Rapport

Handy-Pedro Valery
Charles Dehlinger

Dimanche 14 juin 2020

Table des matières

1 introduction

Dans ce projet, nous avons été amenés à chercher la meilleur façon de détecter la langue maternelle d'un locuteur lorsqu'il s'exprime en anglais.
Nous avons choisit de regarder les mots typiques utilisés par certains locuteurs. En effet certains mots sont transparents dans certaines langues et sont donc plus utilisés par ceux qui la pratiquent.

2 le traitement des données

Les données sont traitées de la façon suivante :
— On sépare les phrases, selon les points (.),
— On découpe chaque phrase en mots grâce à une expression régulière,
— On obtient un tableau de phrases, chaque phrase étant un tableau de mots.

À la suite de cela, on peut travailler phrase par phrase.

3 analyse des données

3.1 les mots

Nous avons d'abord essayé un procédé fondé sur de l'inférence bayésienne. On commence par calculer la probabilité qu'un mot se trouve dans un texte

d'une langue donnée $P_l(m)$, puis la probabilité qu'un texte soit de cette langue $P(l)$ (uniforme dans le cas présent), en suite on calcule la probabilité que ce mot se trouve dans un texte quelconque $P(m)$, enfin on peut estimer la probabilité qu'un texte soit dans un texte sachant qu'un mot s'y trouve $P_m(l)$, grâce à la formule de bayes :

$$P_m(l) = \frac{P_l(m) \cdot P(l)}{P(m)}$$

Malheureusement ce procédé nous a donné des résultats désastreux (à peine plus que le hasard).

3.2 les purs/impurs

Nous avons créé trois catégories de mots pour une langue donnée :

les mots pures , c'est à dire les mots trouvés uniquement dans les textes d'une seule langue.

les mots impures , qui sont des mots trouvés dans certaines langues (dont celle actuellement étudiée),mais pas toutes

les mots globaux , qui ne sont pas présents dans la langue étudiée, mais qu'on peut trouver dans d'autres textes

À partir de là, le fait de trouver un mot pure dans un texte augmente les chances que ce texte appartienne à la langue correspondante. Inversement la présence de mots impures réduit les chances que le texte appartienne à une langue donnée. Ce procédé nous a donné de bons résultats : 30% au début puis 51% après une transformation, en effet, ne travailler qu'avec les mots pures donne un modèle fiable à seulement 25% tandis qu'en introduisant les globaux on augmente le score à 30%. Finalement en introduisant un fonction de minimisation des mots pures, impures et globaux on obtient 51% de bonnes prédictions

3.3 les forêts aléatoires

Pour améliorer les scores, nous avons utilisé les variables précédemment nommées ainsi que des variables générales, comme la longueur du texte ou le nombre le mots par phrase pour nourrir un algorithme de forêt aléatoire. Ce procédé nous a permis d'atteindre un score de 41% environ.

4 Conclusion

Le meilleur modèle est le modèle heuristique, ce qui nous a surpris, et même un peu déçu. Étant donné qu'avant la transformation de celui-ci la forêt aléatoire nous avait donné un résultat de 10 points supérieur, nous nous attendions à un score entre 55 et 60%. Mais cette transformation ne l'a pas affecté.

Cependant nous sommes surpris qu'un modèle aussi simple parvienne a dépasser la barre des 50%, nous pensons même qu'en le consolidant avec un autre modèle qui analyserait un autre aspect des textes (analyse des types d'erreurs, construction syntaxique des phrases, ...) nous aurions pu attendre un score nettement supérieur.