K CHALKLEY

Applied Scientist, Machine Learning Engineer, Natural Language Processing Specialist

k@kchalk.com | 585-209-3390
kchalk @ Github and LinkedIn | Remote from Portland, OR

SEEKING…

- Design and experimentation oriented role, applying machine learning techniques to language
- Collaborative, open, and engaged team culture

FAMILIAR WITH…

- Python
- Generative Language Models, GPT 4+, LLMs, Classification, Entity Recognition, IR, tf-idf, SVM, Deep Neural Networks, linear regression…
- Spacy, Pytorch, HuggingFace, Transformers, nltk
- Elasticsearch, Lucene, vector search
- Pyspark, Dask, Kubernetes, Microsoft and Amazon cloud platforms
- Science, research, linguistics

SENIOR APPLIED SCIENTIST, LEAD MACHINE LEARNING ENGINEER JAN 2023 – OCT 2024

RELATIVITY ODC www.relativity.com

Most used technologies: Python, GPT 4(+), Spacy, Scikit-learn; Pyspark, K8s, Azure Cloud, Open telemetry, New Relic, SVM, Classification, Ranking

REVIEW CENTER – MONITORING, TECH DEBT, SCALING

- Designed, implemented, & deployed the company's first model quality monitoring for our active-learning-based classifier. Collaborated legal and data governance committees to navigate requirements concerning client data.
- Resolution of tech debt resulted in reduction from multiple incidents per week to 99% stability, while also providing 10x scalability

PILOT EXPERIMENTATION – CLIENT ENGAGEMENT, OBJECTIVE EVALUATION

- Engaged clients to evaluate and tune and advanced access stage modeling project, involving generative models, natural language prompting, and difficult to optimize ranking targets.
- Revised evaluation protocols to provide more comprehensive and generalizable metrics.

RESPONSIBLE AI INVENTORY – RISK, DOCUMENTATION

- Determined model documentation standards in accordance with NIST AI Modeling Risk Management Framework

SENIOR MACHINE LEARNING ENGINEER

Sept 2021 – Dec 2022

SparkCognition Government Systems

www.sparkgov.ai

Most used technologies: Python, Pandas, Spacy, Scikit-learn, Elastic Search, Azure DevOps, Dask, Document Similarity, Classification

Form Embedding

- Historical records containing heterogeneous data (long & short text, categorical fields, document codes, entities, etc.) embedded for ‘similar document’ ranking.
- Customized for a variety of datasets including: satellite anomaly reports (25k-50k records), autoclave maintenance requests (3k), aircraft maintenance requests (60k)
- Assist engineering in deploying vectorization apps and nearest neighbor search in databases

Data Interpolation

- Predict missing categorical data from text fields for use in downstream efforts
- Design modeling solutions for small (20k) training set

Product support

- Owned and contributed to data science libraries as part of internal and parent company products: SparkCognition Manufacturing Suite, DeepNLP^{TM}, Digital Maintenance Advisor

ENGINEER II, MACHINE LEARNING

Mar 2020 – Aug 2021

Comcast NBC Universal

Most used technologies: Python, PyTorch, AWS EC2 and S3, Gensim, Jenkins, Github, LSTM

Project: Logical Form Prediction

- Upgraded model to support variable length utterances
- Improved prediction of misspelled queries from 23% to 42%, while maintaining 80% accuracy on the rest of the domain
- Curated datasets for model evaluation.

Project: Xfinity Assistant Ontology Development

- Redesigned ontology of intent classifications and extracted entity types to rebalance training data and improve scalability.
- Completion of each project milestone resulted in immediate and sustained improvement on key metrics (containment rate, classification accuracy).

Publication: Voice QAC @ EMNLP

Voice Query Auto Completion. Tang et al., EMNLP 2021. https://aclanthology.org/2021.emnlp-main.68

MS COMPUTER SCIENCE 2017-2019

Oregon Health and Science University

Most used technologies: Python, R, PySpark, GGplot, Seaborn, Bokeh

Project: Absolutist Language Use Across Subreddits

- Text representation by dictionary frequency (also LDA)
- T-SNE dimensionality reduction
- Bokeh for data visualization

Project: Aphasia Classification

- Bi-directional LSTM representing actual response and target response
- ARPAbet and CMU Dict phonemic transcription
- 6 class categorizer of error types (phonetic, semantic, mixed, etc.)

BA LINGUISTICS 2010-2015

Reed College

Most relevant skills: Pattern analysis, theory crafting, research papers, data analytics, wrangling people and ideas in meetings, LaTeX

- Thesis -- Applied Asymmetries: Syntax of applicative constructions in Tukang Besi
- Linguistic focus in Syntax, Morphology, Morphosyntactic typology
- Really loved being a nerd at nerd school

VALUES

- Honesty and transparency
- Seeking input from diverse perspectives
- Thinking! Theory, research, creative problem solving, etc.
- Being my whole, best self and doing my best work