Université de Montréal

IFT6285 (TALN) — Devoir3
Plongements de mots statiques avec gensim

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 : 31 octobre 2023 (23:08)

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 tout ou partie du 1B-word corpus (Chelba et al. 2013) dont une copie est disponible au DIRO (voir devoirs passés).

À faire

Utilisez gensim pour entraîner des représentations vectorielles sur tout ou partie du 1BWC. Produisez un rapport pdf qui contient les informations suivantes :

1. 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 une phrase. Vous pouvez par exemple y aller par tranche, voire à une granularité plus large (ex : 10, 20, 30, ... tranches). Vous prendrez soin d'indiquer la configuration de l'ordinateur sur lequel vous avez réalisé votre devoir.
2. une étude en au plus deux pages de l'influence de méta-paramètres comme la taille du contexte, la dimension d'un vecteur, le nombre d'exemples négatifs. Votre but sera d'optimiser la tâche du benchmark TOEFL, mais vous pouvez aussi étudier d'autres benchmarks populaires sur lesquels les plongements statiques ont été testés.
3. 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-<noms>` (où `<noms>` est remplacer par votre/s nom/s) au format spécifié par cet exemple (utf8) 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.
4. quel est selon vous le pourcentage de synonymes présents dans cette liste de voisins sémantiques ? Indiquez brièvement comment vous avez estimé ce pourcentage.

Remise

La remise est à faire sur Studium sous le libellé `devoir3` sous la forme d'une archive de nom `devoir3-<noms>.tar|tar.gz|zip|gzip` où `<noms>` est à remplacer par le nom des personnes concernées par la remise. Cette archive doit contenir votre code, votre rapport (de nom `devoir3-rapport-<noms>.pdf` au format pdf, texte en anglais ou en français) et le fichier `voisins-<noms>`.

Le devoir est à remettre en groupe d'au plus deux personnes au plus tard jeudi 9 novembre à 23h59.

Note : Aucun modèle n'est demandé : juste le rapport, le code et le fichier `voisins-<noms>`.

CHELBA, Ciprian et al. (2013). "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling". In : *CoRR* abs/1312.3005. URL : http://arxiv.org/abs/1312.3005.