 0.05, .., 0.8]):

- Pour le représenter on peut sélectionner les deux dimensions les plus discriminantes.
- Résultat de l'entrainement d'algorithme sur des milliers de lignes de texte issues de wikipédia.



**Figure 3:** Représentation vectorielle d'un dataset après réduction à 2 dimensions (initialement 300). Chaque point est un mot. Le point rouge est la moyenne de ces mots et représente le dataset en entier.

Maintenant qu'un point représente un jeu de données (plusieurs mots), on peut comparer les jeux de données entre eux, à partir d'une requête.



**Figure 4:** Chaque point bleu est un jeu de données, le point rouge est le mot 'association' et les points verts sont les 10 jdd les plus proches du mot 'association' (par rapport à la moyenne de leurs mots)

## Regroupements automatiques de jeux de données (LDA)

---

### **Pour avoir un aperçu des jdd disponibles**

- Quels mots-clés ressortent le plus souvent ?
- Quels sujets sont abordés ?
- Y-a-t-il des thèmes non catégorisés ? Négligés ?

### **Pour comprendre les producteurs de données**

- Qui sont les acteurs privés de l'open-data ?
- Quelles données fournissent-ils ?



## Information et paramètres des thèmes.

- *Nombre de thèmes* doit être choisi. Permet d'affiner la séparation des thématiques.
- *Proximité* entre les différents thèmes créés permet de corriger éventuellement les erreurs
- Certains *mots-clés* appartiennent à plusieurs thèmes
- Importance *relative* d'un thème par rapport à l'autre

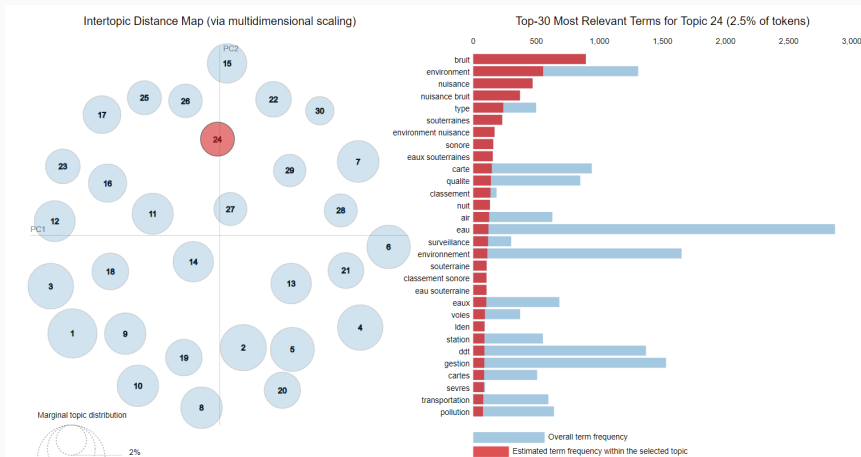
## Problèmes

- *Mots-clés bruités* les thèmes ne font pas toujours sens
- *Thèmes non choisis*. Ne permet pas d'imposer un ensemble de thèmes
- Pas déterministe. Résultats variables. Variance entre les thèmes détectés

---

<sup>1</sup>Latent Dirichlet Allocation

# Exemple d'une thématisation automatique



**Figure 5:** Résultat de l'algorithme LDA (vectorisation TF-IDF, Représentation 2D: MMDS, Thèmes: 30, Source: texte sans description)

## Compréhension sémantique des phrases et textes longs

---

**Au lieu de rassembler la signification des mots indépendamment les uns des autres, extraire directement le sens de la phrase.**

- Compréhension beaucoup plus fine de la langue.
- Mieux gérer les synonymes, les paraphrases et les liens logiques dans le texte.
- La brique technologique majeure de ces dernières années (2018-2019) est l'*attention layer* et les *transformers networks*. En France, ces modèles de réseaux de neurones sont connus sous le nom de CamemBERT et FlauBERT.

