Josh Malina

Charlottesville, VA / 305 333 3255 / joshuamalina@gmail.com

Data scientist, full stack software engineer and NLP specialist

WORK

Data Scientist / Software Engineer

CCRI - Charlottesville, VA - since Jan 2016 (1 year 11 months)

- Architected and implemented Scala NLP pipeline from raw text -> word2vec -> document embeddings -> { metadata classifiers, resolved entities, linked entities to Wikipedia}, package in Docker
- Evaluated semantic search algorithms: Tf-Idf Weighted Word Embeddings vs Tf-Idf vs BM25
- Used XGBoost to model virality of posts / comments on Reddit.com
- Implemented Word Mover Distance algorithm to make semantic document recommendations explainable to the user
- Evaluated anomaly detection and unsupervised learning models for finding needles in hay

Sole Developer, Owner

LaoshiList.com (extinct) - Beijing - Aug 2013 to Nov 2015 (2 years 2 months)

- Built Chinese / English web application to serve employment needs of Beijing's expat teacher community
- Taught myself PHP / MySQL / Jquery (version 1), then later: Python / Angular / Mongo (version 2)
- Became financially profitable; trained and managed teachers; developed business with picky Chinese parents / schools; Provided hundreds of foreign teachers with jobs

Front End Application Developer

CRM Factory - Beijing - November 2014 to May 2015 (7 months)

- Built consumer facing Angular applications for SaaS platform

EDUCATION

Master's in Software Engineering

Harvard Extension - Cambridge, MA | Sept 2013 to Ongoing: Machine Learning, Statistics, CS Fundamentals, Web Application Development

Bachelor's in Philosophy

Washington University in St Louis | 2006 - 2010

PERSONAL

chirbah.com

Blog aimed at solving the fruit information asymmetry problem / having fun

aqcast.com (extinct)

Random forest / flask api / angular web app to train and predict 10 days of air pollution in Beijing

RESEARCH

Statistical Analysis of Pollution in China

Analyzed temporal trends in particulate matter concentrations in five Chinese cities

Comparison of Information Retrieval Algorithms

Measured search + recall performance of term frequency weighted search versus against embedding based methods

Scala, SQL, Python, Git, Solr, Docker, Random / Boosted Forests, Neural Networks, K means, OrientDB, Javascript, Angular, AWS