Projet NLP : Quora Question Pairs
Démarche de Travail

Justin AYIVI, Samy FERRAT, Othmane ZARHALI and Sebastien CHAILLOU
Université Paris Dauphine

I. Description du Jeu de données

Le jeu de données est un ensemble de questions issues du site Quora, il est décomposé en un ensemble train de 404290 lignes et en un ensemble test de 1048574 lignes.
Le jeu de donnés est composé de 4 colonnes :
Les réponses obtenues pour la variable cible proviennent d'humains et donc elles sont sujettes à d'éventuelles erreurs ou d'avis subjectif qui ne rendent pas parfois réellement compte du sens observé.
Le but est donc de déterminer si deux questions ont un sens similaire.

II. Analyse exploratoire

La tâche à laquelle nous faisons face est une classification de textes. Mais avant de proposer des modèles de Machine Learning adaptés pour le traitement des données séquentielles, nous effectuerons une brève analyse de ces données textuelles. Les méthodes d'analyse factorielle sont ainsi un excellent moyen d'avoir une vue d'ensemble sur les données et favorisent la compréhension de ces données d'un point de vue multidimensionnel. Nous proposons ainsi la méthode t-SNE qui est une méthode efficace pour représenter dans un espace de faible dimension l'essentiel de l'information contenue dans un espace de dimension élevée. Une fois qu'on a une visualisation et donc une information sommaire sur les données, on pourra passer à la phase de pré-processing.

III. Preprocessing

La phase de pré-processing est une étape très importante dans tout processus d'apprentissage et détermine notamment la qualité de ce dernier. Nous travaillons sur les données textuelles et comme ces données ne sont pas directement exploitables par les modèles de Machine Learning, une première phase consistera à transformer le jeu de données en tokens et à affecter un vecteur de poids à ces tokens. Ce processus de transformation est connu sous le nom de la vectorisation (Embedding en anglais). On distingue néanmoins différents types de vectorisation notamment la vectorisation classique et la vectorisation pré-entraînée. Le choix de la vectorisation dépend notamment du projet spécifique. Nous proposons plutôt d'utiliser des modèles pré-entraînés pour notre projet. Ce choix se justifie en grande partie par l'existence de ces modèles sur des sujets semblables à notre cadre d'étude. Dans cette optique, les modèles issus de Word2vec nous semblent très pertinents. Avant d'alimenter notre modèle de Deep learning, nous constituerons un ensemble de validation qui servira à l'évaluation de notre modèle.

IV. Choix du modèle RNN

Le traitement des données textuelles intègre une caractéristique importante qui est celui d'un lien chronologique dans les données. C'est pour cela que le modèle qui nous semble adéquat est celui d'un réseau neuronal récurrent. Sa principale force réside dans le fait que contrairement aux réseaux de neurones classiques (forward based), il permette d'intégrer une sorte de lien entre les différentes entrées fournies au modèle. Cependant, en pratique, il lui est impossible d'apprendre des dépendances à long terme dans les données. On

utilise donc des modèles alternatifs tels les modèles LSTM et les GRU. Ainsi, les premiers modèle d'entrainement qui seront utilisés sur le jeu de données sont les modèle LSTM / GRU. Une temps précieux sera accordé au réglage des hyperparamètres du modèle et de l'architecture du réseau récurrent.

V. Validation du modèle

Une fois que ce modèle a été appliqué aux données, la phase suivante sera consacrée à l'évaluation de la performance du modèle sur l'ensemble de validation. Suivant la valeur obtenue par les différentes métriques, différents diagnostic seront effectués sur le modèle. On se concentrera sur deux problèmes principaux que sont le surajustement et le sous ajustement. Pour éviter le surajustement en particulier, on aura recours à différentes architectures telles que les modèles Bi-Directionnels récurrents ou encore l'ajout des piles de couches récurrentes. La technique du dropout sera également appliqué afin de régulariser le modèle. En ce qui concerne le sous ajustement, nous serions peut être amené à apporter un regard plus pertinent à la structure de la variable cible (contient potentiellement des informations mal répertoriées).