# Exemple Comparé 1

Quelles sont les piscines de la ville de Toulouse ? : *piscines toulouse*

ElasticSearch – mots-clés	SBERT – mots-clés	SBERT - phrase
<p>🔍 piscines toulouse</p> <p>Trier par <b>PERTINENCE</b> ⌵</p>	<p>🔍 piscines toulouse</p> <p>Trier par <b>PERTINENCE</b> ⌵</p>	<p>🔍 Quelles sont les piscines de la ville de Toulouse ?</p> <p>Trier par <b>PERTINENCE</b> ⌵</p>
<p> <b>05 Station météo Toulouse Nakache</b> Ce jeu de données est issu du capteur n° 5 situé proche du site de la piscine Alfred Nakache (sur ○ Inconnu ⚡ 0 ★ 0</p>	<p> <b>Piscines - Toulouse</b> Localisation des piscines municipales sur la commune de Toulouse avec notamment ○ Inconnu ⚡ 0 ★ 0</p>	<p> <b>Piscines - Toulouse</b> Localisation des piscines municipales sur la commune de Toulouse avec notamment ○ Inconnu ⚡ 0 ★ 0</p>
<p> <b>53 Station météo Toulouse Ponsan</b> Ce jeu de données est issu du capteur n° 53 situé sur le site de la Piscine Bellevue (Quartier ○ Inconnu ⚡ 0 ★ 0</p>	<p> <b>Piscines municipales</b> Localisation des piscines municipales sur la commune de Toulouse avec notamment ○ Hebdomadaire ♀ Toulouse ♂ Autre ⚡ 0 ★ 0</p>	<p> <b>Piscines municipales</b> Localisation des piscines municipales sur la commune de Toulouse avec notamment ○ Hebdomadaire ♀ Toulouse ♂ Autre ⚡ 0 ★ 0</p>
<p> <b>Piscines - Toulouse</b> Localisation des piscines municipales sur la commune de Toulouse avec notamment ○ Inconnu ⚡ 0 ★ 0</p>	<p> <b>Organisations</b> ⌵ Mairie de Toulouse 1 Toulouse métropole 1</p>	<p> <b>Organisations</b> ⌵ Mairie de Toulouse 1 Toulouse métropole 1</p>
<p> <b>Piscines municipales</b></p>		

Figure 6: Exemple 'Piscines'

# Exemple Comparé 2

ElasticSearch – mots-clés	SBERT – mots-clés	SBERT - phrase
<input type="text" value="Vacances Scolaires"/>	<input type="text" value="Quand est-ce que les enfants sont en vacances ?"/>	<input type="text" value="Quand est-ce que les enfants sont en vacances ?"/>
Trier par <span>PERTINENCE</span>	Trier par <span>PERTINENCE</span>	Trier par <span>PERTINENCE</span>
<div><p><b>Vacances scolaires par zones</b> Contient les vacances scolaires des zones A, B et C en France. Description des zones La</p><p>1990–2021 • Annuelle • France • Autre</p></div> <div><p><b>Simulateur CALENDRIER DES VACANCES ...</b> Le site officiel de l'administration française service-public.fr référence une soixantaine de</p><p>Ponctuelle • France • Département français</p></div> <div><p><b>Le calendrier scolaire – Format iCal</b> Le calendrier scolaire officiel, publié par le Ministère de l'Éducation nationale, de</p><p>2018–2020 • Annuelle • France • Autre</p></div> <div><p><b>Recensement des équipements sportifs, e...</b> Le recensement des équipements sportifs</p></div>	<div><p><b>Scolaire – Etablissement</b> Liste des établissements scolaires</p><p>0 • 0 • 0</p></div> <div><p><b>Vacances scolaires par zones</b> Contient les vacances scolaires des zones A, B et C en France. Description des zones La</p><p>1990–2021 • Annuelle • France • Autre</p></div> <div><p><b>Le calendrier scolaire</b> Le calendrier scolaire officiel, publié par le Ministère de l'Éducation nationale, de</p><p>2018–2020 • Annuelle • France • Autre</p></div> <div><p><b>Le calendrier scolaire – Format iCal</b> Le calendrier scolaire officiel, publié par le</p></div>	<div><p><b>Vacances scolaires par zones</b> Contient les vacances scolaires des zones A, B et C en France. Description des zones La</p><p>1990–2021 • Annuelle • France • Autre</p></div> <div><p><b>Le calendrier scolaire</b> Le calendrier scolaire officiel, publié par le Ministère de l'Éducation nationale, de</p><p>2018–2020 • Annuelle • France • Autre</p></div> <div><p><b>Le calendrier scolaire – Format iCal</b> Le calendrier scolaire officiel, publié par le Ministère de l'Éducation nationale, de</p><p>2018–2020 • Annuelle • France • Autre</p></div>

