Quelques limites à l'algorithme implémenté dans Sicore

Théo Leroy

JMS - Mercredi 30 mars 2022

Sicore

Un outil générique de codification automatique

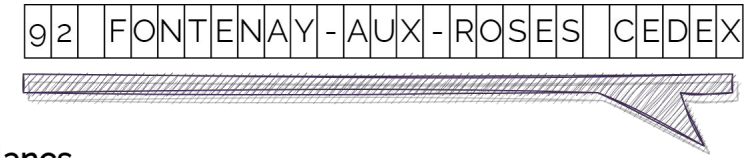
Input (libellé à coder) ⟶ *Output (code)*

08 REMAUCOURT	08356
08 REMILLY AILLICOURT	08357
08 REMILLY POTHEES	08358

- Né en 1993 grâce aux travaux de Pascal Rivière
- Utilisé pour de nombreux cas d'usage : professions, activités d'entreprise, communes, pays, produits, activités quotidiennes…
- Système expert : construit autour de 3 moteurs de règles

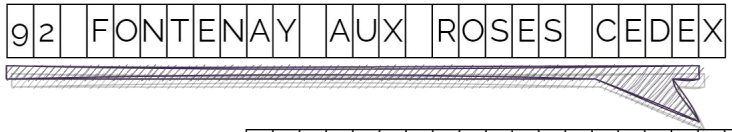
Règles de normalisation

Chaîne en entrée



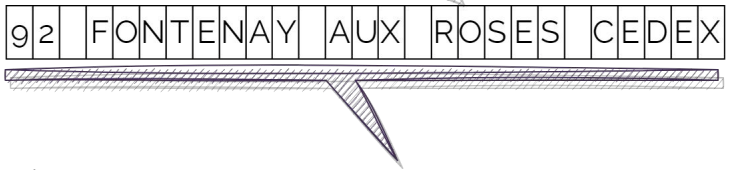
Traitement des caractères blancs

Dans cet environnement, les caractères ci-après sont remplacés par des espaces : ()-_',/"



Traitement des caractères vides

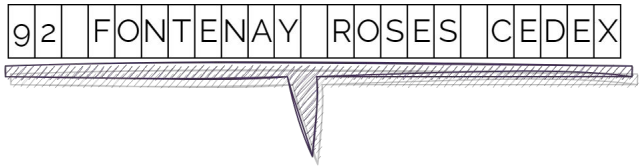
Le point est supprimé. Comme ce caractère est absent, cette chaîne à coder reste donc inchangée par rapport à l'étape précédente..



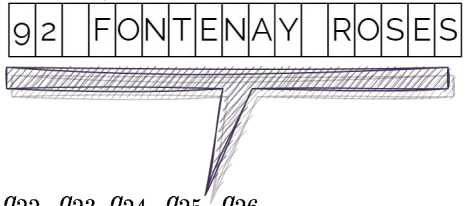
Application d'expressions synonymes

Des expressions de synonymes sont appliquées successivement. Plus de deux cents sont définies pour cet environnement Sicore. Seules deux transforment la chaîne de caractère pour ce cas.

