Université de Montréal

IFT6285 (TALN) — Devoir5
Plongement de mots

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 (14:55)

Contexte

La représentation vectorielle de mots est une avancée majeure en traitement des langues. Dans ce devoir, vous allez manipuler la librairie gensim afin d'entraîner des plongements de mots (*embeddings*) statiques sur le corpus de blogs auquel vous avez accès dans votre projet 1.

Données

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. Vous pouvez vous concentrer sur le sous-ensemble `train`. Je vous rappelle qu'une copie locale est disponible sur les machines du DIRO :
`/u/felipe/HTML/IFT6285-Automne2020/blogs/`.

À faire

1. vous devez écrire du code permettant d'entraîner avec gensim des représentations vectorielles sur tout ou partie du sous-ensemble `train`.
2. Vous devez produire un rapport d'au plus 3 pages (format pdf, en français ou en anglais) qui contient les informations suivantes :
   — une courbe montrant les temps d'entraînement en fonction du nombre de phrases traitées (vous n'avez bien sûr pas besoin d'y aller par pas de 1 phrase !).
   — une étude de l'influence de meta-paramètres comme la taille du contexte, la dimension d'un vecteur, le nombre d'exemples négatifs. Cette étude peut porter sur la qualité des plongements obtenus, mais il vous faut alors une mesure de cette qualité, ou peut porter sur les temps d'entrainement.
   — pour un modèle que vous avez entraîné (et dont vous spécifierez les détails) un calcul des (au plus) 10 mots les plus proches des mots de cette liste de mots (utf8). Vous devez remettre un fichier voisins (utf8) dont le format est spécifié par l'exemple, à savoir un mot (de la liste) par ligne, suivi d'une tabulation, suivie des 10 mots

les plus proches avec leur score de similarité entre crochets. Si un mot de la liste n'est pas connu de votre modèle, ne mentionnez pas ce mot dans votre fichier.

Remise

La remise est à faire sur Studium sous le libellé `devoir5`. Vous devez remettre votre code, votre rapport (format pdf, texte en anglais ou en français) et le fichier `voisins` dans une archive (gzip, tar, tar.gz) dont le nom est préfixé de `devoir5-name1` ou `devoir5-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 23 octobre à 23h59.

Note : Aucun modèle n'est demandé : juste le rapport, le code et le fichier `voisins`