Jacob Bayer

LinkedIn • GitHub • PyPi • Homepage
(914) 268-7131
jacobbayer@proton.me
Brooklyn, NY

Education
MS – Statistics | GPA: 3.71/4 | January 2022
City University of New York (CUNY), Baruch College

BS – Economics | GPA: 3.87/4 (Summa cum Laude) | May 2020
State University of New York (SUNY) at New Paltz

Experience

Data Scientist at *\/prompt.* • New York, NY — **Apr ‘23 – Present**

- Developing a machine learning pipeline to identify topics in natural language data using Top2Vec, a technique that utilizes Doc2Vec embeddings and HDBSCAN for clustering.
- Wrote a proprietary extension for Top2Vec to assist two data scientists in ad-hoc parameter optimization for HDBSCAN and UMAP as part of our topic modeling workflow.
- Designed a prompt engineering workflow for optimizing results from OpenAI’s GPT-4 model to successfully create a hierarchy of topics from hundreds of topic clusters.
- Created engaging network graph visualizations of the clustered topic results and presented these data to clients as part of new business pitches.

Creator of *Sunbelt* — **Dec ‘22 – Apr ‘23**

- Created Sunbelt, a social media conversation crawler, database, and API that connects machine learning and analytics pipelines with conversational natural language data from the internet.
- Worked independently on the development of Sunbelt for two months to support two users.
- Sunbelt is a Flask application using SQLAlchemy and a PostgreSQL database. It processes incoming data using a Redis queue and serves data to end-users via GraphQL and REST API endpoints.

Data Engineering Contractor at *Govini* • Remote — **Jun ‘22 – Dec ‘22**

- Built 5 major ETL pipelines using AWS Glue and Apache Spark to extract tens of millions of rows from Govini’s data warehouse to the application database.
- Designed target database schemas to optimize storage space and API read efficiency.
- Collaborated with the front-end and data science teams to ensure output data meets product requirements.
- Orchestrated migrations and ETL jobs to synchronize with ingestion of new data into the data warehouse.

Data Scientist at Phosphorus • New York, NY — **Aug ‘20 – May ‘22**

- Led the effort to extract insights from Phosphorus’s operational data, resulting in a reduction in laboratory sample turnaround time by 75% while sample volume doubled.
- Built a data reporting system which included a Flask app linked to an analytics database that ingested data in real time from the main application database.
- Started an initiative to change the data architecture to improve analytics capabilities, and worked closely with the engineering team to implement this migration.
- Built a logistic regression model to flag certain samples in the laboratory and alert a slack channel that these samples may require additional attention.

Software Languages and Skills

Python (3 years), SQL (3 years), Pandas (3 years), Flask (2 years), PostgreSQL (3 years), Redis, SQLAlchemy, Plotly Dash, PySpark, NumPy, Ruby-on-Rails, Git, Jira, Agile, Linux, Top2Vec, HDBSCAN, LLMs, GraphQL

Coursework

Machine Learning for Data Mining, Data Mining for Business Analytics, Foundations of Statistical Inference, Applied Probability, Applied Natural Language Processing, Multivariate Statistical Methods

Hobbies

Skiing, jiu-jitsu, wrestling