Cours TAL – Labo 5 : Le modèle word2vec et ses applications

Objectif

Le but de ce labo est de comparer un modèle *word2vec* pré-entraîné avec deux modèles que vous entraînerez vous-mêmes, sur deux corpus de tailles différentes. La comparaison se fera sur une tâche de similarité mots et sur une tâche de raisonnement par analogie, en anglais. Vous utiliserez la librairie Gensim de calcul de similarités pour le TAL.

Consignes

- Veuillez suivre les étapes indiquées ci-après, en écrivant votre code, vos résultats et vos réponses aux questions dans un notebook Jupyter, que vous soumettrez à la fin sur Cyberlearn.
- Il sera utile de bien lire la documentation de Gensim sur word2vec, et surtout celle sur la classe KeyedVectors (qui représente des vecteurs de mots), qui fournissent des exemples utiles.
- Les différentes tâches se feront soit sur votre propre ordinateur (si possible avec au moins 8 Go de RAM), soit sur le service en ligne Google Colab.

1. Tester et évaluer un modèle déjà entraîné sur Google News

Installez `gensim`, une librairie Python qui fournit des outils pour travailler avec Word2Vec (avec `conda` ou avec `pip`). Attention : la dernière version 4.2.3 de `gensim` est incompatible avec la librairie `scipy` version 1.13, donc il faut installer la version 1.12 de `scipy` ; la variable `Path` doit contenir `C:\ProgramData\Miniconda3\Library\` et `C:\ProgramData\Miniconda3\Library\bin\`.

Obtenez depuis `gensim` le modèle `word2vec` pré-entraîné sur le corpus Google News en écrivant : w2v_vectors = `gensim.downloader.load("word2vec-google-news-300")`, ce qui téléchargera le fichier la première fois.

Après avoir téléchargé le modèle, vous pouvez utiliser ainsi votre copie locale :
`w2v_vectors = KeyedVectors.load_word2vec_format(path_to_file, binary=True)`.

a. Quelle place en mémoire occupe le processus du notebook avec les vecteurs de mots ?

b. Quelle est la dimension de l'espace vectoriel dans lequel les mots sont représentés ?

c. Quelle est la taille du vocabulaire connu du modèle ? Veuillez afficher 5 mots anglais qui sont dans le vocabulaire et deux qui ne le sont pas.

d. Quelle est la distance entre les mots *rabbit* et *carrot* ? Veuillez expliquer en une phrase comment on mesure les distances entre deux mots grâce à leurs vecteurs.

e. Considérez au moins 5 paires de mots anglais, certains proches par leurs sens, d'autres plus éloignés. Pour chaque paire, calculez la distance entre les deux mots. Veuillez indiquer si les distances obtenues correspondent à vos intuitions sur la proximité des sens des mots.

f. Pouvez-vous trouver des mots de sens opposés mais qui sont proches selon le modèle ? Comment expliquez-vous cela ? Est-ce une qualité ou un défaut du modèle word2vec ?

g. En vous aidant de la documentation de Gensim sur `KeyedVectors`, calculez le score du modèle word2vec sur les données **`WordSimilarity-353`**. (Cette documentation vous indiquera aussi comment récupérer le fichier.) Expliquez en 1-2 phrases comment ce score est calculé et ce qu'il mesure.

h. En vous aidant de la documentation, calculez le score du modèle word2vec sur les données **`questions-words.txt`**. *Attention, cette évaluation prend une dizaine de minutes, donc il vaut mieux tester d'abord avec un fragment de ce fichier, extrait par copier/coller.* Expliquez en 1-2 phrases comment ce score est calculé et ce qu'il mesure.

2. Entraîner deux nouveaux modèles word2vec à partir de deux corpus

a. En utilisant `gensim.downloader`, récupérez le corpus qui contient les 10^8 premiers caractères de Wikipédia (en anglais) avec la commande : `corpus = gensim.downloader.load('text8')`. Combien de phrases et de mots (*tokens*) possède ce corpus ?

b. Entraînez un nouveau modèle word2vec sur ce nouveau corpus (voir la documentation de Word2vec). Si nécessaire, procédez progressivement, en commençant par utiliser 1% du corpus, puis 10%, etc., pour contrôler le temps que cela prend.

- Veuillez indiquer la dimension choisie pour le *embedding* de ce nouveau modèle.
- Combien de temps prend l'entraînement sur le corpus total ?
- Quelle est la taille (en Mo) du modèle word2vec résultant ?

c. Mesurez la qualité de ce modèle comme en (1g) et (1h). Ce modèle est-il meilleur que celui entraîné sur Google News ? Quelle est selon vous la raison de la différence ?

d. Téléchargez maintenant le corpus quatre fois plus grand constitué de la concaténation du corpus *text8* et des dépêches économiques de Reuters (413 Mo) fourni en ligne par l'enseignant et appelé `wikipedia_augmented.dat`. Entraînez un nouveau modèle word2vec sur ce corpus, en précisant aussi la dimension choisie pour le plongement (*embedding*).

- Utilisez la classe `Text8Corpus()` pour charger le corpus et pour faire la tokenisation et la segmentation en phrases.
- Combien de temps prend l'entraînement ?
- Quelle est la taille (en Mo) du modèle word2vec résultant ?

e. Testez ce modèle comme en (1g) et (1h). Est-il meilleur que le précédent ? Pour quelle raison ?