Prédiction du sexe à partir des données personnelles

Yseult Masson - yseult.masson@ensae.fr - Code associé au projet

Introduction

Le but de ce projet est de prédire le sexe des individus recensés entre 1836 et 1936 en France, à partir des données personnelles dont on dispose sur le document de recensement. En effet, le sexe n'est pas toujours renseigné dans ces documents.

1 Présentation des données

Nous disposons d'un tableau contenant les informations personnelles de 241 personnes, recensées en France entre 1836 et 1936. Nous avons accès, selon les personnes, au nom, au prénom, à la date de naissance, au lieu de naissance, à l'employeur, à la relation au chef de famille, à la profession, à l'état civil et à l'éducation.

Ces informations proviennent d'archives papier, et ont donc dû être retranscrites automatiquement, ce qui correspond à la colonne 'prediction' du tableau. Ceci peut être effectué sur chaque document d'archive. Pour ces 241 personnes en particulier, nous avons aussi accès à la 'groundtruth', qui correspond à une retranscription manuelle du document de recensement. Cela ne peut pas être fait à grande échelle, car il s'agirait d'un énorme travail de retranscription. Enfin, nous avons accès au sexe de ces personnes, qui ne figure pas sur les recensements d'origine mais qui a été rajouté manuellement. Par exemple :

- **groundtruth** : surname: Chardon firstname: Marie occupation: idem link: fille age: 30
- **prediction** : nom: Chardon prénom: Marie date_naissance: 30 lieux_naissance: ”
- **sex** : femme

Comme on peut le voir, les transcriptions automatiques ne donnent pas toujours toutes les informations, ni les bonnes informations. Elles peuvent même induire en erreur quant au sexe de la personne, comme "Simon, cultivateur" qui a été retranscrit en "Simone, cubsinière" par le modèle. Cependant, il est nécessaire d'utiliser ces transcriptions et non pas la 'groundtruth' pour pouvoir ensuite appliquer les modèles à des documents qui ont été retranscrits automatiquement et non pas manuellement.

Cette base de données contient 125 hommes, 107 femmes et 9 personnes dont le sexe est ambigu. On retire ces 9 personnes de la base pour faire nos prédictions. De plus, on conserve seulement les informations sur le prénom, la profession et la relation au chef de famille. En effet, les autres données ne sont pas informatives pour prédire le sexe (nom de famille, lieu de naissance) ou n'ont pas été renseignées pour la grande majorité des personnes recensées et contiennent des valeurs aberrantes (éducation, état civil, employeur). De plus, la date de naissance contient parfois l'âge et parfois la date de naissance, et il est impossible de déduire la date de naissance à partir de l'âge et donc d'avoir une colonne 'date de naissance' cohérente.

Nous avons aussi accès, pour une grande partie des prénoms, à la quantité de femmes et d'hommes qui portent ce prénom, ce qui nous permet de calculer à quelle fréquence ce prénom est porté par un homme ou par une femme. On inclut cette information dans notre base.

2 Modèles utilisés

Il semblait intéressant de tenter différentes approches, de la plus naïve à la plus poussée.

2.1 Approche déterministe

La première approche consiste à classifier les individus de la manière suivante:

1. Si le prénom est à plus de 90% masculin ou féminin, on renvoie le sexe indiqué.
2. Si la relation contient 'mère', 'mere', 'fille', 'soeur', 'femme' ou 'brue', on renvoie 'femme', et si elle contient 'père', 'pere', 'fils', 'frere' ou 'frère', on renvoie 'homme'.
3. Si la profession finit par 'ière' ou 'euse', on renvoie 'femme', et si elle contient 'fils', 'ier', 'eur' ou 'patron', on renvoie 'homme'.

4. Si aucun des points précédents n'a donné de résultat, on renvoie 'femme' si le prénom est à plus de 50% féminin, et 'homme' sinon. Pour les prénoms pour lesquels on n'a pas accès à cette information, on renvoie 'homme' (choix arbitraire, car il y a 52% d'hommes dans les données).

L'intérêt de cette méthode est qu'elle est très simple à mettre en place, ne nécessite pas d'entraînement et donc n'a pas de coût de calcul. Elle ne permet cependant pas de déceler des liens moins évidents entre le sexe d'une personne et ses autres informations personnelles.

2.2 Approche *Machine Learning*

La deuxième approche consiste à utiliser un modèle de *Machine Learning* à partir des 4 principales données dont nous disposons : le prénom, la profession, la relation au chef de famille, et le taux de masculinité/féminité du prénom. Étant donné que quasiment toutes les professions sont différentes, et que beaucoup sont mal retranscrites, on conserve seulement l'information "en activité professionnelle" ou "sans activité professionnelle".