Figure 7: Exemple 'Vacances Scolaires'

# Exemple Comparé 3

Le coût de l'essence a-t-il augmenté ? : *prix des carburants*

ElasticSearch – mots-clés	SBERT – mots-clés	SBERT - phrase
<p>🔍 prix des carburants</p> <p>Trier par <b>PERTINENCE</b> </p> <p> <b>Prix carburant jour</b> Inconnu 0 0 0</p> <p> <b>Prix des carburants - J-7 en Corse</b> Les informations sont extraites du système d'information « Prix Carburants » à J-7. Inconnu 0 0 0</p> <p> <b>Prix des carburants en France</b> Les données mises à disposition au téléchargement sont les informations extraites Inconnu 0 Point d'Intérêt 19 14</p>	<p>🔍 prix des carburants</p> <p>Trier par <b>PERTINENCE</b> </p> <p> <b>Prix carburant jour</b> Inconnu 0 0 0</p> <p> <b>Prix du pétrole brut</b> Prix du pétrole brut sur les marchés Inconnu 0 0 1</p> <p> <b>Prix des carburants en France</b> Les données mises à disposition au téléchargement sont les informations extraites Inconnu 0 Point d'Intérêt 19 14</p>	<p>🔍 Le coût de l'essence a-t-il augmenté ?</p> <p>Trier par <b>PERTINENCE</b> </p> <p> <b>Prix du pétrole brut</b> Prix du pétrole brut sur les marchés Inconnu 0 0 1</p> <p> <b>Prix des carburants en France</b> Les données mises à disposition au téléchargement sont les informations extraites Inconnu 0 Point d'Intérêt 19 14</p> <p> <b>Prix carburant jour</b> Inconnu 0 0 0</p>

Figure 8: Exemple 'Prix du gasoil'

