Atelier Pseudonymisation et Anonymisation des données

IRD

Kim Montalibet

Septembre 2021

etalab gouv.fr

Outline

Outline

Présentation d'Etatab

- Département de la DINUM, **Etalab** coordonne la conception et la mise en oeuvre de la stratégie de l'Etat dans le domaine de la donnée.
- Une action tout au long du cycle de vie de la donnée :
  - **Ouverture** des données (data.gouv.fr)
  - **Circulation et partage** des données (api.gouv.fr)
  - **Exploitation** des données et **algorithmes publics**

Le Lab IA, mission au sein d'Etalab

- Le **Lab IA**, créé au sein d'Etalab en 2019, a vocation à accompagner les **administrations dans la mise en oeuvre de solutions d'IA**
- Développe des **outils mutualisés et guides pour les administrations**, dont l'outil de pseudonymisation est un exemple (guide, repertoire de code et application dont les liens sont donnés en annexe)
- Travaux initiés avec le **Conseil d'Etat** pour la pseudonymisation des décisions de justice administratives

Contexte

- L'open data par défaut concerne à la fois les administrations et le monde de la recherche, comme le montre l'initiative de la **science ouverte**
- L'ouverture doit pouvoir concilier les obcjectifs de transparence et de protection des données personnelles
- La réglementation sur la protection des données personnelles (RGPD) implique souvent d'**anonymiser ou de pseudonymiser** des données avant diffusion, tâche qui peut s'avérer complexe, et ce particulièrement lorsque les **données sont non structurées** (texte, voix par exemple)
- Exemple dans l'administration : les décisions de justice

Définitions 1/2

- **Pseudonymisation** : est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans avoir recours à des informations supplémentaires. En pratique la pseudonymisation consiste à remplacer **les données directement identifiantes** (nom, prénom, etc.) par **des données indirectement identifiantes** (alias, n, etc.). Il est toutefois bien souvent possible de retrouver l'identité de ceux-ci grâce à des données tierces. C'est pourquoi des données pseudonymisées demeurent des données personnelles. **L'opération de pseudonymisation est réversible**, contrairement à l'anonymisation.

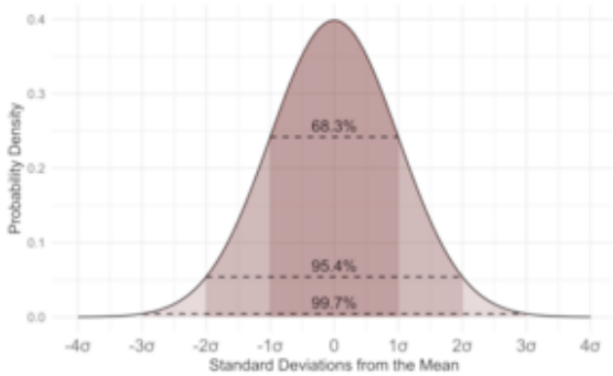
Définitions 2/2

- **Anonymisation** : Processus consistant à traiter des données à caractère personnel afin d'**empêcher totalement et de manière irréversible l'identification d'une personne physique**. L'anonymisation suppose donc qu'il n'y ait plus aucun lien possible entre l'information concernée et la personne à laquelle elle se rattache.
- **Différence entr les 2** : La différence entre anonymisation et pseudonymisation réside ainsi dans **le caractère réversible ou non** de la dissimulation des données à caractère personnel.

Quelles données personnelles retirer ?

- Cela dépend du contexte règlementaire
- Il y a en général un arbitrage entre **protection des données personnelles et complétude de l'information** contenue dans les données pseudonymisées
- En pratique, une complète anonymisation est difficile à atteindre et à évaluer et peut aboutir à une trop grande **perte d'informations**

Anonymat : quelle gradation ?



Données à caractère personnel

Nominative

Aucun effort pour identifier

Pseudonyme

Effort plus ou moins important nécessaire pour ré-identifier

Anonyme ?

