

Fonctionnement des identificateurs intelligents

Contenu

[Introduction](#)

[Numéros de carte de crédit](#)

[Numéros de sécurité sociale aux États-Unis](#)

[Numéros CUSIP](#)

[Numéros de routage ABA](#)

Introduction

Ce document décrit les identificateurs intelligents, qui sont des modèles d'analyse de contenu intégrés qui détectent certains types de données. Pour cette version, le système implémentera des identificateurs intelligents pour les numéros de carte de crédit, les numéros de sécurité sociale américains, les numéros CUSIP et les numéros de routage ABA.

En interne, un identificateur intelligent consiste en une expression régulière qui correspond aux chaînes de candidats, ainsi qu'en une fonction de validation qui vérifie d'une manière ou d'une autre la correspondance de candidats. Par exemple, la fonction de validation d'un numéro de carte de crédit garantit que le chiffre de contrôle est correct.

Les expressions régulières pour chaque identificateur intelligent incluent des ancres de limite de mot ('b') aux deux extrémités. (Cela empêche le système de correspondre à un numéro de sécurité sociale américain, par exemple, au milieu d'une chaîne de chiffres plus longue.) Pour des raisons de simplicité, elles sont omises dans les descriptions ci-dessous.

L'implémentation des identificateurs intelligents doit être prudente quant au chevauchement des correspondances, car une sous-chaîne trouvée par l'expression régulière peut ne pas être validée. Par exemple, un filtre recherche les numéros de carte de crédit par rapport à la chaîne 9999 4321 9999 999 9995 1234 5678 9000 doit trouver le numéro de carte de crédit valide 4321 999 999 995, même si une simple analyse d'expression régulière pour des nombres possibles trouverait 9999 4321 999 9999 et 9995 1234 5678 9000.

Numéros de carte de crédit

Un numéro de carte de crédit commence par un type de carte de longueur variable, qui indique si le numéro est un VISA, MasterCard, AMEX, etc., et se termine par un chiffre de chèque. Différents types de cartes utilisent des nombres différents de chiffres dans l'ensemble du nombre, mais le calcul du chiffre de contrôle est le même dans chaque cas.

Notez que les cartes enRoute ou JCB ne correspondent pas. De plus, il n'existe pas de numéro de VISA à 13 chiffres et il n'y aura pas de correspondance dans notre mise en oeuvre.

Les numéros de carte de crédit à 16 chiffres correspondent à l'une des expressions régulières suivantes :

$[0-9]\{4\}-[0-9]\{4\}-[0-9]\{4\}-[0-9]\{4\}-[0-9]\{4\}$
 $[0-9]\{4\}\.[0-9]\{4\}\.[0-9]\{4\}\.[0-9]\{4\}$
 $[0-9]\{4\} [0-9]\{4\} [0-9]\{4\} [0-9]\{4\} [0-9]\{4\}$
 $[0-9]\{16\}$

Avec le préfixe « 4 », « 51 »-« 55 », ou « 6011 ».

Les numéros AMEX à 15 chiffres correspondent à l'une des expressions régulières suivantes :

$[0-9]\{4\}-[0-9]\{6\}-[0-9]\{5\}$
 $[0-9]\{4\}\.[0-9]\{6\}\.[0-9]\{5\}$
 $[0-9]\{4\} [0-9]\{6\} [0-9]\{5\}$
 $[0-9]\{15\}$

Avec les préfixes autorisés « 34 » ou « 37 ».

Les numéros du Diners Club à 14 chiffres correspondent à l'une des expressions régulières suivantes :

$[0-9]\{4\}-[0-9]\{6\}-[0-9]\{4\}$
 $[0-9]\{4\}\.[0-9]\{6\}\.[0-9]\{4\}$
 $[0-9]\{4\} [0-9]\{6\} [0-9]\{4\}$
 $[0-9]\{14\}$

Les préfixes autorisés sont « 300 »-« 305 », « 36 » ou « 38 ».

Notez que les expressions régulières définissent un regroupement spécifique de chiffres pour une longueur de carte de crédit donnée et que si des signes de ponctuation se produisent entre les chiffres, ils doivent être identiques dans l'ensemble.

Le dernier chiffre d'un numéro de carte de crédit est un chiffre de contrôle créé à l'aide de l'algorithme Luhn. À partir de l'extrémité droite du nombre, double chaque deuxième chiffre. Ensuite, additionnez les chiffres individuels des nombres résultants (à la fois ceux qui ont été doublés et ceux qui ne l'ont pas été). Si le résultat est un multiple de 10, le nombre est valide.

Par exemple, avec le numéro 1234 5678 9012 3456 :

1 2 3 4 5 7 8 1 2 3 5 6

Double : 2 2 6 4 10 6 14 8 18 0 2 2 6 4 10 6

Ajout de 2 + 2 + 6 + 3 + 1 + 0 ... + 1 + 0 + 6 donne 64, ce qui n'est pas un multiple de 10, donc le numéro n'est pas valide.

Compte tenu du numéro 1234 5678 9876 3333 :

1 2 3 4 5 7 8 7 3 3 3

Double : 2 2 6 4 10 6 14 8 18 8 14 6 6 3 6 3

Ajout de $2 + 2 + 6 + 3 + 1 + 0 \dots + 6 + 3$ donne 80, soit un multiple de 10, donc le numéro est valide.

