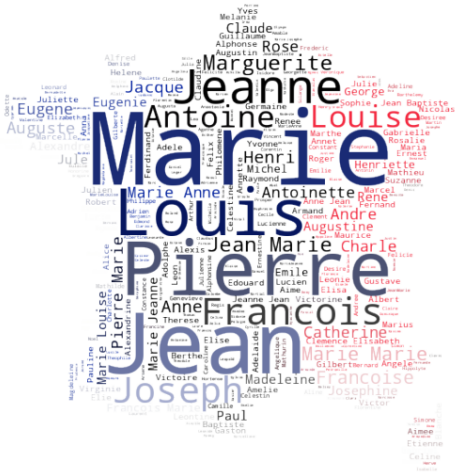
Claire Dechaux

ENSAE Paris : 3ème année
Année scolaire : 2023 - 2024

Projet de Machine Learning for Natural Language Processing

Sujet 3 : Typage des informations



Professeur : Christopher KERMORVANT
Chargée de TP : Yasmine HOURI

Table des matières

1 Présentation et description des données **1**
1.1 Statistiques descriptives des données disponibles 1
1.2 Estimation des données cibles 2
1.3 Identification autres données externes utilisables 2

2 Analyse des modèles **3**
2.1 Identification du type de tâches à réaliser 3
2.2 Analyse des forces et faiblesses des modèles utilisables 3
2.3 Choix d'un modèle 4

3 Expérimentation **4**
3.1 Description du protocol expérimental 4
3.2 Analyse des résultats obtenus 4

4 Conclusion **5**

Références **6**

Code **6**

Annexes **7**

1 Présentation et description des données

1.1 Statistiques descriptives des données disponibles

Nous travaillons sur une base de données issue du projet Socface qui vise à analyser les documents de recensements français et en extraire des informations à très grande échelle. Il s'agit ici de données générées par reconnaissance automatique de l'écriture (OCR) à partir des listes nominatives manuscrites des recensements de 1836 à 1936. Plus précisément, nous possédons dans notre base de données initiale 1 218 documents, soit 25 448 individus recensés. Pour chaque individu, plusieurs informations sont censées être complétées, ces informations sont catégorisées par 14 labels : age (*age*), année de naissance (*birth_date*), statut civil (*civil_status*), employeur (*employer*), prénom (*firstname*), nom de famille (*surname*), nom de jeune fille (*maiden_name*), nom de famille du chef de ménage (*surname_household*), lien de parenté avec le chef de ménage (*link*), lieu de résidence (*lob*), nationalité (*nationality*), niveau d'éducation (*education_level*), occupation (*occupation*) et observation (*observation*). Nous traitons les données de manière à passer de 14 labels à 11, en supprimant le niveau d'éducation et le nom de jeune fille puisque ces catégories ne sont pas une seule fois renseignées. De plus, nous regroupons les informations de *surname* et *surname_household* en une seule catégorie *surname*, puisqu'aucune de ces deux données n'est complétée en même temps, il semble au final s'agir de la même information.

Table 1 – Nombre d'occurences, d'uniques et de valeurs manquantes par label

	age	birth_date	civil_status	employer	firstname	link	lob	nationality	observation	occupation	surname
Nb d'occurences	16436	7344	10705	2910	24931	20736	9232	13314	596	16178	24798
unique	248	158	6	1074	2373	888	2889	70	307	1959	9464
NA (%)	35.5	71.2	57.9	88.6	1.7	18.5	63.7	47.6	97.6	36.4	3.3

Comme nous pouvons le voir Table 1, plusieurs catégories sont assez mal renseignées. On observe notamment qu'il y a 97.6 % de valeurs manquantes pour le label *observation*. Nous pourrions avoir l'idée ultérieure d'augmenter les données pour sur-représenter ces occurences (en dupliquant les lignes considérées ou en générant de nouvelles lignes en mélangant les informations avec d'autres lignes). Cependant, la catégorie *observation* semble être en grande partie des informations censées être contenues dans d'autres catégories (*cf. annexe 2* (4)), et donc provenants d'erreurs de recensements ou d'erreurs de l'OCR. Les informations de cette catégorie pourraient donc être recatégorisées par notre modèle de labelisation.