Impossible de réidentifier ou de déduire de l'information sur un individu

La pseudonymisation en pratique

- Plusieurs méthodes sont possibles :
  - Annotation **manuelle**
  - Automatisation par **moteur de règles**
  - Automatisation par méthodes d'**apprentissage automatique** (traitement du langage naturel)

L'anonymisation en pratique

- Plusieurs méthodes sont possibles et doivent évoluer au fil du temps pour s'adapter à des nouvelles techniques de réidentification
  - **K-anonymisation** : publier des informations sur des groupes qui doivent contenir au moins K individus
  - **L-deversité** : Tout groupe de quasi-identifiants doit comprendre + de L valeurs sensibles distinctes
  - **Confidentialité différentielle** : données essentiellement identique en retirant n'importe quel individu de la source

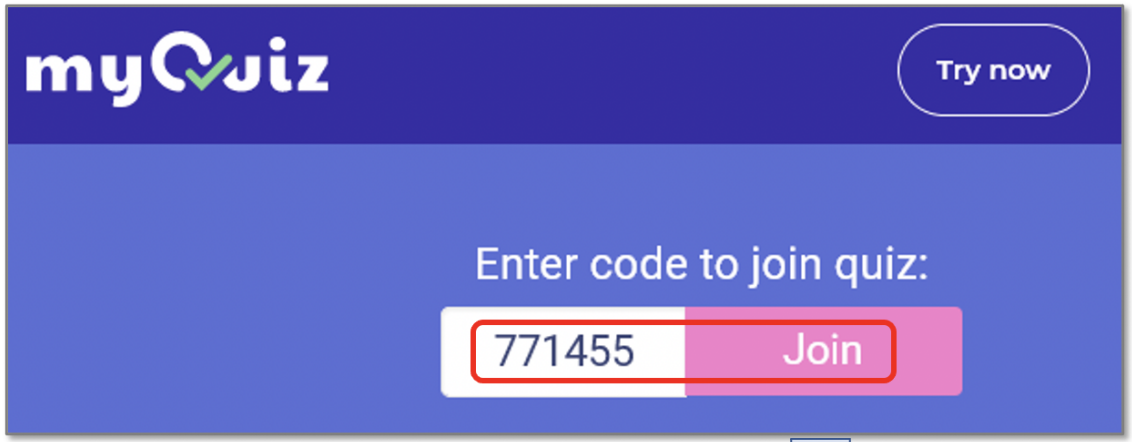
Evaluer les risques

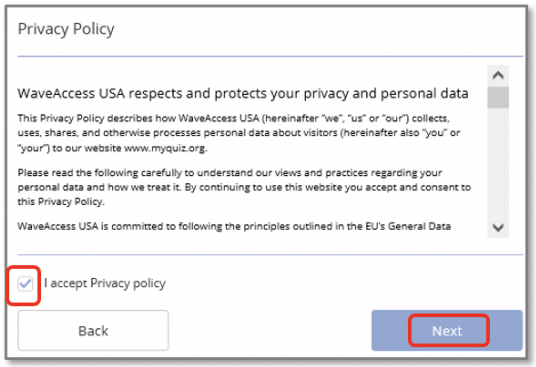
- Les propriétés des différentes méthodes sont connues, mais le choix de la méthode dépendra des cas d'usage et des contraintes associées (nécessité d'une précision élevée par exemple)
- Les responsables du traitement et les sous-traitants doivent prendre en compte la **finalité et le contexte global** afin de choisir la méthode la plus appropriée

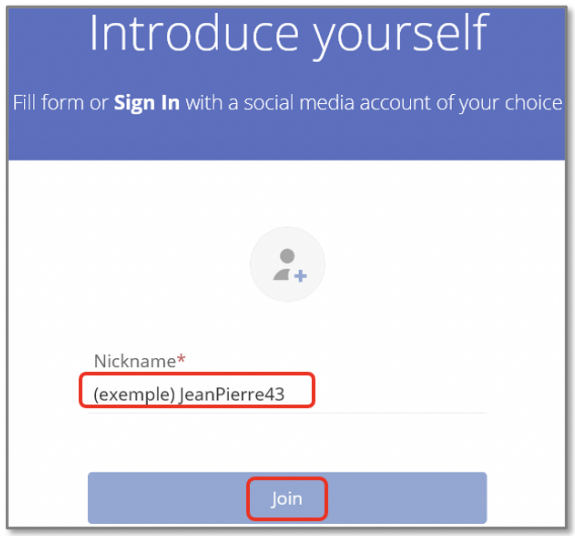
Outline

A vous de jouer !

https://myquiz.org code 771455







Outline

Pseudonymisation : la tâche d'apprentissage automatique

Document annoté | Document pseudonymisé

Lecture du mercredi 28 février 2018

REPUBLIQUE FRANCAISE

AU NOM DU PEUPLE FRANCAIS

Vu la procédure suivante:

Mme Aleron PRENOM Landry NOM , demeurant au 123 rue Fausse Ville-Fantastique-Sur-Saône ADRESSE , a demandé au juge des référés du tribunal administratif de Poitiers, sur le fondement de l'article L. 521-1 du code de justice administrative, de suspendre l'exécution de la décision du 21 juillet 2012 par laquelle le recteur de l'académie de Bordeaux a rejeté sa demande d'inscription en première année de licence de sciences et techniques des activités physiques et sportives (STAPS) à l'université de Bordeaux pour l'année 2002/2003 et d'enjoindre au recteur de l'inscrire temporairement au sein de cette formation dans un délai de quinze jours, sous une astreinte de 10 euros par jour de retard.

Par une ordonnance n° 1703763 du 21 mars 2011, le juge des référés du tribunal administratif a suspendu l'exécution de cette décision et a enjoint au recteur de l'académie de Bordeaux de procéder à l'inscription de Mme Landry NOM en première année de licence de STAPS dans l'attente qu'il soit statué au fond sur sa légalité.

Le résultat : texte pseudonymisé

Document annoté | Document pseudonymisé

Lecture du mercredi 28 février 2018

REPUBLIQUE FRANCAISE

AU NOM DU PEUPLE FRANCAIS

Vu la procédure suivante:

Mme A... PRENOM B... NOM , demeurant au ... ... ADRESSE , a demandé au juge des référés du tribunal administratif de Poitiers, sur le fondement de l'article L. 521-1 du code de justice administrative, de suspendre l'exécution de la décision du 21 juillet 2012 par laquelle le recteur de l'académie de Bordeaux a rejeté sa demande d'inscription en première année de licence de sciences et techniques des activités physiques et sportives (STAPS) à l'université de Bordeaux pour l'année 2002/2003 et d'enjoindre au recteur de l'inscrire temporairement au sein de cette formation dans un délai de quinze jours, sous une astreinte de 10 euros par jour de retard.

Par une ordonnance n° 1703763 du 21 mars 2011, le juge des référés du tribunal administratif a suspendu l'exécution de cette décision et a enjoint au recteur de l'académie de Bordeaux de procéder à l'inscription de Mme B... NOM en première année de licence de STAPS dans l'attente qu'il soit statué au fond sur sa légalité.

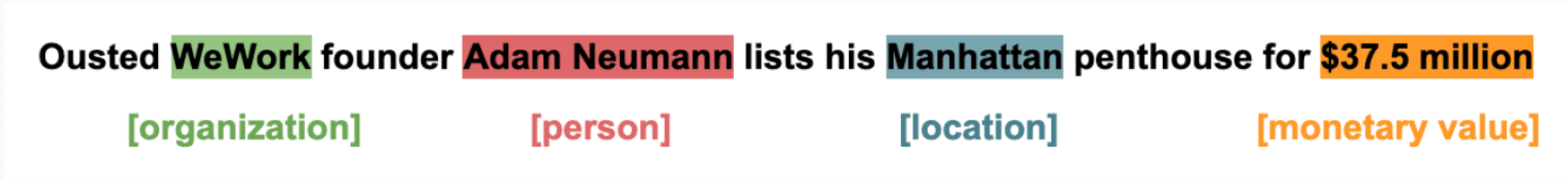
Par un pourvoi, enregistré le 3 janvier 2009 au secrétariat du contentieux du Conseil d'Etat, la ministre de l'enseignement supérieur, de la recherche et de l'innovation demande au Conseil d'Etat d'annuler cette ordonnance.