Expression 24/201
remplacement du mot "AUX" par le mot vide

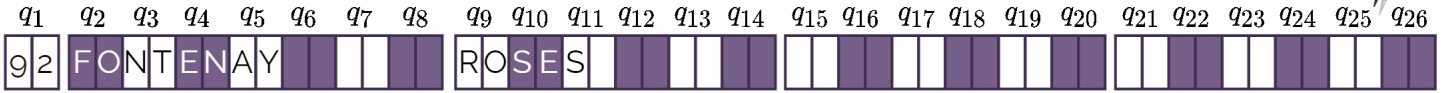


Expression 95/201
remplacement du mot "CEDEX" par le mot vide

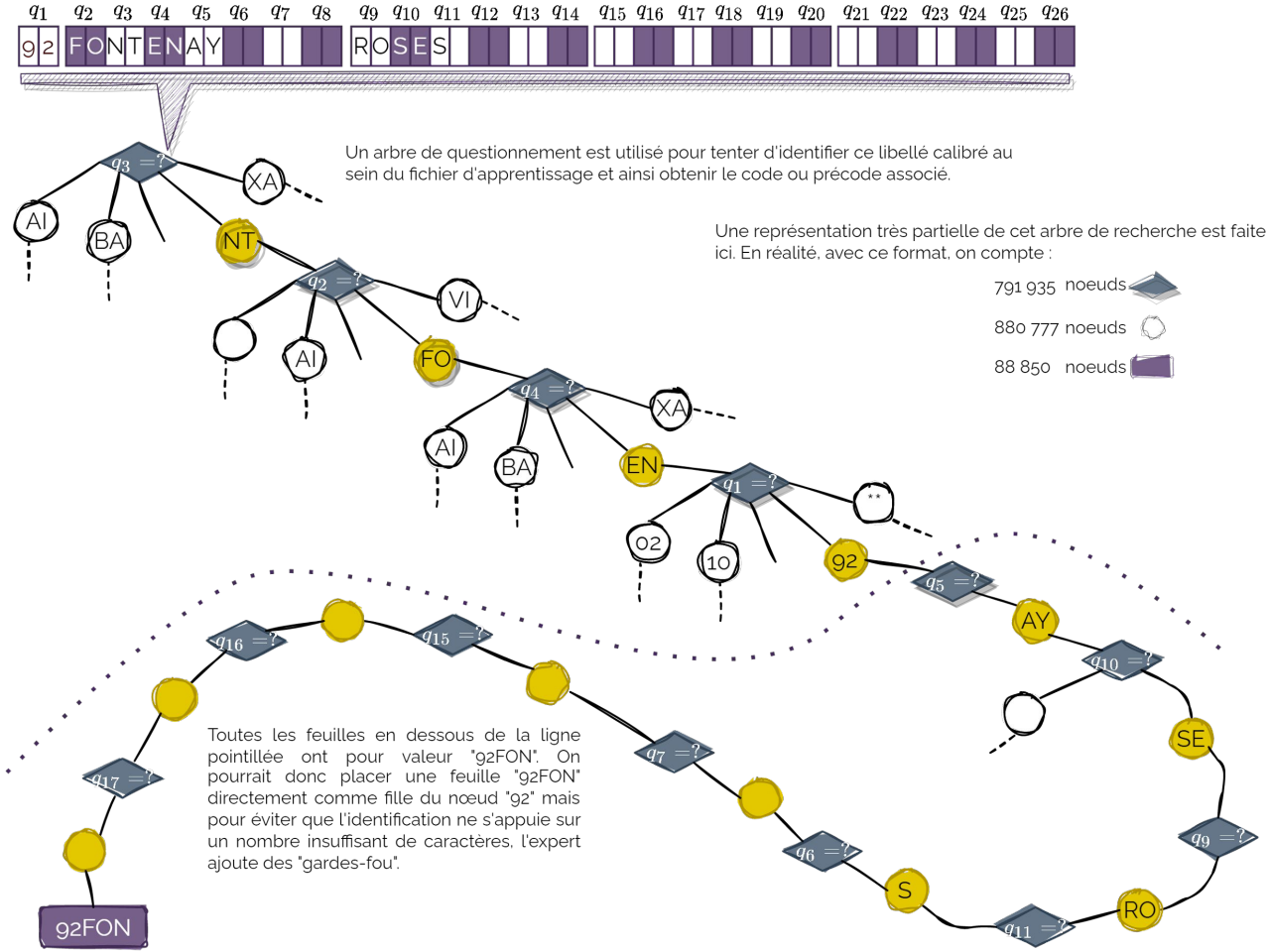


Calibrage

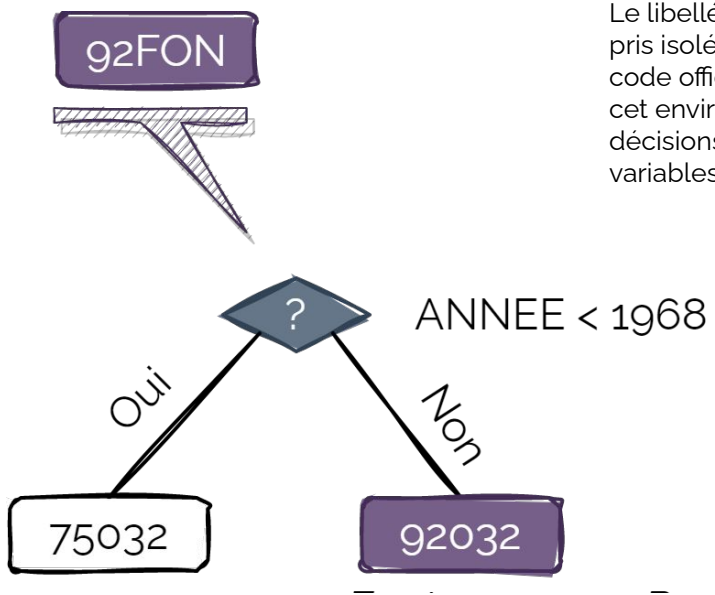
Le libellé est ensuite modifié pour occuper une taille fixe. Ici le libellé sera formé de 5 mots, le premier de 2 caractères, le deuxième de 14 et les 3 derniers mots de 12 caractères. Ce formatage induit souvent de tronquer des mots ou au contraire d'ajouter des espaces.