Numéros de sécurité sociale aux États-Unis

Les numéros de sécurité sociale sont divisés en un numéro de zone à trois chiffres, qui est attribué géographiquement, un numéro de groupe à deux chiffres attribué dans un ordre particulier d'une zone, et un numéro de série à quatre chiffres attribué séquentiellement.

Notre mise en oeuvre utilisera les expressions régulières suivantes :

```
[0-9]{3}-[0-9]{2}-[0-9]{4}
[0-9]{3}\.[0-9]{2}\.[0-9]{4}
[0-9]{3} [0-9]{2} [0-9]{4}
```

Voici quelques exemples des expressions ci-dessus :

```
555-55-5555
555.55.5555
555 55 5555
```

L'administration de la sécurité sociale tient à jour une liste des numéros de zone/groupe qui ont été attribués : SSN émis [3]. Mais comme ce document change périodiquement, nous ne pouvons pas nous en fier pour sa validation. La fonction de validation vérifie qu'aucun des 3 champs ne contient uniquement des zéros et que les 3 premiers chiffres sont inférieurs à 800. (La référence précédente utilise 771 comme limite, mais le SSA a déjà attribué des numéros avec les 3 premiers chiffres 771 et 772.)

(Les numéros commençant par 666 ne sont pas attribués et les numéros compris entre 987-65-4320 et 987-65-4329 sont réservés à la publicité. De plus, 078-05-1120 est le SSN le plus utilisé ; c'était le numéro de série d'un secrétaire d'une société de portefeuille, qui a utilisé le numéro comme exemple.)

Numéros CUSIP

Les numéros du CUSIP (Committee on Uniform Security Identification Procedure) sont 9 identificateurs alphanumériques qui identifient les titres nord-américains de divers types. Le numéro est divisé en un numéro d'émetteur de 6 caractères, qui identifie de manière unique l'émetteur (par exemple, une société), un suffixe de 2 caractères qui identifie le titre en question ; p. ex., actions ordinaires, par rapport aux actions privilégiées par rapport aux options par rapport aux instruments à revenu fixe.

Le code d'identificateur intelligent CUSIP utilise les expressions régulières suivantes :

```
[0-9]{3}[0-9a-zA-Z]{3} [0-9a-zA-Z]{2} [0-9]
[0-9]{3}[0-9a-zA-Z]{3}-[0-9a-zA-Z]{2}-[0-9]
[0-9]{3}[0-9a-zA-Z]{3}[0-9a-zA-Z]{2}[0-9]
```

La fonction de validation est similaire à celle utilisée pour les numéros de carte de crédit. La seule

différence est que les lettres du numéro CUSIP sont converties en valeur numérique en attribuant A=10, B=11, ..., Z=35.

Un exemple du site cusip.com utilise le numéro CUSIP 392690 QT 3 :

3 9 2 6 9 0 T 3

Convertir les lettres : 3 9 2 6 9 0 26 29 3

Double : 3 18 2 12 9 0 26 58 3

Ajout de $3 + 1 + 8 + 2 + 1 + 2 + \dots + 5 + 8 + 3$ donne 50, c'est-à-dire un multiple de 10, donc le numéro d'origine était valide.

Numéros de routage ABA

Un numéro de routage ABA (American Banking Association) est une valeur à 9 chiffres. Les quatre premiers chiffres sont le symbole de routage de la Réserve fédérale, les quatre suivants l'identifiant de l'institution et le dernier un chiffre de chèque.

Le code d'identificateur intelligent du numéro de routage ABA utilise les expressions régulières suivantes :

$[0-9]\{4\} [0-9]\{4\} [0-9]$
 $[0-9]\{4\}-[0-9]\{4\}-[0-9]$
 $[0-9]\{9\}$

La fonction de validation impliquait de multiplier chaque chiffre par 3, 7, 1, ...; si la somme des produits est un multiple de 10, le nombre est valide.

Prenons par exemple le numéro 123 456 789 :

1 2 3 4 6 7 9

Multiplier par : 3 7 1 3 7 1 3 7 1 1 1 1 1 1

Produit : 3 14 3 12 35 6 21 56 9

Ajouter $3 + 13 + 3 + 12 + 35 + 6 + 21 + 56 + 9$ donne 159, ce qui n'est pas un multiple de 10, donc le numéro d'origine n'était pas valide.

Compte tenu du nombre 322 271 627 :

3 2 2 7 1 6 2 7

Multiplier par : 3 7 1 3 7 1 3 7 1 1 1 1 1 1

Produit : 9 14 2 6 49 1 18 14 7

Ajouter $9 + 13 + 2 + 6 + 39 + 1 + 18 + 13 + 7$ donne 120, qui est un multiple de 10, donc le numéro d'origine était valide.

(Bien que certaines plages de symboles de routage de la Réserve fédérale soient réservées, et donc non attribuées, l'algorithme de validation ne vérifiera pas les numéros réservés, pour éviter

d'avoir à les réviser si l'ABA modifie sa politique.)