# Exemple Comparé 4




ElasticSearch – mots-clés	SBERT – mots-clés	SBERT - phrase
<div><input type="text" value="épidémie incidence"/></div> <div>Trier par <span>PERTINENCE</span></div>	<div><input type="text" value="épidémie incidence"/></div> <div>Trier par <span>PERTINENCE</span></div>	<div><input type="text" value="Le virus va-t-il provoquer une deuxième vague ?"/></div> <div>Trier par <span>PERTINENCE</span></div>
<div><b>Indicateurs de suivi de l'épidémie de COVI...</b> Présentation des indicateurs de suivi Le 28 mai 2020, le gouvernement a présenté dans le cadre <span>Inconnu</span> <span>0</span> <span>0</span></div> <div><b>Indicateurs de suivi de l'épidémie de COVI...</b> Présentation des indicateurs de suivi Le 28 mai 2020, le gouvernement a présenté dans le cadre <span>Inconnu</span> <span>0</span> <span>0</span></div> <div><b>Capacité analytique de tests virologiques ...</b> Les actions de Santé publique France Santé publique France pour mission d'améliorer et <span>Inconnu</span> <span>0</span> <span>0</span></div> <div><b>Indicateur Avancé Sanitaire IAS® - INCIDE...</b> Pour la première fois en France, il est possible de suivre au jour le jour l'évolution des <span>2009-2016</span> <span>Quotidienne</span> <span>Région française</span></div>	<div><b>Taux d'incidence de l'épidémie de COVID-1...</b> Les actions de Santé publique France Santé publique France pour mission d'améliorer et <span>Inconnu</span> <span>2</span> <span>3</span></div> <div><b>Mortalité par cause</b> Export : CSV, HTML et XLS Source : www.ecosante.fr l'irdes d'après données CépiDC <span>1979-2011</span> <span>Annuelle</span> <span>Département français</span></div> <div><b>Chiffres-clés concernant l'épidémie de CO...</b> L'information officielle sur la progression de l'épidémie en France est assez fragmentée, et <span>Quotidienne</span> <span>France</span> <span>Autre</span> <span>20</span> <span>16</span></div> <div><b>Estimation d'incidence des syndromes gri...</b> Estimations d'incidence hebdomadaire des syndromes grippaux basées sur les données des <span>1984-2014</span> <span>Hebdomadaire</span> <span>Région française</span></div>	<div><b>Fr-SARS-CoV-2</b> Tentative de classement du nombre de cas confirmés du virus SARS-CoV-2 sur le territoire <span>2020</span> <span>Ponctuelle</span> <span>Région française</span> <span>0</span> <span>0</span></div> <div><b>Taux d'incidence de l'épidémie de COVID-1...</b> Les actions de Santé publique France Santé publique France pour mission d'améliorer et <span>Inconnu</span> <span>2</span> <span>3</span></div> <div><b>Données relatives aux résultats des tests ...</b> Les actions de Santé publique France Santé publique France pour mission d'améliorer et <span>Inconnu</span> <span>17</span> <span>4</span></div> <div><b>Cas confirmés d'infection au COVID-19 pa...</b> Santé Publique France a publié un point quotidien sur cette page indiquant le nombre de <span>2020</span> <span>Quotidienne</span> <span>France</span> <span>Région française</span></div>

Figure 9: Exemple 'Covid-19'

## Exemple du mécanisme d'attention

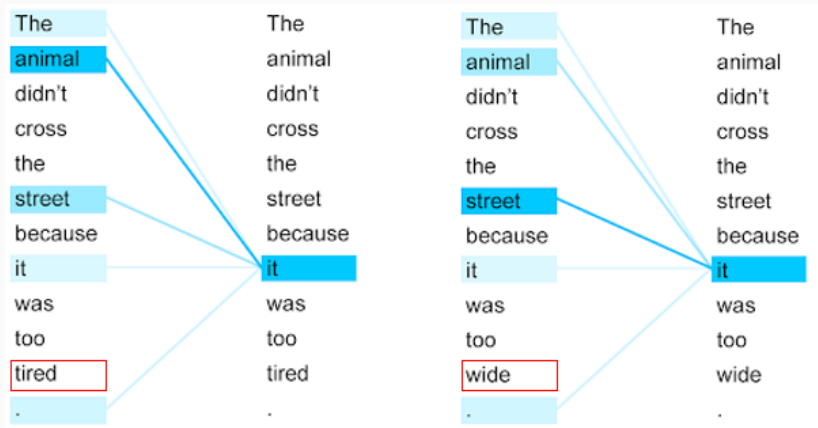


Figure 10: Exemple : Attention Head [3]

Mécanisme d'attention : Chaque *Attention head* est une couleur. Essayez de déterminer si **it** concerne la *rue* ou l'*animal*.



## Jeux de données anglophones

- Wikipedia
- Common Crawl
- BooksCorpus
- SQuAD, SQuAD v2
- SNLI, MultiNLI
- QNLI, CoLA, STS-B, RTE, MRPC, SST-2, QQP .....

