Prof. Amal Zouaq – INF8460 (Automne 2019)
Chargé de laboratoire : Félix Martel

École Polytechnique de Montréal
Département Génie Informatique et Génie Logiciel
INF8460 – Traitement automatique de la langue naturelle

TP3

Objectifs d'apprentissage

- Implanter des modèles de classification neuronaux
- Utiliser des plongements lexicaux pré-entrainés

Logiciels

- Keras
- Scikit-learn
- NLTK
- ipython Notebook

Pour ce TP, il est **fortement recommandé** d'utiliser une machine avec un GPU. Vous pouvez par exemple utiliser les machines du lab L-4818 ou travailler sur Google Colab.

Modalités de remise du TP

Vous devez compléter le squelette `inf8460_tp3.ipynb` et le soumettre sous le nom `matricule1_matricule2_matricule3_TP3.ipynb` qui reprend les différentes questions, et implante les fonctionnalités requises.
Vous devez retourner un zip qui contient :

- Le notebook complété
- Une version HTML de votre notebook

Critères d'évaluation

- L'exécution correcte du code et sa qualité
- Le professionnalisme du notebook
- La clarté des explications et commentaires qui l'accompagnent

Corpus et code

Le zip du TP contient :

- Le corpus de revues de films du TP2
- Le squelette du notebook qui doit être complété

Prof. Amal Zouaq – INF8460 (Automne 2019)
Chargé de laboratoire : Félix Martel

Travail à faire

Dans ce TP, on travaille à nouveau sur la classification des revues de films, avec les mêmes données qu'au TP précédent.

Pour chacun des modèles suivants, indiquez ses performances et ses spécifications (nombre d'époques, régularisation, optimiseur, nombre de couches, etc.). N'hésitez pas à expérimenter avec différents paramètres. Vous ne devez reporter que votre meilleure expérimentation.

a) En utilisant Keras, on vous demande de développer un perceptron multi-couches simple pour l'analyse de sentiments. (*30 points*)

b) En utilisant Keras, on vous demande de développer un bi-LSTM pour l'analyse de sentiments. (*30 points*)

c) On souhaite améliorer la représentations des mots en utilisant des vecteurs d'un modèle word2vec pré-entrainé English Wikipedia (enwiki_upos_skipgram_300_5_2017) disponible à vectors.nlpl.eu/explore/embeddings/en/models/ . Ré-entraînez votre modèle bi-LSTM avec ces vecteurs. Essayez d'obtenir le meilleur modèle possible en vous basant sur les pistes d'amélioration suggérées (*30 points*)

d) Indiquez les performances des modèles, leurs spécifications, la durée d'entraînement et commentez ces résultats. Quelle est l'influence des *word embeddings* sur les performances du LSTM ? Comparez les performances de tous vos modèles avec celles du modèle n-grammes du TP2 et du modèle Naive Bayes. Quel est votre meilleur modèle ? (*10 points*)