Sujet de stage de Master 2 Recherche

Méthodes d'apprentissage profond pour le traitement automatique de la langue et l'analyse textuelle

Structure d'accueil

— Laboratoire : Laboratoire LJAD, UMR CNRS 7135, Université Côte d'Azur
— Encadrants :
  — Charles Bouveyron (charles.bouveyron@unice.fr)
  — Marco Corneli (Marco.corneli@unice.fr)
  — Pierre Latouche (pierre.latouche@parisdescartes.fr)
— Localisation : Laboratoire LJAD, Faculté des Sciences, Parc Valrose, 06000 Nice
— Page web : http://math.unice.fr/~cbouveyr/

Renseignements relatifs au stage

— Période : 6 mois (avril – septembre 2018)
— Rémunération : Gratification usuelle de stage (~ 500€ nets / mois)

Sujet du stage

Dans tous les aspects de la vie quotidienne, on assiste à une digitalisation des systèmes qui est de plus en plus importante. Une des conséquences de ce phénomène est la production massive de données, en particulier de données textuelles. Par exemple, certains sites internets de e-commerce (tels que Amazon ou TripAdvisor) demandent à leurs clients de commenter les biens/produits qu'ils ont achetés.

Comme les textes sont parfois très longs et/ou hétérogènes, l'analyse statistique se révèle un instrument très utile pour fournir une vision synthétique des corpus textuels. Les buts sont multiples : sélectionner les mots les plus représentatifs dans chaque texte, estimer la probabilité d'extraction d'un mot par rapport à un sujet de conversation, détecter les sentiments de l'auteur (par exemple avis favorable ou défavorable), etc. Une des approches les plus connues et utilisées pour la classification non supervisée des textes est Blei et al. [2003]. En revanche, cette méthode ne prend pas en compte l'ordre des mots et, d'un point de vue génératif, elle peut produire des textes qui n'ont pas de sens (les mots sont tirés par hasard dans un dictionnaire). D'autres méthodes plus récentes sont liées au traitement automatique de la langue [NLP Chowdhury, 2003], un domaine de l'intelligence artificielle qui vise à rendre les ordinateurs capables de comprendre et manipuler le langage humain. En ce qui concerne le NLP statistique, les réseaux de neurones profonds (deep learning) sont utilisés pour faire de l'analyse et de la prévision textuels [voir par exemple Collobert and Weston, 2008].

L'objet du stage sera, dans un premier temps, de faire l'état de l'art sur les techniques de deep learning pour le NLP. Dans un deuxième temps, le stage portera sur la programmation d'un algorithme d'analyse statistique des textes qui sera basé sur une de ces techniques. Une piste serait d'utiliser des auto-encoders pour réduire la dimension des corpus et sélectionner les thèmes (topics) les plus importants.

Profil du candidat

Le/la candidat(e) devra être étudiant(e) en master 2 recherche en Mathématiques Appliquées ou Informatique. Il/Elle devra maîtriser les notions et langages suivants :

— statistique multivariée, régression logistique et analyse des données (classification, clustering et idéalement réseaux de neurones),
— estimation (algorithme EM, algorithme de Netwon-Raphson),
— logiciel R ou Python (C/C++ sont également appréciés)

Références

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3 :993–1022, 2003.

Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1) :51–89, 2003.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.