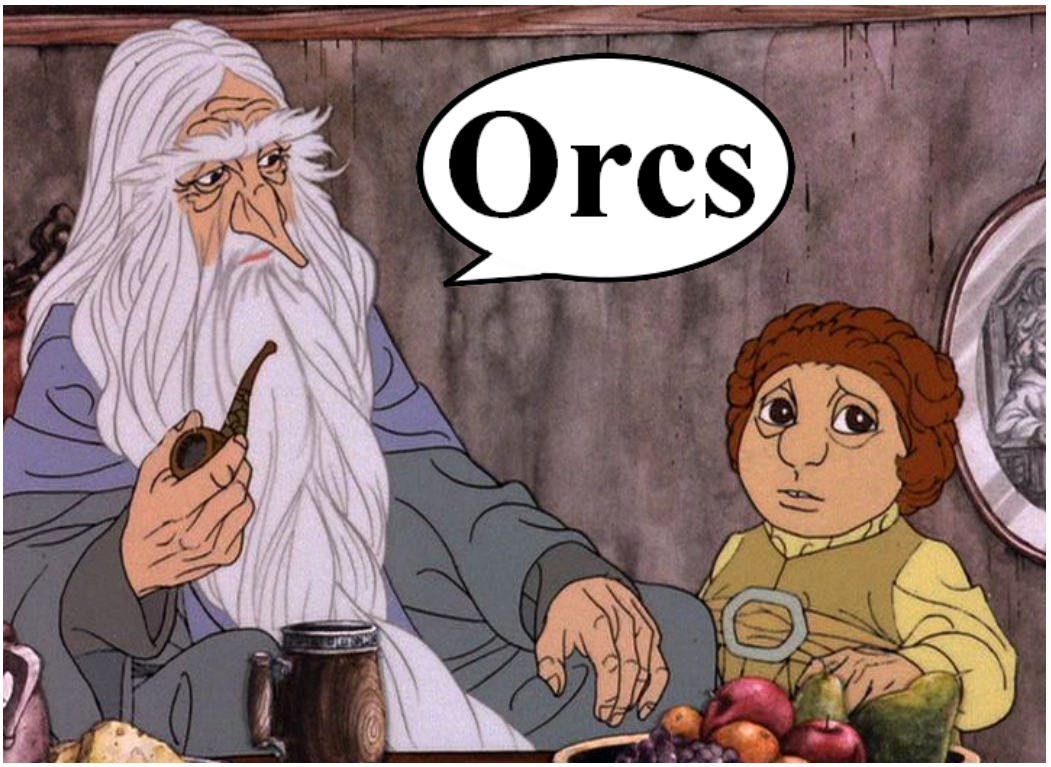
Le Mesle Maxime ET4-INFO

Roblin Clément 2020-2021

Projet d'apprentissage automatique



Prédiction de la race des personnages de l'univers de J. R. R. Tolkien

à partir de leur nom

Introduction :

Notre choix de projet s'est porté sur une base de données regroupant tous les personnages issus du seigneur des anneaux. Cette base associant le nom à la race contenait environ 900 entrées à l'origine, avec des données très hétérogènes, que ce soit sur les noms (plusieurs communautés mélangées dans le cas des hommes et des elfes) ou sur le nombre d'individus (deux corbeaux uniquement contre une quatre centaines d'humains).

Pre-Processing :

Du fait de cette hétérogénéité des données, nous avons décidé dès le départ d'écrémer la base pour retirer les races pas assez significatives afin de limiter leur impact sur la prédiction.

Le sexe n'a également pas été gardé (du moins pas exploité dans le code) malgré plusieurs tentatives de séparer les hommes des femmes. Il n'y a vraisemblablement pas de lien suffisamment important entre le nom et le genre.

Nous avons également réduit le nombre de caractères aux minuscules car les majuscules n'apportent rien de plus. Les chiffres romains ont également été convertis en "*" pour l'analyse. Ce n'est pas important de savoir si tel ou tel chiffre est dans le nom, mais plutôt s'il y a un chiffre.

Par la suite et dans le sens d'optimiser la prédiction nous avons également procédé à régularisation des données en filtrant les personnages avec des noms très éloignés de ce que leur race a habituellement.