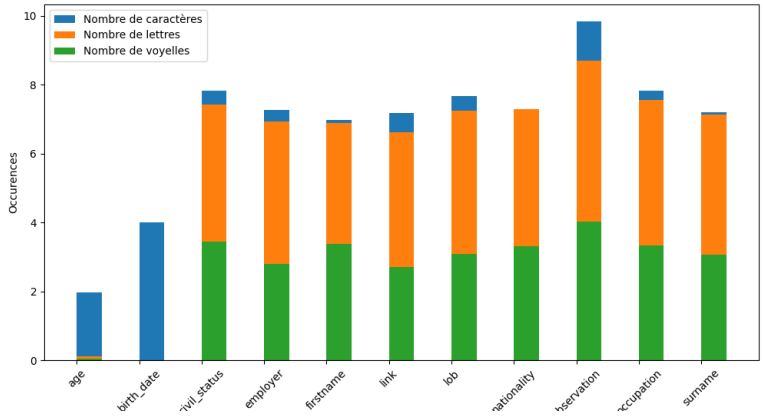
Table 2 – Tableau des **moyennes** des statistiques descriptives par label des chaînes de caractères de la base de données

	Longueur de chaine	Nombre de mots	Nb de carac. speciaux	Nombre de lettres	Nombre de voyelles	Nombre de consonnes
age	1.962947	1.030604	1.820273	0.112071	0.051046	0.061025
birth_date	3.997549	1.000408	3.994145	0.002996	0.001362	0.001634
civil_status	7.818216	1.397384	0.000000	7.420831	3.437366	3.983466
employer	7.255670	1.316151	0.014777	6.923368	2.792440	4.130928
firstname	6.985520	1.092295	0.000000	6.893185	3.376359	3.516826
link	7.184414	1.568142	0.003954	6.612220	2.699942	3.912278
lob	7.675802	1.387890	0.004008	7.253033	3.087088	4.165945
nationality	7.285564	1.001728	0.000075	7.283761	3.300811	3.982950
observation	9.837248	1.864094	0.255034	8.696309	4.026846	4.669463
occupation	7.823526	1.258746	0.000494	7.562925	3.321671	4.241254
surname	7.207759	1.083595	0.000000	7.124083	3.070852	4.053230

Nombre de mots : au sens, chaînes de caractères séparées par un espace.

La Table 2 présente les moyennes de certaines variables représentant les caractéristiques des chaînes de caractères renseignées, par catégorie (*cf. annexe 1* (4)) pour un tableau plus précis des statistiques). Nous pouvons déjà observer de grandes différences entre certaines catégories, notamment par rapport à la longueur de la chaîne de caractères, le nombre de caractères spéciaux, et le nombre de lettres. La Figure 1 représente ces différences de composition des chaînes de caractères entre les labels. Le nuage de points annexe 3 (*cf. annexe 3* (4)) représente aussi la répartition des catégories entre nombre de lettres et nombre de caractères.

FIGURE 1 – Barplot



1.2 Estimation des données cibles

Dans le contexte du projet Socface, une des tâches principales consiste en la reconnaissance et la typage des informations extraites des recensements français. Ici nous nous concentrons sur l'identification et l'extraction des entités, puis à l'attribution des étiquettes ou des tags à chaque information extraite. Ces étiquettes sont diverses, il peut s'agir de noms de famille, de prénoms, d'âges, d'années de naissance, de professions, de nationalités, etc. Il s'agit donc de labeliser les données extraites à partir de la reconnaissance optique de caractères (OCR), ligne par ligne. Cette phase de typage des informations revêt une importance particulière pour garantir la qualité et la précision des données extraites, ce qui facilitera les analyses statistiques et historiques ultérieures.

1.3 Identification autres données externes utilisables

Outre les transcriptions manuscrites des recensements et les définitions de tags associées, l'enrichissement du corpus peut être accompli en exploitant des données externes complémentaires. Parmi celles-ci, nous pourrions notamment penser à des données historiques, géographiques ou démographiques.

