
EXTRACTION AUTOMATIQUE DES QUANTITÉS DE STUPÉFIANTS, DANS LES RÉSUMÉS DE PROCÉDURE ENREGISTRÉS PAR LES FORCES DE SÉCURITÉ INTÉRIEURE, À L'AIDE DE MODÈLES LÉGERS

Mathias ROBERT (*)

(*) SSMSI, Bureau des études et statistiques sur la criminalité organisée

mathias.robert@interieur.gouv.fr

Mots-clés (6 maximum) : Analyse textuelle, machine learning, extraction, modèles de langage, entités nommées, stupéfiants

Domaines : Science des Données – Analyse du Langage Naturel.

Résumé

Au sein du service statistique ministériel de la sécurité intérieure (SSMSI), le bureau des études et statistiques sur la criminalité organisée (Besco) étudie les champs infractionnels de la criminalité organisée, dont les infractions à la législation sur les stupéfiants (ILS), principalement le trafic et l'usage de stupéfiants. Le SSMSI dispose de données administratives issues des logiciels de rédaction des procédures (LRP) de la police et de la gendarmerie nationales (LRPPN, LRPGN). Ces logiciels alimentent des puits de données qui fournissent diverses informations générales sur les mis en cause, les victimes et les circonstances des faits commis. À partir des données brutes issues des puits sont constituées des bases statistiques, facilement exploitables par les chargés d'études. Le présent travail a vocation à enrichir les données sur les ILS disponibles dans ces bases, notamment par l'ajout de variables indiquant le ou les types de stupéfiants en cause, ainsi que leur quantité associée, généralement indiquée sous la forme d'un poids en grammes.

Au sein des puits de données, plusieurs sources permettent d'identifier les stupéfiants : la base des procès-verbaux électroniques, la base OSIRIS gérée par l'Office anti-stupéfiants (OFAST) et la base des « objets » issue de LRPPN. Dans ces bases, quand elle est présente, l'information sur le type et la quantité est inscrite dans des champs spécifiques et est donc facile à extraire. Cependant, ces informations ne sont pas renseignées pour tous les mis en cause pour ILS ; quand le type de stupéfiant et la quantité associée sont inconnus, l'étude des résumés de procédure, petits textes rédigés par les policiers et gendarmes décrivant les manières d'opérer (« *manop* »), devient indispensable. Ces résumés de procédure sont de qualité inégale et ne sont pas toujours remplis.

Les *manop* issues de LRP concernant les mis en cause pour trafic ou usage de stupéfiants mentionnent dans 80 % des cas le(s) type(s) de stupéfiants concerné(s) ; moins de 5 % des *manop* sont vides. Il s'agit le plus souvent de résine de cannabis, d'herbe de cannabis, de cocaïne, d'héroïne

et d'ecstasy. À l'aide d'expressions régulières et de listes de mots-clés, il est aisé de détecter quels types de stupéfiants sont impliqués. En revanche, il est plus difficile de repérer la quantité associée. Il s'agit en effet d'identifier un nombre, une unité de mesure et le stupéfiant associé.

Une approche naïve peut consister en l'identification de *patterns* décrivant des poids (« 30 g », « 1,2 kilo », « 250grm », etc.) à l'aide d'expressions régulières. La proximité des différents poids avec les types de stupéfiants détectés dans la manière d'opérer permet de déterminer le stupéfiant le plus probable auquel ce poids appartient. Le cas où autant de stupéfiants que de poids sont détectés est le plus simple. En revanche, pour M types de stupéfiants et N poids, l'affectation est plus complexe à résoudre et s'apparente à un problème d'affectation linéaire, où le coût d'une affectation correspond à la distance dans le texte entre un stupéfiant et sa quantité associée supposée. Pour 10 stupéfiants et 9 quantités, il y a ainsi $\binom{10}{9} \times 9!$ (soit plus de 3 millions) affectations différentes possibles. L'algorithme hongrois permet notamment de résoudre ce problème en temps polynomial mais de nombreuses autres méthodes existent, dont celle de R. Jonker et T. Volgenant [1].

Puisqu'elle est basée sur des expressions régulières, cette méthode est peu flexible. Il y a de nombreux cas où la quantité de stupéfiant ne saurait être détectée par des expressions régulières (« trente grammes », « 120 de cannabis », « cocaïne : 3+5g », etc.), ou au contraire identifiée fallacieusement (« 30kg de légumes », « véhicule immatriculé ab 230 gr », « poids total de 850 g », etc.). Le présent travail présente ainsi différentes méthodes de traitement automatique du langage naturel, permettant l'extraction de quantités dans les textes et des substances auxquelles elles sont liées. Les méthodes présentées ici se veulent légères et adaptées à des capacités computationnelles limitées qui sont celles mises à disposition dans l'environnement professionnel du SSMSI.

Ce problème s'apparente notamment à une situation de reconnaissance d'entités nommées (*Named-Entity Recognition*, NER) [2]. Les modèles de NER permettent de reconnaître certaines entités prédéfinies comme des prénoms ou des lieux par exemple. Il est possible d'entraîner un modèle NER sur la détection de l'entité « quantité de résine de cannabis » ou d'une autre substance que l'on sait présente dans une *manop*. L'entraînement du modèle, nécessitant par ailleurs une étape de labellisation manuelle, est une tâche relativement lourde en ressources, c'est pourquoi il est possible de se tourner vers un modèle tel que GLiNER (*Generalist and Lightweight model for Named-Entity Recognition*) [3]. Ce dernier permet en « *zero-shot* », soit sans entraînement, de détecter directement l'entité choisie – néanmoins, entraîner un tel modèle reste avantageux pour ses performances.

Enfin ce problème peut également être résolu en mettant à profit les performances des modèles de langage. Si les grands modèles de langage (LLM) sont davantage performants, l'utilisation de plus petits modèles de langage (SLM) de 100 millions à 1 milliard de paramètres est souvent suffisante pour des tâches simples comme l'extraction de quantités, que ce soit en *zero-shot* ou bien après un *fine-tuning*, plus coûteux mais qui donne généralement de meilleurs résultats [4].

Bibliographie

[1] Jonker, R., & Volgenant, T. (1986). Improving the Hungarian assignment algorithm. *Operations research letters*, 5(4), 171-175.

[2] Mikheev, A., Moens, M., & Grover, C. (1999). Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

[3] Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2023). Gliner: Generalist model for named entity recognition using bidirectional transformer. *Association for Computational Linguistics*.

[4] Schick, T., & Schütze, H. (2021). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *Association for Computational Linguistics*.