Règles d'identification du libellé dans un index



Règles logiques faisant appel à des variables annexes



Le libellé "92 FONTENAY AUX ROSES CEDEX", pris isolément ne permet pas de coder dans le code officiel géographique. L'e concepteur de cet environnement, définit des règles de décisions pour coder en utilisant d'autres variables : ici le millésime du COG

Pertinence de la reconnaissance du libellé sous forme d'arbre de questionnement

D'un index à un arbre (peu profonds)

Trouver la meilleure position de bigramme localement sur un critère d'entropie

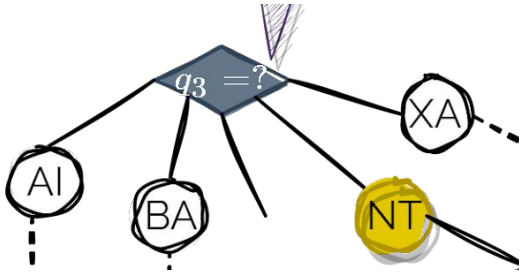
Code	Bigramme							
(T)	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
01001	01	AB	ER	GE	ME	CL	EM	EN
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
08355	08	RE	GN	IO	WE	__	__	__
08356	08	RE	MA	UC	OU	__	__	__
08357	08	RE	MI	LL	Y_	__	__	__
08358	08	RE	MI	LL	Y_	PO	TH	EE
08360	08	RE	NN	EV	IL	__	__	__
08361	08	RE	NW	EZ	__	__	__	__
08362	08	RE	TH	EL	__	__	__	__
08363	08	BO	IS	__	__	BR	YA	S_
08363	08	LA	__	__	__	BO	UV	ER
08363	08	LE	S_	__	__	BR	OU	TA
08363	08	OR	ZY	__	__	__	__	__
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98CLI	98	IL	E_	__	__	CL	IP	PE

Proportion des libellés de l'index appartenant au noeud y conditionnellement à x

$$\underset{q_j}{\operatorname{argmin}} \sum_{y \in \Gamma_{q_j}(x)} P(y|x) H(T|y)$$

Entropie de la variable T conditionnellement au noeud y

Ensemble des noeuds fils de x en choisissant le jème bigramme

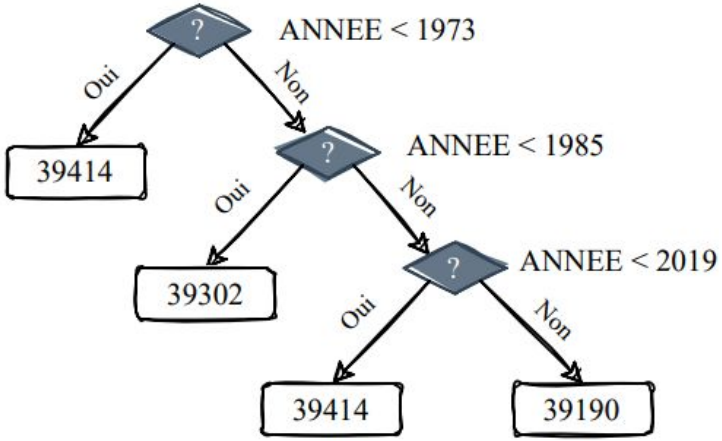


Une méthode d'identification finalement peu efficace

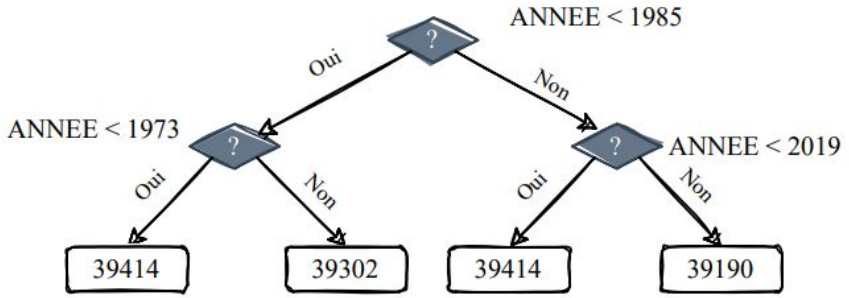
	Nombre d'observations	Base de connaissances Sicore	Pourcentage d'identification (efficacité de la méthode)		
			Appariement strict sur le libellé non normalisé	Appariement strict sur le libellé normalisé	Arbre de reconnaissance Sicore
Commune de résidence (Base tous salariés 2017)	27 000 000	Commune version 2021	69,5	97,4	98,2
Pays de naissance (EAR 2020)	900 000	PAYS version 2020	94,1	98,4	99,2
Profession actuelle des salariés (EAR 2020)	1 500 000	PCS2003R et REFNC version 2020	33,0	76,3	87,5
Activité économique pour les liasses provenant des chambres de métiers et de l'artisanat (Sirene 2020)	200 000	NR21C version 2021	35,3	49,0	63,3
Activité économique pour les liasses provenant des chambres des URSSAF (Sirene 2020)	500 000	NR21U version 2021	6,3	81,1	81,1