Les données géographiques offrent par exemple un cadre spatial crucial pour contextualiser les informations extraites des recensements. En effet, les données que nous utilisons ici datent de 1836 à 1936, et les noms de communes ont pu évoluer depuis, mais également au sein même de cette période. D'après DELATTRE Eric[1], il y aurait eu plusieurs décrets durant la première moitié du XXe siècle qui ont officialisé un nombre important de changements, le nombre de changements serait néanmoins moins important au cours du XIXe siècle. Cependant, Le Code Officiel Géographique (COG)[2] tenu par l'INSEE a permis de comptabiliser 1863 changements de nom de communes françaises métropolitaines entre 1943 et 2006, soit environ 5 % des communes. Ainsi, prendre en compte ces changments pourrait être utile dans la tâche d'identification des communes de résidence déclarées par les personnes recensées.

Par ailleurs, des données démographiques pourraient également aider à l'identification de la nationalité des recensés, en nous permettant d'avoir une idée plus exacte des proportions de chaque nationalité vivant en France durant cette période.

D'un autre côté, les données généalogiques constituent une source précieuse d'informations sur les individus répertoriés dans les recensements, en particulier pour valider ou compléter les données extraites. Les bases de données généalogiques et les arbres généalogiques en ligne regorgent de détails sur les relations

familiales, les événements de la vie (naissance, mariage, décès) et d'autres aspects de la vie des individus. En croisant ces données avec les transcriptions des recensements, il devient possible d'enrichir les profils individuels et de mieux comprendre leur contexte familial et social. Pour cela nous pourrions scrapper des données de sifférents sites de généalogie en France [3] [4], ou le portail généalogie du Ministère de la Culture [5].

En intégrant ces types de données à notre corpus, nous pourrions enrichir notre système de typage et fournir une vision plus holistique et contextuelle des individus répertoriés dans les recensements.

Par ailleurs, nous observons qu'aucun des labels suivants n'est identifié dans nos données actuelles : le niveau d'éducation (*education_level*) et le nom de jeune fille (*maiden_name*). Par conséquent, lors de l'inférence sur d'autres ensembles de données de recensement contenant ces labels, nos modèles risquent de ne pas être en mesure de les catégoriser correctement. Il pourrait donc être nécessaire d'enrichir nos données avec des informations telles que des noms de jeunes filles et des niveaux d'éducation correspondant à cette période.

Nous pourrions également augmenter nos données en créant artificiellemnt de nouvelles lignes à partir de celles déjà disponibles.

2 Analyse des modèles

2.1 Identification du type de tâches à réaliser

Dans le cadre de notre projet, la tâche que nous entreprenons est de la classification de tokens, et plus particulièrement de la reconnaissance d'entités nommées (NER, *Named Entity Recognition*)[6]. La NER consiste à identifier et à classifier des entités spécifiques dans un texte, telles que des noms de personnes, d'organisations, de lieux, de dates, de quantités, et d'autres informations importantes. Par exemple, dans un document sur l'histoire familiale, la NER peut être utilisée pour repérer et étiqueter automatiquement les noms des membres de la famille, les dates de naissance, les lieux de résidence, etc. Cette tâche permet de structurer efficacement les données textuelles, facilitant ainsi leur analyse, leur extraction d'informations et leur utilisation ultérieure dans divers contextes. En exploitant la NER, notre objectif est d'enrichir les transcriptions des documents de recensement avec des informations typées, telles que les noms, prénoms, âges, professions, etc., afin de faciliter une analyse statistique précise et approfondie des données historiques.

2.2 Analyse des forces et faiblesses des modèles utilisables

Différentes approches ont été développées pour détecter les "entités nommées", allant des règles manuelles aux méthodes supervisées. Celles-ci comprennent des algorithmes d'apprentissage supervisé avec un feature engineering textuel minutieux et des approches basées sur l'apprentissage profond, qui génèrent automatiquement leurs propres features pour classifier les tokens dans des catégories d'"entités nommées". Ces dernières années, les méthodes basées sur l'apprentissage profond ont connu une croissance rapide, offrant des performances de pointe [7]. Cette étude conclut que l'adaptation fine des modèles de langue contextuel généraux avec des données spécifiques au domaine est susceptible de fournir de bons résultats pour les cas d'utilisation avec des textes spécifiques au domaine et peu de données d'entraînement. Cependant, cette approche a été utilisée pour extraire des entités nommées dans des textes historiques OCRisés en allemand, français et anglais [8], et il a été montré que la performance de la détection des entités nommées diminue significativement avec la baisse de la qualité de la reconnaissance optique de caractères.

