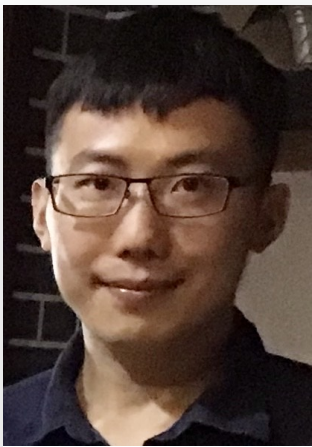




JINGSHU LIU

Data scientist

✉ jingshu.liu@ls2n.fr

📍 France

in <https://www.linkedin.com/in/jingshu-liu-6ba9b679>

🐙 <https://github.com/jingshu-liu>

📁 <https://bitbucket.org/stevall/data-crawler>

🔔 [869436889](#)

RÉSUMÉ

- Travaille actuellement chez Easiware-Dictanova en tant que data scientist en préparant mon doctorat en traitement du langage naturel dirigé par Emmanuel Morin.
- Faire des recherches sur NLP et l'apprentissage automatique, en se concentrant sur les applications multilingues et la modélisation de séquences avec l'apprentissage par transfert en utilisant des modèles de langue pré-entraînés.
- Avec une capacité pratique sur l'apprentissage automatique et l'apprentissage en profondeur, je m'intéresse largement à l'apprentissage automatique appliqué pour les scénarios industriels et aux systèmes distribués.

EXPERIENCE

Data scientist/Machine learning

Easiware

2017 - Présent

- Construire à partir de zéro un cadre de mapping de mots et de phrases basé sur un réseau neuronal bilingue en Java avec Deeplearning4j-0.91.
- Implémenter les pipelines pour la modélisation des thématiques en utilisant le clustering sur des représentations de phrases unifiées pré-entraînés.
- Effectuer des optimisations pour les matrices creuses et d'autres opérations mathématiques pour Nd4j-0.91.
- Dessiner et construire un framework encodeur-décodeur pour la modélisation de séquences avec Pytorch-1.2. Compatible en mode CPU et GPU exécuté sur le serveur cloud OVH à l'aide des outils de gestion openstack et nvidia gpu cloud.
- Incorporer des modèles de langue basés sur Transformer pré-entraînés dans nos réseaux de neurones pour des scénarios réels.
- Fine-tuner des modèles de langue pré-entraînés pour des dialogues.

Achievements:

- Améliorer l'induction de lexique bilingue des multi-mots et des mots simples par une moyenne de 22 points en MAP sur les données client.
- Le nouveau système de modélisation de thématiques a remplacé le système existant.

Environements:

Java Python Pytorch Deeplearning4j Keras Scikit Learn OpenStack-Docker

Stagière en Natural Language Processing

Dictanova

2016

- Implémenter un système d'extraction de termes et d'analyse de sentiments basée sur l'aspect(ABSA) pour le chinois simplifié et traditionnel avec l'architecture UIMA et le stockage Elasticsearch.
- Améliorer le prétraitement en chinois (étiquetage POS) pour FNLP. En plus, nous avons inventé un lemmatiseur original pour les mots redupliqués.
- Nettoyage et visualisation des données avec Pandas et R.

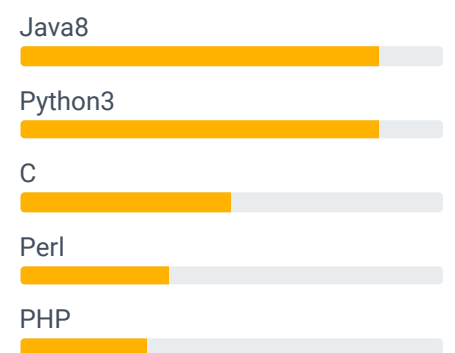
Accomplissement:

- Résultats état-de-l'art pour l'analyse des sentiments basée sur l'aspect sur le défi Semeval2016.
- Améliorer la précision pour l'extraction terminologique de 50%.

