

WORK HISTORY & EXPERIENCE

RESEARCH ENGINEER

Restless Bandit, Inc.

Sep. 2014 - Jul. 2019

As the first member of the data science team after co-founders, I worked on the initial architecture of the data infrastructure built on Python and AWS to support the job-candidate match scoring engine, the main business offering before the major pivot. I designed, programmed, and maintained several microservices, exposing the REST interface (flask) for interprocess communication and user interaction:

- Data ingestion service integrating a dozen external job application tracking systems APIs, e.g., Taleo, iCIMS, Bullhorn, and JobVite, ingesting canonical job and candidate entities (PostgreSQL, RDS) as well as batch input for bulk processing (S3); implemented the replay-via-cache layer to save API calls
- Job scheduling/monitoring service for chained mapreduce/Spark workflows (mrjob, pyspark, boto3, moto, EMR, S3)
- Scraper to gather > 20 million resumes for machine learning model training (scrapy)
- Autoscaler/load balancer for a cluster of Daxtra resume parser EC2 instances (SOAP)

The services above were heavily threaded (gevent) and IPC were often done by messaging via Redis, RabbitMQ, SQS, and Lambda. Linux DevOps, testing, as well as miscellaneous quality-control contributions include:

- CI/CD to AWS (CircleCI, Drone CI, Docker, Ansible) and Docker Compose to help local development and integration tests
- PyPI server for privately maintained Python packages, integrated with GitHub hook
- Full migration of Python from version 2.7 to 3.7
- Active participation in code documentation (reStructuredText, RESTPlus), testing (pytest), and review (GitHub)

Data science-oriented contributions include:

- Job title parser library (hidden Markov model and Viterbi algorithm)

DATA SCIENTIST

Bright Media Corp.

Nov. 2012 - Mar. 2014

- Prepared Hadoop-streaming infrastructure on EMR to improve data processing capacity
- Implemented a MinHash deduper on CouchDB/Hadoop to reduce similar job postings
- Run hierarchical clustering on logs to improve categorization of jobs and candidates
- Analyzed quantitative and textual data using numpy, scipy, matplotlib to support operations, marketing, and internal data standardization and visualization

SOLE PROPRIETOR

Okome Studio

Sep. 2007 - Current

- Freelance consulting in software development and data science
- English-Japanese technical translation

FORMAL EDUCATION

Ph.D. in ASTROPHYSICS

University of California, Santa Barbara

2007