Par ailleurs, l'article *"A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories"*[9] se rapproche grandement de notre cadre d'étude, puisqu'il s'agit de détecter des "entités nommées" à partir d'OCR de documents français datants du 19ème siècle. Les auteurs ont notamment montré que la reconnaissance d'entités nommées basée sur BERT peut bénéficier d'un pré-entraînement et d'un affinage sur un corpus produit avec le même processus que les textes à annoter, ce qui est intéressant dans notre cas. Ils concluent également qu'avec un score F1 de 92 % avec seulement 49 exemples d'entraînement, le modèle pré-entraîné CamemBERT est un très bon choix de modèle pour cette tâche.

2.3 Choix d'un modèle

Nous avons décidé d'utiliser le modèle pré-entraîné CamemBERT [10] pour notre tâche, comme recommandé dans la littérature pour la reconnaissance d'entités nommées en français. CamemBERT est un modèle de langage basé sur l'architecture BERT, spécialement adapté pour la langue française. Il a été pré-entraîné sur un grand corpus de texte en français et a démontré d'excellentes performances dans différentes tâches de traitement automatique du langage naturel, y compris la NER. En particulier, CamemBERT a été fine-tuné sur des données spécifiques à la langue française, ce qui lui permet de mieux capturer les caractéristiques linguistiques de ce domaine. Cette capacité lui permet d'obtenir des résultats précis même avec des ensembles de données d'entraînement relativement petits.

3 Expérimentation

3.1 Description du protocol expérimental

Notre protocole expérimental repose sur un fine-tuning du modèle Camembert d'HuggingFace [11] afin d'adapter ce dernier à nos propres données en vue d'étiqueter de nouvelles données de recensement. Dans un premier temps, les données sont prétraitées : le texte brut est séparé des étiquettes correspondantes, puis ces dernières sont retirées du texte. Ensuite, une méthode B, I est appliquée pour dupliquer les tags en utilisant les tags B (début) et I (intérieur). Les données sont ensuite divisées en ensembles d'entraînement et de test dans un rapport de 70/30, suivies de la création de jeux de données distincts pour chaque ensemble, incluant les tokens et leurs étiquettes respectives. Les mots sont ensuite tokenisés en sous-mots à l'aide de l'autotokeniseur pré-entraîné associé à CamemBERT. Compte tenu de cette tokenisation, il est nécessaire de réaligner les tokens et les étiquettes, car un mot unique correspondant à une seule étiquette peut maintenant être divisé en deux sous-mots en raison de l'introduction des tags B et I. Les données d'entraînement sont alors formatées à l'aide du *DataCollator* avant d'être transmises au modèle pendant l'entraînement. Enfin, les métriques, telles que la précision, le rappel, le score F1 et l'accuracy, sont définies à l'aide du framework *Seqeval* pour évaluer les performances du modèle sur les données d'entraînement et de test. Les paramètres d'entraînement comprennent un taux d'apprentissage (learning rate) de 2exp −5, une taille de lot d'entraînement et d'évaluation de 10, un nombre d'époques d'entraînement de 10, et une pénalité (weight decay) de 0.05 pour prévenir le surajustement (overfitting)et réduire la magnitude des poids.

En utilisant ce protocole, nous visons à exploiter pleinement le potentiel du modèle CamemBERT pour la labelisation des entités nommées dans les données de recensement, conformément aux recommandations de la littérature spécialisée dans ce domaine.

3.2 Analyse des résultats obtenus

Table 3 – Métriques par catégories

	precision	recall	f1
age	0.997766	0.997969	0.997868
birth_date	0.997725	0.997725	0.997725
civil_status	0.999062	1.000000	0.999531
employer	0.953488	0.969595	0.961474
firstname	0.993170	0.993037	0.993103
link	0.970939	0.985652	0.978240
lob	0.969042	0.973599	0.971315
nationality	0.982806	0.977205	0.979998
observation	0.830986	0.617801	0.708709
occupation	0.964050	0.960280	0.962161
surname	0.994754	0.996094	0.995423