Critères d'analyse :

Pour réaliser notre algorithme de prédiction, nous nous sommes basés sur des critères que nous avons jugé pertinents, applicables à nos données, à savoir :

- 1. Fréquence moyenne d'apparition de chaque lettre : chez certaines races, certaines lettres apparaissent souvent, comme le "o" chez les orcs, et certains caractères sont exclusifs à certaines races, comme les chiffres aux nains, hobbits et humains.
- 2. Distribution du nombre de syllabes : les hobbits sont la seule race possédant un prénom et un nom de famille, ainsi leur nom se décompose en beaucoup de syllabes. Au contraire, les nains ont assez peu de syllabes.
- 3. Probabilité d'avoir une lettre sachant une autre lettre dans le mot : en plus de connaître la fréquence de chaque lettre, on peut aussi en déduire quelles lettres sont souvent employées ensemble, comme par exemple les "a" et "u" chez les orcs, les "i" et "n" chez les nains.
- 4. Fréquence d'apparition d'une lettre dans le mot sachant la lettre précédente : les humains et elfes ont très souvent un genre de motif dans leur nom, comme "tar-". De même les noms de famille des hobbits sont souvent les mêmes pour un grand groupe d'individus.
- 5. Fréquence moyenne d'apparition de chaque ensemble phonétique : les sonorités des orcs sont surtout gutturales et vélaires, quand celles des Ainurs sont plus linguales et palatales.

Un critère que nous pensions initialement utile mais qui s'est avéré trop généraliste était de mettre en avant les caractères rares (accents, espaces).

Scoring et hyper paramètres

_____A partir des différents critères implémentés nous nous sommes intéressés à leur impact sur chaque race :

En premier lieu, nous avons établi un système de score:

On calcule les données sur chaque critère pour chaque nom, et on les compare aux données de chaque race. On établit alors un score qui est 1 - somme(abs(x1 - x2)), où x1 est la valeur des données pour une lettre / un phonème du nom, et x2 son équivalent dans les données de la race. Seul le critère sur les syllabes fonctionne différemment. Le score renvoyé est la fréquence du nombre de syllabes chez la race en question.

On a ensuite établi un système de ranking, qui trie les scores obtenus pour le nom pour chaque race du plus faible au plus fort, et distribue des points allant de 0 pour le plus faible à 5 pour le plus fort. Pour éviter de donner du crédit aux valeurs aberrantes, 0 points sont distribués si le score est négatif. L'idée de ce ranking est de ne pas donner plus de poids à certains critères dès le départ.

En analysant toutes les données on obtient en moyenne pour chaque race les scores :

Orcs : distribution = 4.4 syllabe = 4.4 presence = 3.45 succession = 4.75 phonetique = 3.8 Men : distribution = 3.62 syllabe = 1.99 presence = 3.5 succession = 3.91 phonetique = 2.91 Hobbits : distribution = 5.0 syllabe = 4.48 presence = 4.82 succession = 4.46 phonetique = 2.91	Elves : distribution = 3.46 syllabe = 2.84 presence = 2.79 succession = 4.11 phonetique = 3.26 Dwarves : distribution = 4.01 syllabe = 3.17 presence = 2.21 succession = 4.92 phonetique = 4.13 Ainur : distribution = 3.16 syllabe = 4.08 presence = 1.94 succession = 4.72 phonetique = 3.37

A partir de ces données nous avons fait le réglage d'hyper paramètres pondérant chaque critère selon l'importance qu'il a pour la race. Les différences entre les hyper paramètres réels et déduits du tableau ci-dessus viennent de l'impact de la concurrence, qui n'a pas été pris en compte au dessus (notamment, les hobbits font pas mal d'ombre aux humains).

Les autres hyperparamètres sont la taille des données d'entraînement, et le seuil d'élimination du bruit.

Résultats

Avant de parler des résultats en eux-mêmes, on peut remarquer que les données supprimées lors du filtrage, présenté dans le pré-processing, sont judicieuses:

```
data removed : 80
Balin : Dwarves
Elendur of Arnor : Men
Gothmog : Ainur
```

Balin ne sonne pas très nain, Elendur of Arnor est particulièrement long pour un humain, et les lettres de Gothmog sont peu courantes pour un Ainur.