Démonstration

https://datascience.etalab.studio/pseudo/

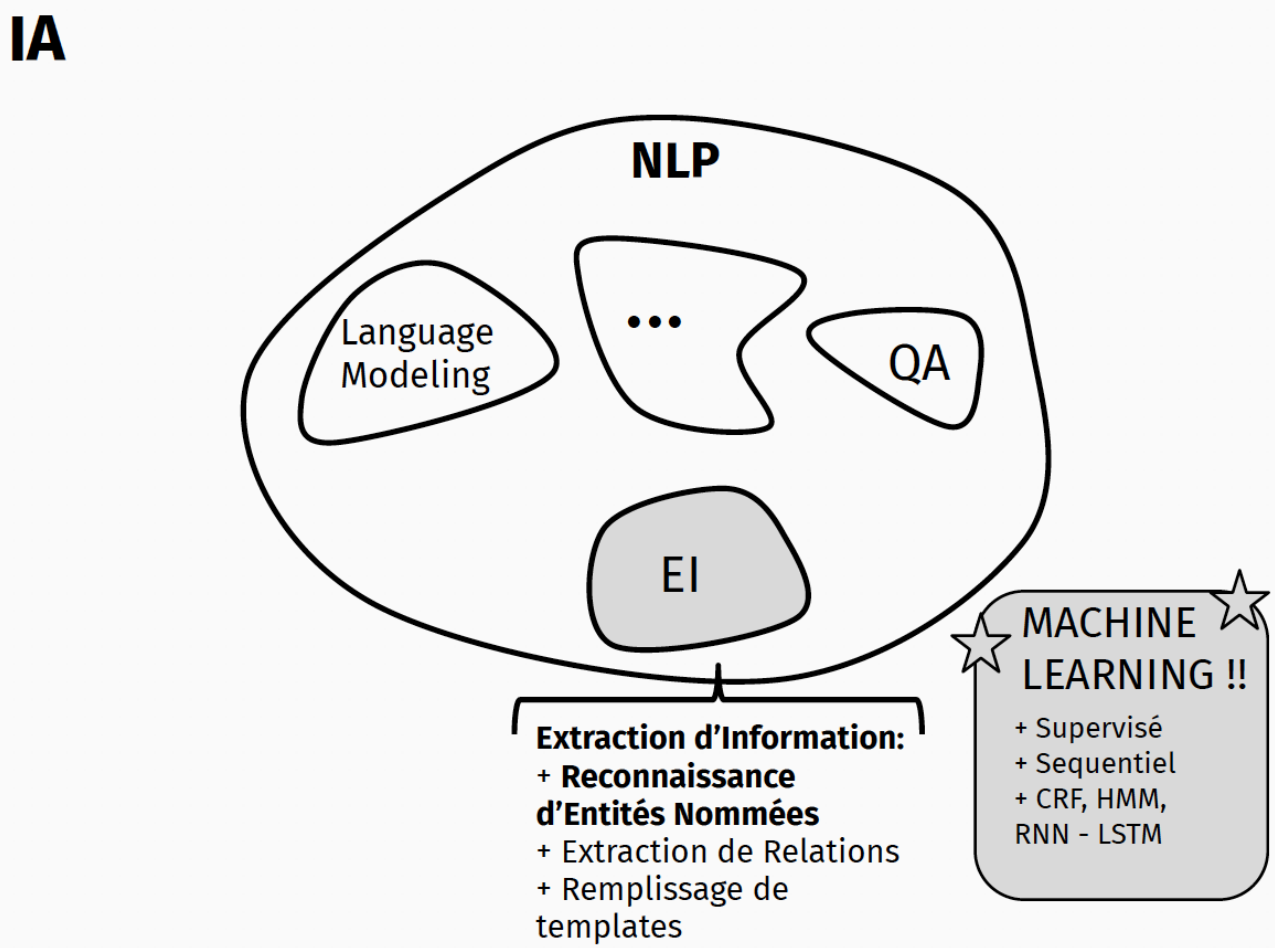
Comment ça marche ?