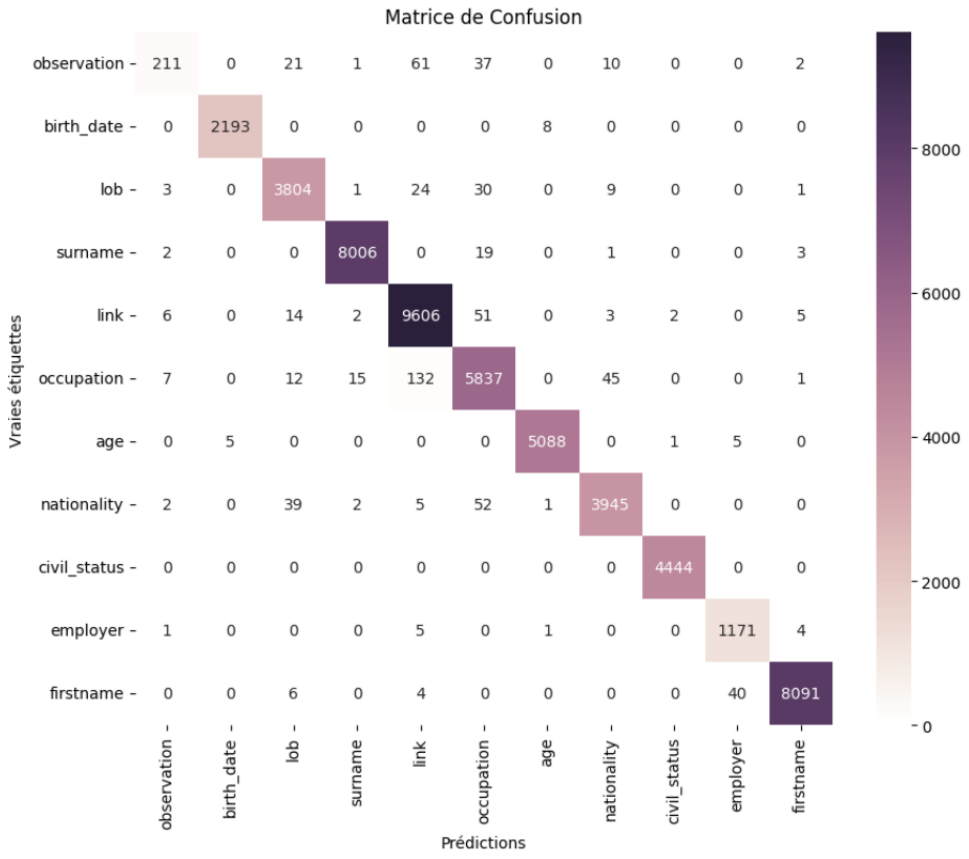
Table 4 – Métriques d'ensemble

precision	0.984455
recall	0.985436
F1	0.984946
accuracy	0.985461

Nous obtenons une accuracy de 0.985 % et un score F1 également de 0.985 %, ce qui est encore mieux que le score F1 de 92 % atteint par Abadie dans l'article *"A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories"* [9].

Nous remarquons cependant que tous les scores F1 sont supérieurs à 0.96 %, excepté pour le label *observation* qui a obtenu un score F1 de 0.708 %. De plus, elle apparaît comme la catégorie la moins bien classifiée également dans la matrice de confusion 2 (*cf. annexe 4* [4] pour la matrice de confusion des labels B-I) . En effet, le modèle classifie mal ce label et le confond souvent avec le lien familial, l'occupation et le lieu de résidence. Nous pouvions nous y attendre étant donné sa très faible proportion dans les données et, comme spécifié dans la partie 1.1 et en annexe 2 [4], il semble que la grande majorité des informations renseignées dans cette catégorie soient mal répertoriées et correspondent en fait à un autre label (mauvais recensement ou erreur d'OCR). Ainsi, le model pourrait aider à la bonne reclassification de ces informations, sous supervision.

FIGURE 2



4 Conclusion

En conclusion, notre étude démontre que le modèle CamemBERT, finement ajusté à nos données de recensement, offre de très bonnes performances dans la reconnaissance et le typage des entités nommées dans le contexte de recensements. Avec une précision et un score F1 de 0,985 %, notre modèle permet de labeliser nos données de test de manière efficace. Cependant, nous notons que la catégorie "observation" présente des défis de classification, suggérant la nécessité d'une réévaluation de ces données, et notamment de l'étude de l'importance de la qualité de l'OCR utilisé, enjeu majeur évoqué par Abadie en 2O22 [9].