## Modèles pré-entraînés [EN]

- BERT
- SBERT
- Multitude de versions fine-tunées

## Jeux de données francophones

- Wikipédia
- Common Crawl (OSCAR, CCNet)
- Piaf [4]

## Modèles pré-entraînés [FR]

- CamemBERT [5]
- FlauBERT [6]
- Très peu de fine-tuning existants

## Pourquoi BERT ne fonctionne pas en recherche de similarité ?

- BERT n'est pas un modèle d'*Embedding*, les poids sur la dernière couche n'ont qu'une signification pour répondre à une certaine tâche.
- BERT est prévu pour produire un résultat et une réponse à un problème sans que l'on puisse réutiliser les étapes intermédiaires de son extraction pour d'autres tâches.
- Une grande partie du fonctionnement de BERT est prévue en dualité avec une seconde phrase
- Sbert propose une architecture différente, une solution, au coût d'une nouvelle étape d'entraînement supervisé...

## Caractéristiques principales

- Retour à une finalité d'embedding : **chaque phrase correspond à un vecteur**
- Changement d'architecture (réseau Siamese)
- Entraînement sur des jeux de données avec une interaction globale entre les phrases
- Pas de jeu de données français pour ré-entraîner l'ensemble

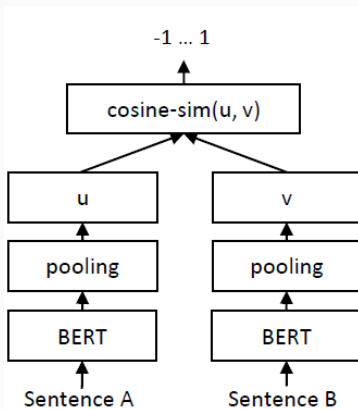


Figure 11: Réseau Siamese de SBERT [1]

## Jeux de données multilingues

- Europarl
- UNPC
- OpenSubtitles
- TED2020

Le principe de l'approche multilingue est d'utiliser l'apprentissage sur des jeux de données annotés monolingues, puis de se servir de textes dont la traduction existe (sans aucune contrainte sur ces textes et sans annotation nécessaire) pour transposer l'apprentissage d'une langue à une autre.

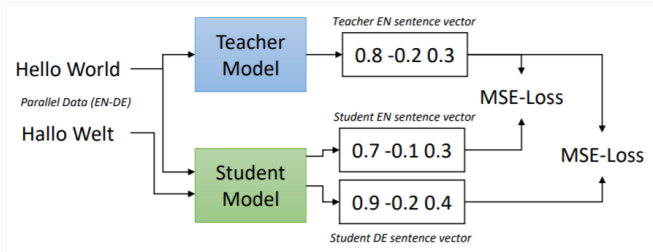


Figure 12: Knowledge Distillation - Teacher/Student [2]

On considère la phrase de recherche comme un document et on retrouve ceux qui sont 'les plus similaires'.

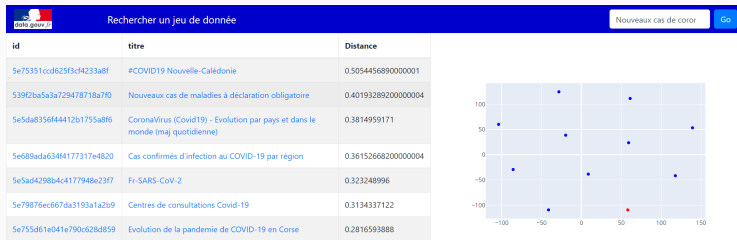


Figure 13: Vue de l'application Dash

Afin d'utiliser au mieux la puissance de SBERT on peut proposer à l'utilisateur un format instinctivement sous forme de phrase, une recherche 'conversationnelle'.

