Université de Montréal

IFT6285 (TALN) — Projet 1
Classification de blogs

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 : 14 octobre 2020 (12:54)

Contexte

Écrire sur les réseaux sociaux n'a jamais été aussi facile et de tels écrits laissent des traces (style, sujets d'intérêt, informations personnelles, etc.) qui font la joie du GAFAM et autres compagnies des technologies du Web. Dans ce projet, vous allez tenter de retrouver des informations à partir d'un *post*, comme la tranche d'âge du blogger, son genre (homme/femme), ou son signe zodiaqual.

Données

Vous avez accès à un corpus extrait de The Blog Authorship Corpus (Schler et al. 2006). Le corpus blog.tar.gz (276 Mo) est composé d'un total de 18 819 bloggers qui ont écrit un peu plus de 657k posts. Les bloggers ont été répartis en `train` et `test`, tel qu'indiqué en Table 1. Les blogs sont encodés en utf-8 et sont sauvés sous format csv (séparateur : virgule). Chaque fichier correspond à un blogger, chaque ligne à un post.

Une copie locale disponible sur les machines du DIRO se trouve ici : `/u/felipe/HTML/IFT6285-Automne2020/blogs/`.

	# bloggers	# posts
test	500	14 303
train	18319	643 256

Table 1 – Caractéristiques du corpus distribué

Voici deux posts pris au hasard, chacun sur une ligne. On y trouve dans cet ordre : l'ID du blogger, le genre, l'âge, le signe zodiaqual et le blog. Vous remarquerez que les posts ont subi quelques pré-traitements. Notamment, certaines urls ont été remplacées par le *token* `urllink`.

```
ZZS7Z21,female,24,Libra,urlLink    The man of the hour  urlLink
ZZS7Z21,female,24,Libra,"A few months ago while living in Austin,
I finally conquered my mini-tragedies and made the decision to
relocate to Silver Spring, Maryland for the summer.  Silver Spring,
Maryland is an affluent and sleepy little suburb at the end of the ...
```

À faire

1. vous devez développer des classificateurs prenant en entrée un post et produisant :

   la tranche d'âge du blogger (0=[0-19], 1=[20-29] et 2=[30 et plus]),

   son genre (male, female),

   son signe zodiaqual (Aquarius, Aries, Cancer, Capricorn, Gemini, Leo, Libra, Pisces, Sagittarius, Scorpio, Taurus, Virgo).

   Les données étant regroupées par blogger, vous pouvez développer **également** des classificateurs qui prennent en entrée tous les écrits d'un blogger (plutôt que chaque post), mais ce qui vous est demandé est d'étiqueter un post.

   Aucune contrainte n'est imposée sur la nature des classificateurs. Vous pouvez utiliser des règles, des classificateurs à base de traits (*features*), des réseaux de neurones profonds, ou une combinaison de tout cela.

2. Vous devez produire un rapport d'au plus 6 pages (format pdf, en français ou en anglais) qui contient au moins les informations suivantes :
   - une description des données et leur particularité.
   - une description des pré-traitements effectués (si pré-traitements il y a).
   - une description d'un ou plusieurs systèmes de base (*baselines*) auxquels vous vous êtes comparé.
   - une section de résultats rapportant les performances des différentes approches testées ainsi qu'une analyse de ces résultats. Vous décrirez les métriques avec lesquelles vous évaluez vos systèmes.
   - vous indiquerez sur quelle architecture machine vous avez travaillé et préciserez les librairies utilisées, ainsi que toute donnée externe utilisée.

3. une semaine environ avant la fin du projet, un autre corpus de test vous sera distribué pour lequel vous aurez à produire un fichier au format csv avec votre meilleur système. Le corpus (au format csv) comprendra des lignes correspondant à un post, et chaque ligne comportera deux champs : ID et texte :

```
BLIDE4B,"EXFGUGA IS currently down do to bandwidth levels*,
so there's not MUCH to talk about.
HBY89FN,Matt is  A losEr -* yeah. kind of im awfully tired
because emily wouldn't hang up.
```

Vous devrez produire un fichier `best.csv` qui contient pour chacune de ces lignes (séparateur : la virgule) une ligne avec comme information : l'identificateur du blogger, son genre (male, female), sa tranche d'âge (entre 0 et 2) et son signe prédits :

```
BLIDE4B,female,0,Scorpio
HBY89FN,male,2,Virgo
```

Ce fichier sera évalué par mes soins à des fins de comparaison de vos approches. Vous accompagnerez ce fichier d'un fichier `readme` qui décrira très brièvement le système utilisé pour produire `best.csv`. Vous prendrez dans ce `readme` d'identifier toute ressource externe utilisée par votre solution (lexiques, modèles, données autres), aussi le cas échéant, d'indiquer si votre approche agrège plusieurs posts d'un même blogger au moment de prendre une décision sur un post particulier.

Notation

La notation n'est pas corrélée à la performance de vos approches, mais à la **curiosité** que vous développez et à votre esprit d'**analyse**. La clarté et l'informativité de vos rapports sont deux critères importants.

Remise

La remise est à faire sur Studium sous le libellé `projet1`. Vous devez remettre votre code, votre rapport (format pdf, texte en anglais ou en français), ainsi que les fichiers `best.csv` et `readme` dans une archive (gzip, tar, tar.gz) dont le nom est préfixé de `projet1-noms`, où `noms` sont à remplacer par l'identité des personnes (`prénom_nom`) impliquées dans le projet. Donc si j'avais à remettre seul mon solutionnaire au projet1, je le ferais sous le nom `projet1-philippe_langlais.tar.gz`. Assurez vous que le nom des

personnes impliquées dans le projet soit indiqué sur tous les documents remis. Le projet est à remettre en groupe d'au plus trois personnes au plus tard mardi 10 novembre à 23h59.
Note : Aucun modèle n'est demandé : juste le rapport, le code, et les fichiers `best.csv` et `readme`.

SCHLER, J. et al. (2006). "Effects of Age and Gender on Blogging". In : *AAAI Spring Symposium : Computational Approaches to Analyzing Weblogs*.