Après 100 itérations de notre code, en recréant à chaque fois les données d'entraînement en prenant des noms aléatoires dans le .csv et en appliquant le même filtrage, sans changer les hyper paramètres, on obtient :

```
A predit ->
Pour |
     ^
matrice de confusion:
[[ 336   86   29   54   47   26]
 [ 874 7122 1209 1826  918  808]
 [ 131  415 6270   58   31   82]
 [ 160 1387  225 1260  193  475]
 [ 107  332   79   87  794   83]
 [  80  491   38  246   57  184]]
accuracy score:
60.02%

Orcs: 58%
Men: 55%
Hobbits: 89%
Elves: 34%
Dwarves: 53%
Ainur: 16%
```

On reconnaît plutôt bien les orcs et les nains, et les hobbits sont la race la mieux reconnue par notre système du fait de leurs noms caractéristiques.

Ce résultat est assez satisfaisant, sachant qu'en faisant le test de notre côté sur un échantillon de 50 noms, nous trouvons toujours entre 60 et 75 % de bonnes réponses.

```
A predit ->
Pour |
     ^
matrice de confusion:
[[ 1  1  0  0  0  0]
 [ 0 19  1  8  0  1]
 [ 0  2 10  0  0  0]
 [ 0  2  0  1  0  0]
 [ 1  0  0  0  2  0]
 [ 0  1  0  0  0  0]]
accuracy score:
66.0%

Orcs: 50%
Men: 65%
Hobbits: 83%
Elves: 33%
Dwarves: 66%
Ainur: 0%
```

Globalement la machine se trompe sur les mêmes choses que nous. Par exemple, elle a tendance à dire que c'est un homme quand il s'agit d'un elfe et inversement. Ce phénomène est lié à la très forte similarité entre des noms d'hommes et ceux d'autres races, en particulier les elfes et les ainurs, ainsi que leur surabondance qui engendre la majorité de nos erreurs.

Là où nous sommes meilleurs que la machine, c'est pour deviner la race d'un individu comprenant des mots en anglais dans son nom, car la traduction de ces mots nous donne un indice sur la race que la machine ne peut pas deviner. Il en découle qu'on a moins tendance à deviner Hobbits quand ce n'est pas le cas.

```
Nom ; Race réelle : Race devinée

Uolë Kúvion ; Men : Hobbits
Dori ; Dwarves : Men
Thingol ; Elves : Ainur
Oromendil ; Men : Elves
Almarian ; Men : Elves
Ecthelion of the Fountain ; Elves : Dwarves
Meleth ; Men : Elves
Déagol ; Hobbits : Men
Walda ; Men : Ainur
Daeron ; Elves : Hobbits
Celebrindor ; Men : Elves
Hob Hayward ; Hobbits : Men
Gríma Wormtongue ; Men : Hobbits
Gil-galad ; Elves : Ainur
Tarannon Falastur ; Men : Hobbits
Lindissë ; Men : Ainur
Soronto ; Men : Hobbits
Amrod ; Elves : Hobbits
Denethor ; Elves : Men
Amarië ; Elves : Men
Damrod ; Men : Hobbits
Larnach ; Men : Dwarves
Ulbar ; Men : Orcs
Ulmo ; Ainur : Orcs
Dwalin ; Dwarves : Hobbits
Lothíriel ; Men : Elves
Will Whitfoot ; Hobbits : Ainur
Hathol ; Men : Orcs
Túrin 1 ; Men : Dwarves
Beril ; Men : Elves
Denethor 2 ; Men : Dwarves
Great Goblin ; Orcs : Hobbits
Hallatan ; Men : Hobbits
Ecthelion 2 ; Men : Dwarves
Ulwarth ; Men : Orcs
Galdor of the Tree ; Elves : Hobbits
Duinhir ; Men : Dwarves
Elanor Gardner ; Hobbits : Men
```

Une partie des erreurs de la machine a été répertoriée dans un fichier .txt, comme sur la capture ci-dessus. On constate que souvent, là où elle s'est trompée, on se serait également fourvoyé.

Conclusion

Nous sommes globalement satisfaits de notre prédiction, les différentes méthodes utilisées pour la prédiction nous semblent les plus pertinentes et efficaces pour ce genre de problème. En revanche, le manque d'homogénéité dans les données semble être le point faible dans cette application. Un découpage plus fin des données (sous-type d'homme et d'elfes) serait par exemple une solution.