datascience.net

Insee

Amélioration de la qualité de saisie en ligne des libellés de profession

Description du challenge

L'Insee développe le recours à la collecte par internet comme un des modes de collecte de ses enquêtes, sans exclusivité. La réponse au recensement de la population par internet est ainsi proposée dans certaines communes et devrait bientôt l'être dans toutes les communes.

Dans les enquêtes auprès des personnes, on recueille des informations sur leur profession sous forme d'un texte court à remplir par la personne interrogée. Grâce à un logiciel de reconnaissance de libellé, l'Insee transforme ensuite ces informations en des codes directement utilisables pour réaliser des analyses.

Une partie des textes (libellés) saisis par les répondants ne sont pas directement exploitables pour être traités par le logiciel de reconnaissance de libellés : ils sont trop vagues, mal orthographiés, correspondent à une double profession, intègrent des professions qui ne sont pas encore ou plus enregistrées dans le référentiel parce qu'il s'agit d'une nouvelle terminologie qui vient d'apparaître ou d'une ancienne qui a disparu. Ces informations sont alors traitées par des agents de l'Insee (codeurs) pour être ensuite réintégrées dans la suite des traitements avec les autres données.

Dans un souci d'innovation pour améliorer l'efficacité des traitements automatiques, la Direction de la méthodologie et de la coordination statistique et internationale de l'Insee souhaite proposer en temps réel des suggestions à l'internaute lorsqu'il saisit des libellés de profession. Il s'agit de proposer trois libellés qui figurent dans la base de référence au plus, que le répondant pourrait choisir si l'un d'eux correspond effectivement à sa profession et qui permettent, s'ils sont choisis, d'obtenir une codification automatiquement. Il s'agira donc de proposer à l'internaute, au moment de la saisie (à l'image de ce qui est proposé couramment sur internet par les moteurs de recherche), des libellés figurant dans une liste de référence fournie dans le challenge.

Quelques précisions sur le traitement des textes saisis

Pour simplifier le problème posé, les traitements automatiques réalisés par l'Insee sur les libellés sont volontairement exposés de façon partielle. Le candidat curieux pourra se reporter à la documentation du logiciel Sicore développé par l'Insee. Ceci n'apportera cependant pas d'éclairage supplémentaire dans le cadre de la résolution de la question posée.

Il est important de noter que le programme de reconnaissance des libellés utilisé par l'Insee procède à une normalisation des libellés qui explique la forme que prennent les libellés de référence. Par exemple, le libellé doit être écrit en lettres capitales et ne doit pas comporter de caractères accentués ou « exotiques », les articles ne sont pas traités, certains caractères sont ignorés, etc. Par ailleurs, pour l'attribution d'un code, le programme n'utilise pas la totalité des caractères du texte saisi pour la profession, et utilise des caractères jokers ($ et *).

Description des données

Le candidat dispose des fichiers suivants :

- referentiel.csv
- professions_non_traitees.csv
- synonymes.csv
- exemples.csv

referentiel.csv

Le fichier « referentiel.csv » liste les 27 190 libellés de référence reconnus lors des traitements et les codes qui y sont associés.

PRECODE;	**Profession;**
T-W001;	CONSTRUCTEUR BATIMENT
T-W001;	CONSTRUCTEUR BATIMENT $3
T-W001;	CONSTRUCTEUR MAISON
T-W001;	CONSTRUCTEUR MAISON $3
T-W002;	CONSTRUCTEUR BETON ARMURIER $
...	...

Les caractères jokers $ et * figurant dans les libellés de profession du fichier référentiel.

Caractère $
Lorsque présence d'un $ dans le libellé, on peut remplacer le $ par n'importe quel mot, le résultat du codage n'est pas affecté. Par exemple DETECTEUR MATELOT MARINE $ est équivalent à

DETECTEUR MATELOT MARINE NATIONALE ou encore DETECTEUR MATELOT MARINE MARCHANDE
Lorsque présence de $3 dans le libellé, on peut remplacer le $3 par 3 mots, aucune incidence sur le résultat du codage.

*Caractère **
Lorsque des * apparaissent dans le libellé, cela indique que l'on peut remplacer celles-ci par n'importe quel caractère. Le choix des caractères n'a aucune incidence sur le résultat du codage. Cette possibilité permet d'ignorer les éventuelles fautes d'orthographe. Par exemple ABAT**** est équivalent à ABATOIR ou encore ABATTOIRE, ou encore ABATOIRE, etc. Autre exemple : CABL**** TELEPHONE est équivalent à CABLEUR TELEPHONE ou bien CABLAGE TELEPHONE.

Règles pour lire les libellés du référentiel

- Le libellé est écrit sur au plus 6 mots (le blanc est le séparateur, une fois supprimé les mots non traités - article, synonyme à blanc -, et les synonymes affectés) et de au plus 20 caractères par mot.
- Les articles ne sont pas traités ; par exemple AGENT DE L'ETAT se transforme en AGENT ETAT. La liste des articles non pris en compte est incluse dans la liste des synonymes.
- Les caractères ignorés, caractères "blancs" sont: ( ) / \ ' - _ + , :
- Un libellé suivi de $ indique qu'il peut être suivi par un autre mot, le résultat sera identique
- Un libellé suivi de $3 indique que celui-ci peut être suivi par 3 autres mots, le résultat sera identique
- Pas de double profession, il est nécessaire d'isoler au préalable les professions. Seule la première profession sera à retenir
- Un mot composant le libellé commençant par des lettres et suivi de ***** indique que seules les lettres importantes sont les lettres renseignées.
- Le libellé doit être écrit en lettres capitales et ne doit pas comporter de caractères accentués ou exotiques (*,/,&,.....). Par exemple, le libellé maçon doit s'écrire MACON

Il peut y avoir plusieurs libellés pour un même précode, ils correspondent à la même profession.

professions_non_traitees.csv

Le fichier « **professions_non_traitees.csv** » contient les 22 024 libellés qui ne peuvent être traités automatiquement, et pour lesquels le candidat devra fournir trois propositions de libellés alternatifs au plus, sous forme de libellés, dont on est sûr qu'ils correspondent à un précode.

synonymes.csv

Le fichier « **synonymes.csv** » est un dictionnaire de synonymes, qui peut être utilisé par le candidat.

exemples.csv

C'est un fichier d'exemples de texte saisis pour la profession, et une correction correcte attribuée par un codeur de l'Insee.

Critère d'évaluation

Format du fichier de résultat

Le candidat doit fournir pour chaque libellé de profession non traité automatiquement, trois propositions de libellés alternatifs dont un au moins donne le bon code. Le format de ce fichier de résultat est :

Identifiant du libellé à améliorer;	Libellé 1;	Libellé 2;	Libellé 3
1;	MEDECIN CHEF;	MEDECIN NUTRITIONIST;	MEDECIN CHIROPRACTEUR
2;	...	...	...

A l'issue du challenge, le candidat devra transmettre une documentation décrivant l'algorithme et le programme associé qui pourra qui devra être réalisé en SAS, Java, ou R. Le logiciel SAS est toutefois fortement préféré.

Critère de performance

Les résultats seront évalués selon le critère suivant : **Taux de codage correct des 22 024 libellés proposés**. Un codage est correct si l'un au moins des trois libellés est correct et donne le bon code.