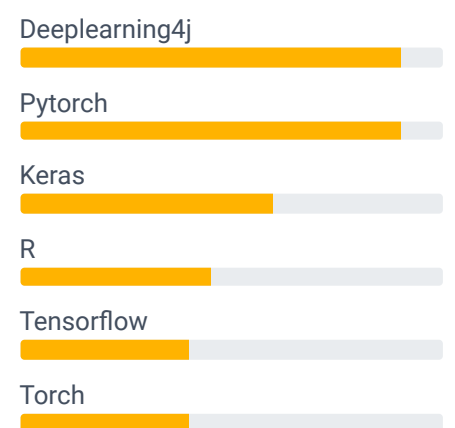
Environements:

COMPÉTENCE & OUTILS

Backend



Machine Learning Framework



Autres

HTML CSS MySQL LaTeX
HDFS Hadoop Git Unit Test
Agile Gradle Lua Neo4j
ElasticSearch

AUTRES PROJETS

🔧 Poem bot

Construire en Java un poem bot qui peut générer la deuxième partie suivante d'un couplet étant donné la première.

🔧 Hackthon CafData 2015

Stagière en Natural Language Processing

LLF

2015

- Collaborer avec des chercheurs pour le Duel Project sur la classification des rôles de parole pour les dialogues.
- Annoter des coprus pour l'analyse de sentiment avec Brat.

Environments:

Java-CorenlpPerlPythonNumpyBrat

EDUCATION

Doctorat en NLP

Université de Nantes

janvier, 2017 - janvier, 2020

Titre de la thèse: Unsupervised cross-lingual representation modeling for variable length phrases.

- Alignement non supervisé des multi-mots bilingues.
- Modélisation de séquences monolingues avec RNN, CNN, LSTM et les architectures modernes basées sur Transformer.
- Bilingual word embedding.
- Augmentation/sélection des données pour les scenarios avec peu de ressource.

Results:

- Améliorer les résultats d'état-de-l'art sur la synonymie des multi-mots de près de 33% sur les corpus de domaine spécialisé à faibles ressources.
- Obtenir les results d'état-de-l'art sur bilingual word mapping.
- Proposer un nouveau réseau neuronal basé sur un graphe sans arbre pour encoder de courtes séquences et des mots simples. Il a surpassé les résultats d'état-de-l'art sur l'alignement non supervisé des multi-mots bilingues par une moyenne de 8,8 points en MAP tout en conservant des résultats comparables pour les mots simples.

Master en NLP

Université Paris Diderot

2014-2016

Cours:

Machine learning; Statistics; Algorithm; 1st order logic; Text mining.

Licence en mathematiques appliquées

Université Paris Diderot

2012-2014

Cours:

Linear algebra; Mathmatical analysis; Java programming; C programming; Probability theory; HTML/CSS/PHP; MySQL

PUBLICATIONS

- Alignement de termes de longueur variable en corpus comparables spécialisés

Jingshu Liu, Emmanuel Morin, Sebastián Saldarriaga

TALN2018
- Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms

Jingshu Liu, Emmanuel Morin, Sebastián Saldarriaga

Coling2018
- Continuous phrase representation learning with wrapped context prediction

In preparation
- A unified and unsupervised framework for bilingual phrase alignment on specialized comparable corpora

Jingshu Liu, Emmanuel Morin, Sebastián Saldarriaga, Joseph Lark

Arxiv
- From unified phrase representation to bilingual phrase alignment in an unsupervised manner

Construire en 48h un système de prédiction du temps d'attente en Python basé sur les données fournies par la Caf dans un concours de hackthon.

Jeu Gounki
Implementer un jeu Gounki en C.

Jeu d'évolution
Implementer un jeu d'évolution en Java avec un UI minimal.

Site web pour les recettes
Créer un site de recettes avec Mysql et PHP hébergé sur le réseau campus de l'Université Paris Dederot. Un étudiant peut s'inscrire pour trouver d'autres personnes qui peuvent lui enseigner les recettes qu'il veut apprendre.

EXPERIENCE VOLONTAIRE

- Coouche customisée pour Deeplearning4j**
Implementerune couche customisée pour Deeplearning4j (avant la version alpha) et la pull request a été fusionnée dans le projet principal.
- Machine Learning Meetup**
Présenter nos travaux à Nantes machine learning meetup 2019.
- Liaison**
Responsable de la communication entre l'équipe du Groupe Edmond de Rothschild et la ville hôte de Extream Sailing Series 2011 à Qingdao.
- Interprète**
Interprète pour Tianhui (SARL) à Salon chinois de l'importation et de l'exportation à Guangzhou, 2010.

LANGUE

Chinois (maternel)
Anglais (Professionnel)
Français (Professionnel)

INTÉRÊTS

Badminton, Basketball, Running