Enfin, nos résultats ouvrent la voie à des possibilités de recherche future, notamment l'exploration de techniques d'augmentation de données et des l'ajout de sources externes complémentaires, telles que des données géographiques, démographiques et généalogiques, pour améliorer la qualité et la précision des résultats obtenus.

Références

[1] DELATTRE Eric. Le changement de nom des communes françaises aspects économiques, marketing et stratégiques. *Revue d'Économie Régionale Urbaine*, pages p. 269–291, 2007.

[2] INSEE. Code officiel géographique (cog). *https ://www.data.gouv.fr/fr/datasets/code-officiel-geographique-cog/.*

[3] Filae. *https ://www.filae.com/.*

[4] Geneanet. *https ://www.geneanet.org/.*

[5] Ministère de la Culture. Portail généalogie. *https ://www.culture.gouv.fr/Espace-documentation/Bases-de-donnees/Fiches-bases-de-donnees/Portail-genealogie2.*

[6] R. Sharnagat. Named entity recognition : A literature survey. *Center For Indian Language Technology*, 2014.

[7] Jianglei Han Jing Li, Aixin Sun and Chenliang Li. A survey on deep learning for named entity recognition. 2020.

[8] Neudecker C. Labusch, K. Named entity disambiguation and linking historic newspaper ocr with bert. 2020.

[9] Carlinet E. Chazalon J. Duménieu B. Abadie, N. A benchmark of named entity recognition approaches in historical documents application to 19 century french directories. *Lecture Notes in Computer Science.*, 2022.

[10] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[11] HuggingFace. Modèle camenbert. *https ://huggingface.co/docs/transformers/en/model$_d$oc/camembert.*

Code

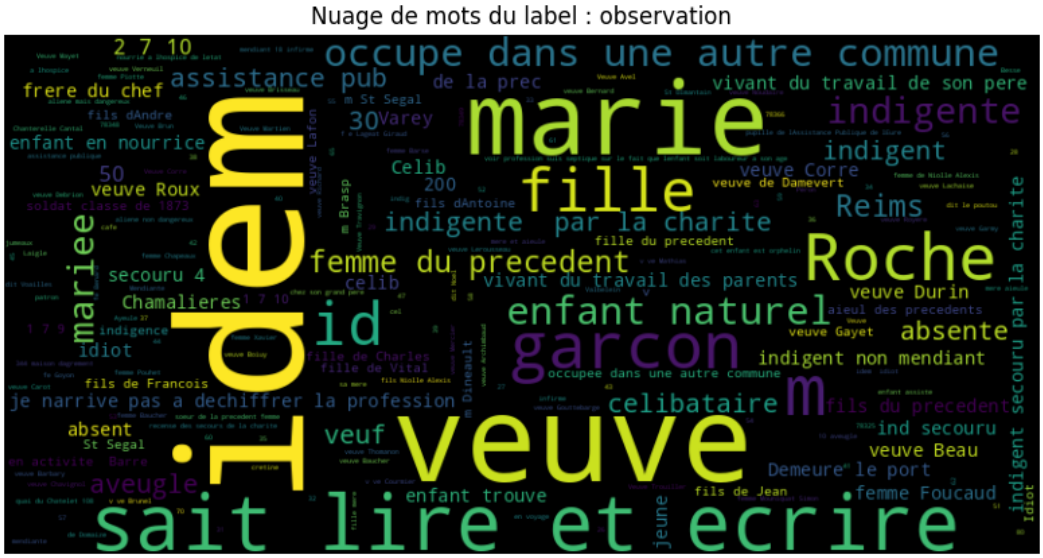
Publicly accessible : https ://github.com/cdechaux/Projet$_N LP_3 A$

Annexes

Annexe 1 - Tableau de statistiques descriptives des chaînes de caractères de la base de données

