Wilson (Wei-Sheng) Wang

Machien Learning Engineer

josixwang@gmail.com
linkedin.com/in/josixwang

josix.tw
github.com/josix

Skills

Programming Language	Python
Data Engineering	Airbyte, Kedro
Backend Development	FastAPI, Django, Redis, ElasticSearch
Developer Experience	invoke, pylint, isort, black, mypy, pre-commit, Commitizen, pytest, loguru, MkDocs
DevOps and Other Tools	Docker, Git, GitHub Actions, GitLab CI, AWS

Technical Experience

Data Engineer / Tresl, Inc.   Palo Alto, CA   Oct 2022 — Nov 2022

airbyte

- Decoupled the customer data extraction and loading workflow from Airflow by using Airbyte. Reduce the complexity of the DAGs and the CPU usage of Airflow by around 20%.
- Built the ETL pipelines that extract data from various sources, load them into the GCS and send them to BigQuery for further analysis.
- Migrated the legacy version of Airbyte to the up-to-date one with 140+ squashed merged commits including 500K+ additions and 280K+ deletions.

ML Engineer / Miso Technologies, Inc.   San Francisco, CA   May 2021 — May 2022

miso-webhook

- Responsible for building the ETL pipelines for ingesting data from Shopify API and sending them into internal data API in around 80 requests per minute from Shopify Webhook API.
- Reduce response time by 80% by using multi-layer caching consisting of memory cache and Redis cache to avoid taking too much time on fetching authentication data from the internal API.

string-index

- Responsible for building a key-value store for mapping productID/userID into integer and ndarray at a performance around 4 ms per 11 million query words, which is the core component used in search and recommendation models.
- Reduce the memory usage by 85% with 11 million tokens input by switching the hash map into DAWG to record the tokens. Persist the tokens in the disk and achieve the purpose of sharing memory through mmap.
- Decoupled the string-index component from the code base to make it easier to maintain and extend. Build a standalone package with complete usage documentation, 100% unit test coverage, and a unified coding style conforming to the black, pylint rules.

python-package-template

- Shorten development environment setup time by 60% by using Cookiecutter. Supporting virtual environment management tools like pipenv/poetry, pre-commit hooks, linting, testing, formatting commands, and GitLab CI/CD, Dockerfile configurations.

Research Assistant / National Chengchi University, CLIP Lab   Taipei, Taiwan   Oct. 2017 — May. 2021

KKTIX Ticket Recommender

- Developed a graph-based hybrid recommendation algorithm that increases MAP@5 by 2.5 times in offline evaluation for solving the complete cold start problem in KKTIX, the online ticketing platform under KKCompany Group.

MOST Review Candidate Recommender

- Developed a multi-path recall recommendation system that re-ranking candidate items from item-to-item [1] and user-to-item [2] strategies for recommending 10K+ PDF documents to 2000+ users.

Server Administration / National Chengchi University, CLIP Lab   Taipei, Taiwan   Aug. 2019 – May. 2021

- Responsible for managing around 50+ accounts across 5 servers through NIS. Setting up Prometheus and Grafana for monitoring the CPU, memory, and disk usage of servers with the alerts sending warning messages to Slack channels.
- Provide external partners with accounts having a limited sandbox environment through ChrootDirectory for uploading data without using external services like Google Drive for saving 3-5 days of uploading/downloading time and communication time.
- Prevent daily 10k+ brute-force attacks on one server by setting no password login and setting up fail2ban to block the attackers.

Cloud Support Engineer Intern / Amazon Web Services, Inc.   Taipei, Taiwan   Feb. 2020 – Feb. 2021

Ticket Category Examinator

- Responsible for building a browser extension used to check the wrong category of tickets and provide the possible options for engineers to choose from. The extension is used by 100+ engineers to reduce the time of categorizing tickets by 14%. Built by using GreaseMonkey and backed by AWS S3 for storing static metadata data.

---

[1] i2i models: ElasticSearch, ElasticSearch LTR, and word2vec
[2] u2i models: CF, graph based CF like LINE and DeepWalk models

EDUCATION

M.S. in Computer Science, *National Chengchi University*, *Taipei, Taiwan*   Feb.2020 — Nov. 2022

- Published *Improving Complete Cold-Start Recommendation via Content-Based Preference Graph Convolution Networks*, which is a **graph-based hybrid recommendation model used to resolve the complete cold-start recommendation issue** that occurred in real-world recommendation systems like ticket-selling platforms and news platforms.

B.S. in Computer Science, *National Chengchi University*, *Taipei, Taiwan*   Sep. 2014 — Feb. 2020

ACTIVITIES

PyCon Taiwan   **2019 — Present**
Web Squad Lead   *2022 — Present*

- Coordinated around 8+ team members to play different roles like UI/UX designers, developers, and reviewers. Make sure the internal communication is smooth by building meeting minutes, onboarding docs, and feature docs to ensure the conversation is public and easy to access.
- Monitor work in the kanban board [3]. Make sure to deliver the features requested by the Program Team, Sponsor Team, Registering Team, etc. Distribute the tasks to the team members, and follow up on the progress.

Infra Squad Lead   *2021 — Present*

- Coordinated around 7 members to fit into the infra squad and help them understand and contribute to the projects by mentoring and 1-1 meetings.
- Maintaining 4 projects [4] for supporting the Program Team, Web Team, and Committee to build an automatic workflow through CLI tools.

Developer   *2019 — Present*

- Contributed to the pycon.tw and pycontw-frontend, including implementing the homepage, about, and sponsor page with the card, modal components, and the flexbox, grid layout.
- Responsible for reviewing the opened PRs. Make sure that code uses the unified naming convention, and shared components, and check the functionalities and appearance work correctly. Giving the feedback following the conventional comments guideline.

Open Source Contributions
@airbytehq

- Implemented the new source connector that ingests 5 different resources from ClickUp API through the low-code CDK with the passed acceptance test and complete documentation. (#17770)
- Reported one extensibility issue that occurred when running acceptance test in python 3.10+ (#17855)

@python

- Co-organized the python documentation localization meetup regularly and improved the amount of translated documents by around 20%

Other Projects

- Improved test coverage and added test cases to Commitizen(#257) and exercism/python(#1053)
- Added the feature of showing all ill-formatted in batch checking mode of Commitizen(#244)
- Translated 100+ english messages into zh-tw in PyPI

---

[3]PyConTW'23 Dev-Web Kanban
[4]session-video-publisher, hackmd-archive-worker, talk-preview-imag-builder, and pycon-archive-past-website