

ZIJIAN WANG

Master NLP & Master Informatique & Ingénieur

EDUCATION

Master Linguistique Informatique, Université Paris Diderot

2020-2021

- Réseaux Neurones, Base des données, Analyses sémantiques

Master Langue et Informatique, Sorbonne Université

2018-2020

- Programmation Java & Python3, NLP, Machine Learning, Deep Learning

EXPERIENCES

Laboratoire STIH, Sorbonne Université / Stage NLP Entités Nommées

- Détecter des Entités Nommées par des outils NLP dans les corpus multilingues.
- Créer des algorithmes pour résoudre des problèmes de frontière.
- Réentraîner des modèles de langues, évaluer des résultats par Matrices de confusion.
- Trouver l'intersection, la complémentarité et dessiner Diagramme de Venn.
- Github <https://github.com/WANGZijian1994/Detect-Named-Entity->.
- Publication : <https://projet.liris.cnrs.fr/aphns-magis/AtelierAssisesMAGIS2020.html>

Shanghai Data Exchange Center / Data Scientist et Java Développeur

07/2019 - 08/2019

- Participation au développement du moniteur système en temps réel
- Encryptage et Decryptage des données par des algorithmes comme MD5
- Test Logiciel automatique avec selenium
- Management des données sur la publicité avec Java Spring
- Github: <https://github.com/WANGZijian1994/fuxing-tools>

PROJETS IMPORTANTS

Comparaison Sémantique et Phonétique (Thèse pour Master 1)

- Création des algorithmes (basés sur Distance de Levenshtein) pour trouver des paires minimales (phonologie).
- Découverte du WordNet Français pour trouver hyponymie et hyperonymie
- Création des services dans le cadre du Spring pour chercher des paires minimales et leurs informations sémantiques.
- Youtube: <https://www.youtube.com/watch?v=fQY9d4JoKxE>

Solr Research

- Nettoyage des données de Ligue 1 pour 5 ans avec Pandas
- Annotation et indexation du corpus par Solr Interface
- Création des services pour chercher par Spring, en utilisant expressions régulières,

Analyse des sentiments des commentaires d'Amazon

- Scrapy des commentaires par Crawler API
- Nettoyage des données (supprimer stopwords, tokenisation, lemmatisation) avec nltk et pandas
- Vectorisation (TFIDF) et Word2Vec (Skip Gram) pour créer des modèles
- Classification des sentiments avec modèles de Machine Learning (86.7% de précision).
- 10 hot topics dans les commentaires négatifs par LDA.

Reconnaissance de la parole

- Collection des données orales par Praat et formatage automatique des données
- Entraînement d'un modèle HMM pour identifier des mots isolés

Algorithmes et Structure des données

- Participation à un projet OpenSource des exercices LeetCode
- Participation au défi mensuel sur LeetCode
- Github: <https://github.com/AlgoStudyGroup/Leetcode>

CERTIFICATIONS

- Coursera SQL pour Data Science
- Coursera C++
- GRE scores: Verbal 152/170 Quantity 170/170 Writing 4/6

Paris, France
+33 07 69 22 16 16
wangzijian1994@hotmail.com
<https://www.linkedin.com/in/wang-zijian-260442165/>
<https://github.com/WANGZijian1994>

Techniques

- Python
- Pandas
- NumPy
- Nltk
- Spacy
- Data Visualization
- Data Cleaning
- Tensorflow
- PyTorch
- selenium
- Bash
- Git
- MySQL
- MongoDB
- neo4j cypher
- Solr
- Java
- Maven and Spring
- StanfordNLP
- C++
- JavaScript
- HTML CSS

Langues

- Français DALF C1 obtenu
- Anglais ILETS 6.5
- Chinois Langue maternelle
- Allemand, Espagnole, et Japonais -Notion