Université de Montréal

IFT6285 (TALN) — Devoir6
Classification d'auteurs

Contact :
Philippe Langlais
RALI/DIRO
Université de Montréal
+1 514 343 61 11 ext: 47494
felipe@iro.umontreal.ca
http://www.iro.umontreal.ca/~felipe/

dernière compilation : 28 octobre 2020 (23:39)

Contexte

Il n'est plus rare que des applications de traitement des langues embarquent des dizaines (voire davantage) de modèles de classification. Classifier un texte est donc une activité courante du traitement des langues. Dans ce devoir, vous allez développer des classificateurs dont le but est d'identifier les auteurs d'un texte. Vous travaillerez pour cela sur un sous-ensemble du corpus de blogs auquel vous avez accès dans votre projet 1.

Données

J'ai extrait du corpus initial différents sous-corpus regroupés dans devoir6.tar.gz (17Mo) qui est également disponible sur les machines du DIRO : `/u/felipe/HTML/IFT6285-Automne2020/blogs/devoir6`.
Dans cette ressource, vous avez accès à différents découpages (*splits*) en *train* et *test* d'un sous ensemble du corpus initial, ou chaque découpage (répertoire) contient les fichiers suivants : `train.csv`, `test-closed.csv` et `test-open.csv` qui sont respectivement le corpus d'entraînement, le corpus de test fermé (les auteurs du test sont ceux vus à l'entraînement) et le corpus de test ouvert (certains auteurs du test n'ont pas été vus à l'entraînement). Les différents sous-répertoires disponibles dans la ressource sont nommés `blogs-`a`-`T`-`t où a désigne le nombre d'auteurs (classes) dans le train, T (resp. t) le nombre de posts de chaque auteur en train (resp. test). Ainsi, le corpus `blog-100-50-10` contient un fichier d'entraînement de 50000 posts (50 posts de 100 auteurs différents), un fichier de test fermé de 1000 posts (10 posts des même 100 auteurs vus en train) et un fichier de test ouvert contenant quelques posts supplémentaires (d'auteurs non vus à l'entrainement). Chaque ligne dans ces fichiers possède le format suivant où la première colonne est l'identificateur de l'auteur, son genre, son âge et le post. Vos classificateurs ne doivent utiliser que le texte, les méta-données sur l'auteur ne sont fournies que pour évaluation et analyse.

```
YZZNOMN,male,15,Ok... Well... Today we had a robotics competition...
```

Notez enfin la présence du répertoire `blog-quiz` pour lequel les informations de l'utilisateur sont masquées dans les fichiers de test. Ces jeux de test seront utilisés pour comparer vos approches.

À faire

1. Vous entraînerez et testerez un classificateur de votre choix pour chaque sous-répertoire distribué.
2. Vous vous intéresserez à la différence de performance entre les conditions ouvertes et fermées et à la taille des modèles utilisés.
3. Vous devez produire un rapport d'au plus 3 pages (format pdf, en français ou en anglais) qui contient les informations suivantes :
   — une description des métriques utilisées pour évaluer les différents classificateurs,
   — les résultats des différents classificateurs,
   — une analyse des résultats
4. Pour le répertoire `blog-quiz`, vous produirez pour `test-open.csv` et `test-closed.csv` la sortie de votre meilleur classificateur dans des fichiers `test-open.quiz` et `test-closed.quiz` au format suivant : pour chaque ligne de test, vous indiquerez l'identificateur de l'auteur présumé si vous pensez qu'il s'agit d'un auteur vu à l'entraînement, ou AUTRE si vous pensez qu'il s'agit d'un autre auteur. Chaque fichier doit avoir le mêne nombre de lignes que le fichier de test correspondant. La nature de votre classificateur doit être identifiée dans votre rapport. Ce fichier sera évalué par nos soins à des fins de comparaison.

Remise

La remise est à faire sur Studium sous le libellé `devoir6`. Vous devez remettre votre code, votre rapport (format pdf, texte en anglais ou en français) et les fichiers `test-open.csv` et `test-closed.csv` dans une archive (gzip, tar, tar.gz) dont le nom est préfixé de `devoir6-name1` ou `devoir6-name1-name2` selon que vous remettez seul ou a deux, où `name1` et `name2` sont à remplacer par l'identité des personnes faisant la remise (`prénom_nom`). Assurez vous que le nom des personnes impliquées dans le devoir soit indiqué sur tous les documents remis (code et rapport). Le devoir est à remettre en groupe d'au plus deux personnes au plus tard vendredi 6 novembre à 23h59. **Note : Aucun modèle n'est demandé.**