DAVID ALFONSO-HERMELO

david.alfonso.hermelo@gmail.com | (+1) 438 399 05 66

david-alfonso-hermelo-6646a1b1 dahrs david alfonso-hermelo dahrs

TL;DR

Data Scientist Senior chez Huawei | 8+ années d'expérience | MSc. TAL et Linguistique | Spécialiste IA, DL, Python, Mégadonnées.

COMPÉTENCES

DOMAINES D'EXPERTISE

• Grands Modèles de Langue (LLM) • Science des données • Analyse de données • Traitement Automatique du Langage (TAL) • Modèles d'Apprentissage Profond • Gestion des données • Compréhension du Langage Naturel • Linguistique

LANGAGES DE PROGRAMMATION

• Python (9 ans) • Bash (2 ans) • Javascript (2 ans) • Java (1 an) • Perl (2 ans) • C# (4 mois) • R (6 mois)

LANGAGES DE BALISAGE

• HTML • LaTeX • XML • XLM • CSS • Markdown

LANGAGES DE REQUÊTES

• SQL • SPARQL • XSLT • OWL

LANGUES NATURELLES

• Espagnol (langue maternelle) • Français (courant, DALF C2) • Anglais (courant, Cambridge C1) • Allemand (débutant, A2)

OUTILS PERTINENTS

• API LLM (2ans) • Pytorch (4ans) • HuggingFace (6ans) • Selenium (6ans) • Pandas (7ans) • AWS BedRock(1an) • Azure AI (1an)

EXPÉRIENCE PERTINENTE

HUAWEI'S NOAH'S ARK LAB | Chercheur en TAL, Data Scientist

Décembre 2019 - Aujourd'hui | Montréal, Québec, Canada

- Supervisé 6 projets d'annotation, gérant des équipes de 2 à 32 annotateurs.
- Co-auteur et publié 9 articles, dont 5 dans des conférences de catégorie A*, avec 2 autres en cours de soumission.
- Évalué un LLM spécialisé via des tâches de QA basées sur des citations, intégrant la récupération web et l'extraction de graphes de connaissances.
- Encadré l'équipe de science des données pour le lancement réussi de 5 produits clients.
- Conçu des métriques et des protocoles d'évaluation pour 4 projets.
- Mené des recherches sur des tâches de TAL : LLM, récupération d'informations, compréhension du langage naturel, traduction automatique, augmentation des données, reconnaissance des entités nommées.
- Formé les nouveaux membres sur la gestion des données, les méthodologies Agile et les bonnes pratiques de codage (PEP-8).

UNIVERSITÉ DE MONTRÉAL | Chercheur

Janvier 2017 - Novembre 2020 | Montréal, Québec, Canada

- Développé des modèles de traduction automatique et d'augmentation des données, conduisant à 2 preuves de concept impactantes.
- Co-auteur et publié 3 articles sur la traduction automatique, les taxonomies et les graphes de connaissances.
- Dirigé la recherche et le développement interne d'une ontologie des compétences pour un client.
- Développé un outil d'extraction de données pour un client, intégré à leur pipeline existant.

ÉDUCATION PERTINENTE

MASTER - TRAITEMENT AUTOMATIQUE DU LANGAGE | Université Sorbonne Nouvelle
2015 - 2017 | Paris, France

MASTER - SCIENCES DU LANGAGE | Université Grenoble-Alpes
2010 - 2012 | Grenoble, France

MASTER - LINGUISTIQUE APPLIQUÉE | Université de La Havane
2010 - 2012 | La Havane, Cuba

PUBLICATIONS DE RECHERCHE

2024
- [1] EWEK-QA : Enhanced Web and Efficient Knowledge Graph Retrieval for Citation-based Question Answering Systems, ACL 2024.
- [2] CHARP: Conversation History AwaReness Probing for Knowledge-grounded Dialogue Systems, ACL 2024.
- [3] EUROPA: A Legal Multilingual Keyphrase Generation Dataset, ACL 2024.
- [4] Efficient Citer: Tuning Large Language Models for Enhanced Answer Quality and Verification, NAACL 2024.
- [5] NoMIRACL: Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation, arXiv preprint.

2023
- [6] MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages, TACL 2023.
- [7] CIRAL at FIRE 2023: Cross-Lingual Information Retrieval for African Languages, FIRE'23.
- [8] Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval, arXiv preprint.
- [9] Evaluating Embedding APIs for Information Retrieval, arXiv preprint.
- [10] The state of OAI-PMH repositories in Canadian Universities, DCMI 2023.

2022
- [11] Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages, arXiv preprint.
- [12] Refining an Almost Clean Translation Memory Helps Machine Translation, AMTA 2022.

2021
- [13] NATURE: Natural Auxiliary Text Utterances for Realistic Spoken Language Evaluation, NeurIPS 2021 Dataset Track.

2020
- [14] Human or Neural Translation?, ICCL 2020.

2019
- [15] Automatically learning a human-resource ontology from professional social-network data, Canadian AI 2019.

PROJETS PERSONNELS

- Développement d'un jeu en GDScript avec Godot
- Développement d'un jeu en C# avec Unity
- Adaptation numérique du jeu de plateau Samurai & Katana pour Tabletop Simulator
- Création d'un résumé quotidien de l'actualité en utilisant des LLMs
- Conception d'un outil de scraping pour l'immobilier
- Développement d'un outil de reconnaissance des chants d'oiseaux
- Implémentation d'un bot Instagram
- Contribution à une cartographie sémantique open source (IEML)
- Apprentissage des outils de conception graphique