Un enrichissement difficile du moteur de règles

Le nombre de noeuds accessibles comme critère de complexité



Représentation sous forme de graphe orientée de la règle logique pour le libellé Petit-Mercey dans l'environnement Sicore Commune

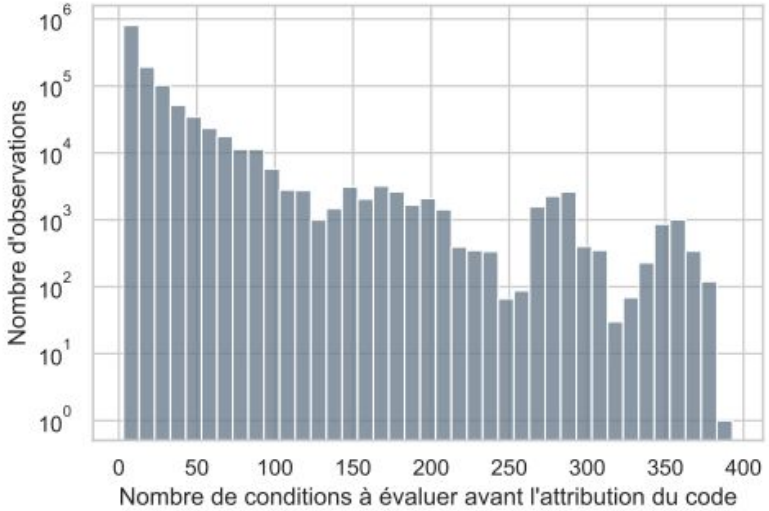


Règle de décision alternative

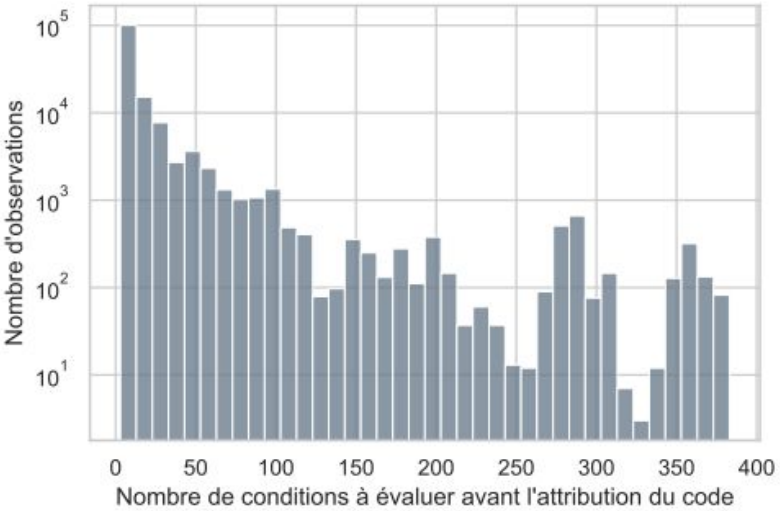
Profondeur maximale : 3 vs 2
Nombre de noeuds de questionnement : 3 dans les deux cas

Des règles de décision parfois très complexes

Environnement Sicore	Nombre de noeuds de conditions	Nombre de noeuds entrants	Statistiques de la distribution du nombre de noeuds de conditions accessibles depuis chaque noeud entrant			
			Minimum	Médiane	Moyenne	Maximum
PCS2003R codification de la PCS 2003 dans les enquêtes ménages et les EAR	14 775	2 446	4	14	194	2 332
REFNC codification en PCS 2003 en complément de PCS2003R pour les EAR uniquement	14 775	1 402	4	16	214	2 328
PCSDSN codification en PCS-ESE lors du traitement des DSN	17 346	934	0	11	278	1 152
ISCO codification en ISCO-08 à partir de la PCS 2003	263 689	2 442	0	29	108	1 159
ISCO2020 codification en ISCO-08 à partir de la liste des professions	69 432	5 786	12	12	12	12
PCS2020R codification en PCS 2020 des professions dans la liste des professions	98 362	5 786	17	17	17	17
COMMUNE codification dans le COG	4 420	4 124	1	1	1	3



(a) EAR 2020 - Profession actuelle des salariés



(b) EAR 2020 - Profession actuelle des non-salariés

Histogrammes du nombre de conditions testées dans les règles avant d'aboutir à un code (échelle logarithmique pour les fréquences)

Conclusion

- Un outil rapide et précieux
- Remplaçable en cas de modification des chaînes de production par des méthodes plus simples ou plus avancées
- Générique mais n’est pas toujours le plus adapté dans toutes les situations