SWORD
UPGRADE YOUR BUSINESS

Stage Master II / Ingénieur

Réf : [SIP-Offre 12192]

Deep learning pour la détection automatique de similarités de noms de marques

Contexte

SWORD-GROUP, ESN déployée à l'international, est un acteur majeur dans le secteur de la propriété industrielle.

Dans le cas d'un dépôt de marque verbale, la recherche d'antériorité doit permettre de s'assurer que le nom de la marque envisagée n'est pas similaire à celui d'une marque existante.

SWORD édite un logiciel permettant de comparer un mot ou un groupe de mots à différentes bases de données contenant des millions de noms de marques et slogans déposés.

L'examen de la similarité de noms de marques nécessite la prise en compte de problématiques « métier » particulières:

- Comparaison cross-lingue, étendue à la phonétique et la translitération des termes
- La similarité perçue dépasse le strict agencement des lettres et mots
  - **FIREFOX** perçu comme plus proche de **FYREFOS** que de **FUREFON**
- L'importance à accorder aux mots est typée (selon leur caractère fantaisiste, suggestif, descriptif, générique, mots outil,…) et dépend du contexte
  - **ADIDAS SPORTSWEAR** en conflit avec **PADIDAS SHOES** mais pas avec **NIKE SPORTSWEAR**
- La synonymie peut être utilement examinée
  - **RED STAR** vs **BLUE MOON**

Objectif

La méthode de comparaison verbale existante s'appuie principalement sur des algorithmes déterministes (expressions régulières, distance d'édition,…).

L'objectif du stage est de compléter ces outils avec des méthodes issues de machine learning permettant une meilleure prise en compte du sens et de l'importance des mots.

Missions

- Prise en main et test approfondi des solutions existantes, mise en place d'un benchmark de mesure de qualité et performance.
- Sélection de données d'apprentissage utiles issues d'un large historique de recherches de noms de marques
- Implémentation d'un moteur de comparaison de noms de marques et slogans :
  - Apprentissage d'une comparaison syntaxique indépendante du contenu et sens du texte, basée sur le char embedding
  - Comparaison sémantique de similarités conceptuelles (Word2vec/fastText)
  - Apprentissage de la pondération des termes pour se focaliser sur le contenu utile
  - Compression de l'information pour passage à l'échelle et performance (Product Quantization)
- Comparaison des résultats obtenus avec les résultats actuels (qualité, rapidité, souplesse)

Profil recherché

- Elève ingénieur 3ème année ou M2, à dominante informatique ou mathématiques appliquées.
- Bonnes connaissances en machine learning et bon niveau de programmation en Python.
- Fort intérêt pour NLP, recherche d'information et deep learning.
- Expérience d'au moins un environnement de développement type TensorFlow ou PyTorch.

Environnement

Environnement technique: Windows server, TensorFlow, Keras, Python, NLTK, Gensim, SQL Server.

Vous serez intégré(e) dans l'équipe produit, au sein de la section R&D DL. Une poursuite en thèse dans le cadre d'un financement CIFRE ou apparenté est envisagée, dans le cadre d'un partenariat avec le CNAM. (http://cedric.cnam.fr/~thomen)

www.career.sword-group.com

PARIS
37 rue de Lyon
75012 Paris
+33 1 44 67 24 00

Suivez **Sword**

ekaterina.untilova@sword-group.com