Ici, nous utiliserons une forêt aléatoire pour classifier les individus. Cette approche de *Machine Learning* a pour avantage de nécessiter moins de données que du *Deep Learning*, et est donc mieux adaptée à nos 241 observations. Cependant, elle ne prend pas en compte la sémantique des mots et peut donc manquer certaines informations (par exemple, les mots terminant par 'ière' ne seront pas automatiquement classifiés comme féminin).

2.3 Approche *Deep Learning*

Cette dernière approche consiste à utiliser un modèle de langage plus adapté à notre tâche. Nous utilisons le modèle Camembert [1], qui est spécialisé dans la langue française, ce qui nous intéresse ici. On démarre avec un modèle préentrainé, et on *fine-tune* les 2 dernières couches, qui sont dédiées à la classification. De cette façon, toutes les couches profondes du réseau gardent leurs poids de pré-entraînement (qui ont été obtenus à partir d'un énorme volume de données et qui sont donc mieux entraînés que ce que l'on pourrait faire avec nos données), et seule la partie "classification" du réseau, soit celle qui doit être adaptée à nos données, est entraînée.

Cette méthode a pour avantage de prendre en compte la sémantique et la relation entre les mots d'une phrase. Cependant, la quantité et la qualité (beaucoup de fautes

d'orthographes dues à la retranscription) des données dont nous disposons ne sont généralement pas suffisantes pour l'entraînement d'un réseau de neurones, quand bien même nous ne réentrainons pas tout le modèle.

3 Expérimentation et résultats

3.1 Approche déterministe

Pour la première approche, étant donné qu'il n'y a pas d'entraînement, nous pouvons directement mesurer la performance sur toute la base de données. Nous obtenons une précision de 0.94. La matrice de confusion est la suivante:

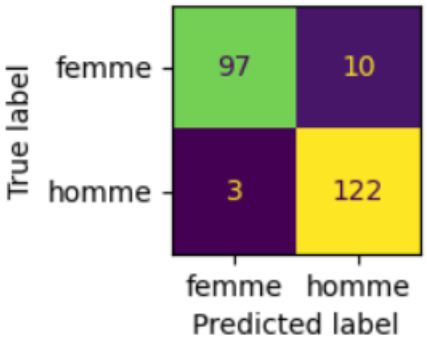

Figure 1: Matrice de confusion - Approche déterministe

Les individus sont donc en majorité bien classifiés.

3.2 Approche *Machine Learning*

Pour cette approche et la suivante, on sépare la base de données en un set d'entraînement (70% de la base, soit 162 individus) et un set de test (70 individus). Afin de trouver les meilleurs hyperparamètres pour la forêt aléatoire, on effectue une *grid search* avec validation croisée. Les détails des différentes étapes se trouvent dans le notebook du projet.

La précision obtenue sur le set de test est, comme pour l'approche précédente, de 0.94.

3.3 Approche *Deep Learning*

Nous utilisons ici, pour chaque individu, une phrase composée des différentes informations utiles dont nous disposons. Par exemple : 'prénom: marie relation: inconnue profession: inconnue fréquence d'hommes : 0.0'. Les détails du *preprocessing* sont indiqués dans le notebook. On entraîne ensuite les couches associées à la classification sur le set d'entraînement.

Afin de tester la robustesse du modèle, nous effectuons plusieurs fois la séparation du jeu de données en sets d'entraînement et de test avec des *seeds* différentes, entraînons le modèle sur le set d'entraînement, et calculons la précision sur le set de test. Les scores obtenus dépendent beaucoup de cette répartition des individus dans les sets d'entraînement et de test : la précision se trouve entre 0.80 et 0.96, sur 8 différents couples entraînement/test. Cette variabilité, qu'on ne retrouvait pas dans l'approche *Machine Learning* (la précision était toujours entre 0.93 et 0.95), est sans doute due au fait que nous ne disposons pas de suffisamment de données pour entraîner correctement un réseau de neurones. Un modèle entraîné sur aussi peu de données ne semble pas pouvoir être généralisé.

Conclusion et recommandations

Au vu des résultats décrits ci-dessus, le modèle le plus adapté serait en fait le plus naïf, à savoir l'approche déterministe. En effet, cette méthode permet d'obtenir une très bonne précision (0.94) et ne nécessite pas d'entraînement, ce qui la rend très peu coûteuse en terme de temps, d'argent et de ressources.

Cependant, si l'on disposait de plus de données, il serait possible d'entraîner un modèle CamemBERT qui pourrait effectuer une analyse plus fine des informations textuelles, et ainsi potentiellement atteindre une meilleure précision.

References

[1] Louis Martin et al. "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.