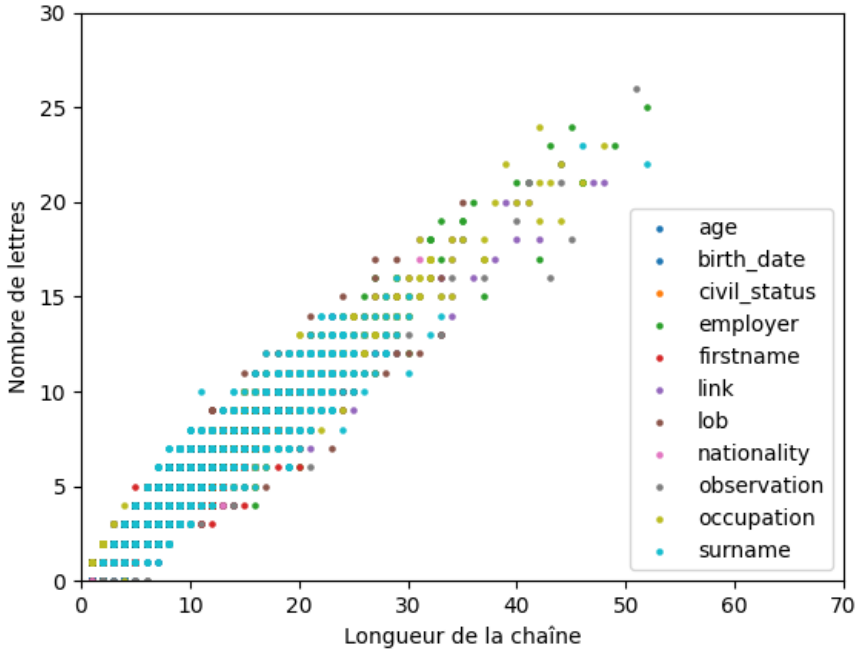
		longueur de chaine	nombre de mots	nb de carc. speciaux	nombre de lettres	nombre de voyelles	nombre de consonnes
age	moyenne	1.962947	1.030604	1.820273	0.112071	0.051046	0.061025
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	14.000000	4.000000	4.000000	10.000000	5.000000	5.000000
birth_date	moyenne	3.997549	1.000408	3.994145	0.002996	0.001362	0.001634
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	9.000000	3.000000	5.000000	9.000000	4.000000	5.000000
civil_status	moyenne	7.818216	1.397384	0.000000	7.420831	3.437366	3.983466
	min	4.000000	1.000000	0.000000	4.000000	2.000000	2.000000
	max	12.000000	2.000000	0.000000	11.000000	6.000000	5.000000
employer	moyenne	7.255670	1.316151	0.014777	6.923368	2.792440	4.130928
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	52.000000	9.000000	3.000000	44.000000	20.000000	25.000000
firstname	moyenne	6.985520	1.092295	0.000000	6.893185	3.376359	3.516826
	min	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000
	max	29.000000	4.000000	0.000000	26.000000	13.000000	14.000000
link	moyenne	7.184414	1.568142	0.003954	6.612220	2.699942	3.912278
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	48.000000	10.000000	2.000000	40.000000	19.000000	22.000000
lob	moyenne	7.675802	1.387890	0.004008	7.253033	3.087088	4.165945
	min	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000
	max	35.000000	7.000000	2.000000	31.000000	14.000000	20.000000
nationality	moyenne	7.285564	1.001728	0.000075	7.283761	3.300811	3.982950
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	31.000000	5.000000	1.000000	27.000000	10.000000	17.000000
observation	moyenne	9.837248	1.864094	0.255034	8.696309	4.026846	4.669463
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	131.000000	24.000000	6.000000	102.000000	48.000000	54.000000
occupation	moyenne	7.823526	1.258746	0.000494	7.562925	3.321671	4.241254
	min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
	max	48.000000	9.000000	4.000000	42.000000	20.000000	24.000000
surname	moyenne	7.207759	1.083595	0.000000	7.124083	3.070852	4.053230
	min	2.000000	1.000000	0.000000	2.000000	1.000000	1.000000
	max	52.000000	10.000000	0.000000	43.000000	21.000000	23.000000

Annexe 2 : Nuage de mots de la catégorie *occupation*



Annexe 3 : Nuage de points la répartition des catégories entre nombre de lettres et nombre de caractères

FIGURE 3 – Nuage de points



Annexe 4 : Nuage de points la répartition des catégories entre nombre de lettres et nombre de caractères