Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

PREMIER MINISTRE

SECRÉTARIAT GÉNÉRAL POUR LA MODERNISATION DE L'ACTION PUBLIQUE

DIRECTION INTERMINISTÉRIELLE DU NUMÉRIQUE ET DU SYSTÈME D'INFORMATION ET DE COMMUNICATION DE L'ÉTAT

Tour Mirabeau
39-43 Quai André Citroën
75015 Paris

Affaire suivie par : Alexis Eidelman
Téléphone : 0140156907
Mél. : alexis.eidelman@modernisation.gouv.fr

Réf. 2017 - Etalab - 28

Paris, le 27 février 2017

L'Administrateur général des données

A

Mme Carole CHAMPALAUNE, Directrice des affaires civiles et du Sceau,
Monsieur Jean-Paul JEAN, directeur du service de documentation, des études et du rapport, Cour de Cassation,
M. Bertrand MUNCH, Directeur de l'information légale et administrative

Objet* : *Avancées en matière de pseudonymisation des décisions de justice

Dans le cadre de la convention de partenariat et de recherche P-2016-004 signée par la DILA, la Cour de Cassation et l'Administrateur Général des Données (AGD), et dans le contexte plus global de la diffusion des décisions de justice en open data prévue par la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, l'AGD a réalisé des premiers travaux de **« pseudonymisation automatique »** : les premiers résultats obtenus sont prometteurs et méritent d'être partagés avec l'ensemble des parties prenantes (*voir l'annexe 1*).

A partir des données fournies par la Cour de cassation (extraction de jurinet 2015) et des données publiées par la DILA, il s'agit d'étudier la possibilité de pseudonymiser automatiquement les décisions présentes dans les bases de données jurisprudentielles de la Cour de cassation et des autres juridictions de l'ordre judiciaire en vue de leur diffusion.

En particulier, l'approche algorithmique peut réaliser la pseudonymisation et faciliter l'application rapide des articles 20 et 21 de la loi pour une République numérique, en déterminant un niveau de risque de ré-identification associé à la mise à disposition des décisions de justice (*voir annexe 2*).

L'Administrateur général des données propose de recourir à une prestation de datascience via l'accord-cadre établi en 2016.
Afin de décider des orientations du projet et d'évaluer les algorithmes qui seront produits, il est nécessaire de définir des critères d'acceptabilité d'une méthode automatique.

Plusieurs orientations peuvent être envisagées et la concertation semble le meilleur moyen de désigner l'option la plus favorable, afin que le travail soit le plus utile possible à la Cour de Cassation, à la DILA, tout comme aux services du ministère de la Justice.

Ainsi, nous souhaiterions que le ministère de la Justice, comme les autres parties prenantes de la convention, participe au cadrage de cette intervention de datascience, notamment pour définir le besoin et fixer les critères de réussite de ce projet, afin qu'il puisse être le plus proche des besoins opérationnels des services.

Nous sommes évidemment disponibles pour vous présenter les résultats détaillés de cette étude et à l'écoute de vos propositions et idées pour alimenter la seconde phase de ce travail, en lien avec la discussion autour du projet de décret d'application des articles « open data jurisprudence ».

Si vous en êtes d'accord, vous pourriez désigner un représentant de vos services pour participer à ces travaux.

Pour toute question relative à ce groupe de travail, les points contacts opérationnels sont les suivants:

- Alexis Eidelman : alexis.eidelman@modernisation.gouv.fr
- Thomas Menant : thomas.menant@modernisation.gouv.fr
- Mathilde Bras : mathilde.bras@modernisation.gouv.fr

Nous restons à votre disposition pour toute information complémentaire.

Je vous prie de croire, Madame, Monsieur, à l'assurance de ma considération très distinguée.

Laure LUCCHESI

Mission Etalab

Annexe 1 : étapes techniques du procédé de pseudonymisation automatique

On peut décomposer les étapes techniques en trois phases importantes :

1. **La première consiste à relier le texte de jurinet avec le texte publié par la Dila correspondant**. Sur les 15 000 textes, actuellement 10 000 sont ainsi retrouvés (il s'agit globalement de décision de la Cour de Cassation). Un peu d'aide permettra certainement de rapprocher les 5 000 autres textes (il s'agit de décisions des cours d'appel) mais 10 000 textes suffisent pour construire et évaluer un algorithme.

2. **L'étape suivante consiste à identifier les mots du texte jurinet qui ont été pseudonymisés dans le texte publié par la Dila**. Cette étape est cruciale car ce sont ces mots que l'on va ensuite essayer de qualifier. Cette étape est importante car c'est elle qui va constituer la base de connaissance des algorithmes d'apprentissages. Les erreurs générées à cette étape introduisent de la confusion entre les mots à pseudonymiser et limitent la capacité des algorithmes à identifier les caractéristiques propres aux mots pseudonymisés.

   L'opération est plus technique qu'il n'y parait. En effet, les différences entre les versions Dila et Jurinet ne sont pas uniquement les noms alors que c'est cette seule partie que l'on doit identifier. Actuellement, on estime que les « mots blancottés » de 7 500 textes ont été correctement associés aux mots correspondants dans jurinet.

3. **Ensuite, la dernière étape consiste à construire un algorithme (par des règles ou par apprentissage automatique) qui identifie les mots.**
   - L'approche par règles consiste à déterminer très directement les caractéristiques des noms à pseudonymiser (première lettre en majuscule par exemple). On peut toujours ajouter de nouvelles règles mais elle fonctionne déjà plutôt bien. Sur une base de test, elle identifie 7 700 noms propres, en laisse passer 12 et retire à tort 10 % des mots.
   - L'approche par *machine learning* (elle aussi en développement), évaluée sur une base de test de 75 000 mots donne contenant 700 noms propres, identifie 97% des mots (erreur de 3%, 20 mots) en pseudonymisant à tort 0,5 % des noms communs (390 mots).

Annexe 2 : Idées de critères d'évaluation d'une méthode de pseudonymisation

- Pourcentage de noms non-pseudonymisés
- Pourcentage de textes ayant au moins un nom non-pseudonymisés
- Pourcentage de mots pseudonymisés à tort
- Pourcentage de textes dont on voudrait une relecture humaine (complémentaire du nombre de textes dont on sait, sans se tromper que l'pseudonymisation est valide). Cela revient à considérer les degrés de certitude de l'algorithme.
- Temps de calcul par décision