- La reconnaissance d'entités nommées

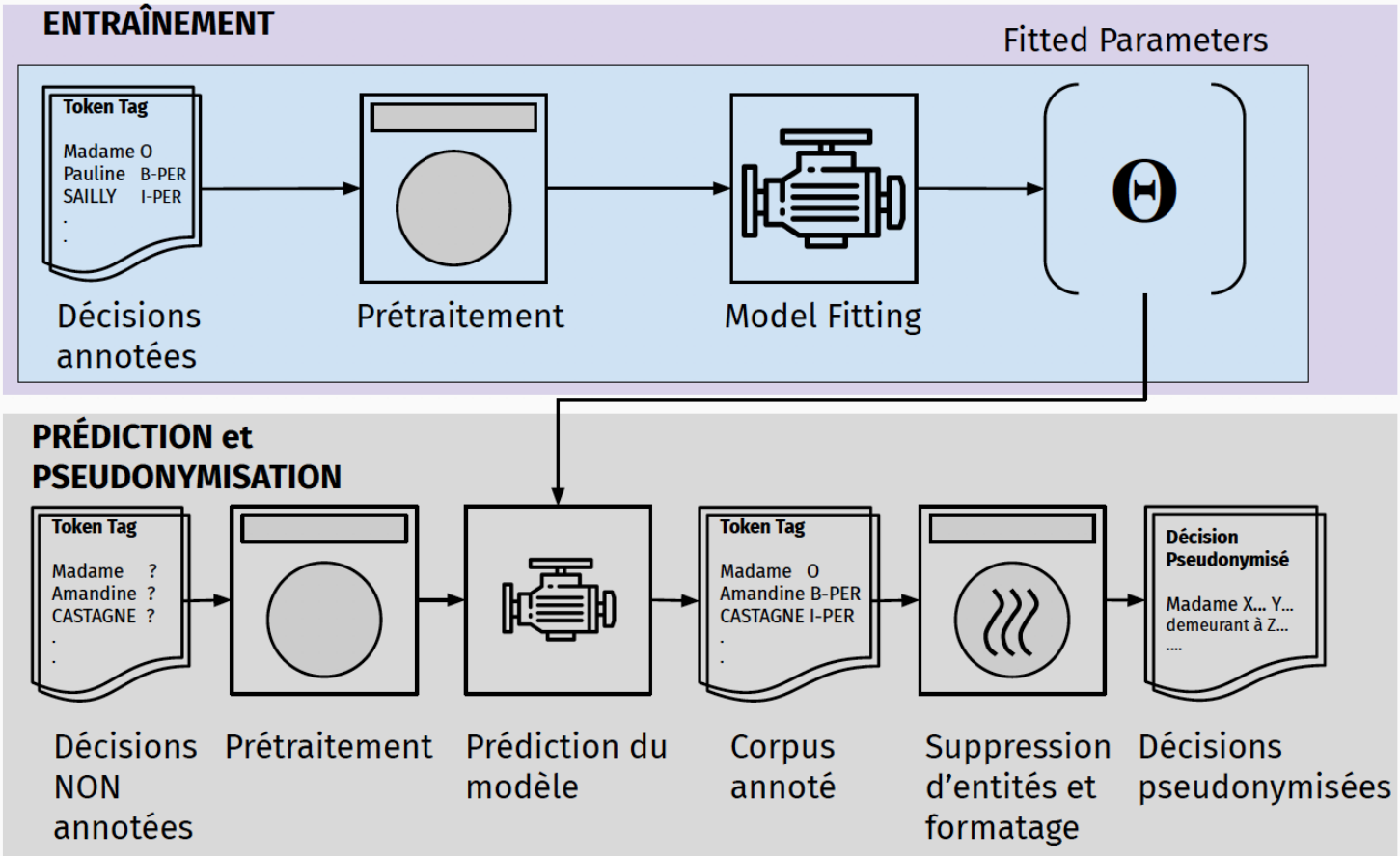


- C'est une tâche standard en **traitement du langage naturel** (NLP)

Le traitement du langage naturel (NLP)



Les différentes étapes de la pseudonymisation



Conclusion sur la pseudonymisation de textes par l'IA

- Les techniques d'apprentissage automatique ne sont pas de la magie -> **il y a des erreurs**
- Entraîner un modèle nécessite des ressources (temps humain pour annoter, temps pour développer et tester le modèle, ressources de calcul pour entraîner le modèle, etc..)
- Ces techniques pourront en général permettre un gain signaficatif lorsque le **volume de textes est élevé**

Outline

Echanges/témoignages

- Avez-vous déjà été confronté à des problématiques de traitement de données à caractère personnel ?
- Quelles solutions avez-vous mises en place ?

Annexes 1 - Ressources Lab IA d'Etalab

- Lien guide pseudonymisation : `https://guides.etalab.gouv.fr/pseudonymisation/#a-quoi-sert-ce-guide`
- Lien application web de pseudonymisatinon : `https://datascience.etalab.studio/pseudo/`
- Lien répertoire de code GitHub : `https://github.com/etalab-ia/pseudo_conseil_etat`

Annexes 2 - Autres ressources

- Guide de la CNIL au sujet de l'anonymisation pour publication en open data : `https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen`
- Guide INSH-CNRS sur les données personnelles et la science ouverte : `https://inshs.cnrs.fr/`