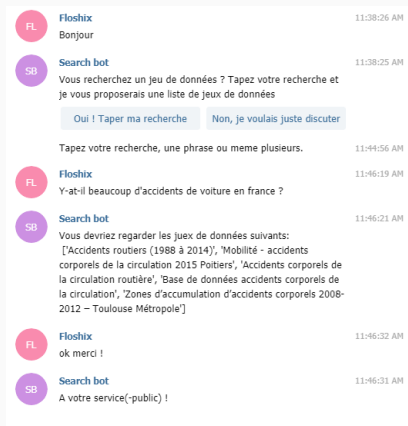


Figure 14: Exemple d'échange avec le chatbot

## Localisation

---

**L'information de localisation est souvent absente de la structure de data.gouv.fr alors qu'elle est facilement accessible textuellement**

"Il faut que les citoyens français sachent ce que l'État fait pour eux chez eux, j'habite dans la commune de Saint-Pol-sur-Ternoise, qu'est ce que le numérique ça veut dire pour moi"

- Madame la Ministre de la Transformation et de la Fonction Publique Amélie De Montchanlin.



## Actuellement

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

## Objectif

- Répertorier à l'échelle la plus fine possible
- Permet de faire une recherche interactive par situation géographique
- Interactif répond à : “Quelles sont les données publique qui concerne mon territoire ?”

## Approche par dataset

- Recherche d'entités nommées avec Spacy (NER) et calcul de distance entre mots.
- Accuracy de 80% (labellisé à la main aléatoirement sur 300 jdd)

## Approche par organisation

- Recherche de subwords étant des lieux dans les noms des organisations.
- Hypothèse : Une organisation avec indication géographique ne publie que des jdd associés à cette zone
- Meilleure précision, accuracy beaucoup plus faible
- Être sûr des indications de localisation que l'on apporte

Pour aller plus loin...

---

### Comment utiliser ces POC

- Ajouter les vecteurs au dictionnaire Elasticsearch
- Ajouter des Localisations (qui ont un fort taux de confiance)
- La recherche de document est une partie importante pour le Q&A, expertise (SBERT) réutilisable pour PIAF

### Continuer à explorer

- Module/Package Python 'OpenDataScience'
- Explorer plus seulement le contexte, mais vectoriser à l'aide du fichier ressource lui-même.
- Étendre PIAF à d'autres types de dataset, pas seulement Q&A ?

## Un avis sur le stage

---

## **Un environnement travail particulier**

- Au rythme des décisions politiques
- Au sein de différentes administrations et ministères

## **Des enjeux uniques**

- Moderniser l'État
- Mise en place directe des politiques publiques et des décisions politiques
- Consultations par les décideurs
- Échelle européenne

## Un rôle transverse





- En tant que organe interministériel
- Clinique Algo / Data drinks / Infolettre ...



## Une fonction interne

- Accompagner des projets mais aussi
- Mener ses propres projets : Piaf

- Intégration dans toutes les fonctions transverses
- Participation à la vie de l'administration
- Equilibre travail/télé-travail flexible
- Beaucoup de libertés
- Merci !



-  N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *arXiv:1908.10084 [cs]*, Aug. 2019.  
**arXiv: 1908.10084.**
-  N. Reimers and I. Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation,” *arXiv:2004.09813 [cs]*, Apr. 2020.  
**arXiv: 2004.09813.**
-  J. Uszkoreit, *Transformer: A Novel Neural Network Architecture for Language Understanding*, August - 2017.
-  R. Keraron, G. Lancrenon, M. Bras, F. Allary, G. Moyse, T. Scialom, E.-P. Soriano-Morales, and J. Staiano, “Project PIAF: Building a Native French Question-Answering Dataset,” in *Proceedings of The 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 5481–5490, European Language Resources Association, May 2020.

-  L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, r. de la Clergerie, D. Seddah, and B. Sagot, “CamemBERT: a Tasty French Language Model,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 7203–7219, Association for Computational Linguistics, July 2020.
-  H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab, “FlauBERT: Unsupervised Language Model Pre-training for French,” *arXiv:1912.05372 [cs]*, Mar. 2020.  
**arXiv: 1912.05372.**