Bill Watson

Data Scientist · Software Engineer

(+1) 516-543-1131 | nextbillyonair@gmail.com | nextbillyonair.github.io | nextbillyonair | nextbillyonair

Education

The Johns Hopkins University — *Baltimore, MD*

M.S.E. in Computer Science — *Aug. 2018 - May 2019*

B.S. in Computer Science, Minor in Applied Mathematics & Statistics — *Aug. 2014 - May 2018*

- Undergraduate GPA: 3.71; Graduate GPA not calculated
- Departmental Honors, Computer Science; University General Honors; Dean's List (2015-2018)

Skills & Abilities

Programming - Python, LaTeX, SQL, JavaScript, HTML/CSS

Frameworks, Libraries, & Tools - PyTorch, React, Flask, OpenCV, TesseractOCR, ElasticSearch, Azure DevOps, Git, UNIX/Linux

Hobbies & Recreation - Baking, Biking

Work Experience

S&P Global — *New York City, NY*

Data Scientist — *Jun. 2019 - Present*

- **Market Insights** - Commercial product that derives investor sentiment & risk factors for companies
  - Created an ingestion pipeline for earning call transcripts that uses NLP techniques to extract salient content for sentiment & location tagging
  - Developed several SQL queries to gather important metadata, including location & analyst information, for each earnings call
  - Contributed enhancements to the project's sentiment model, based on FinBERT, & indepentently conducted ablation studies to increase precision
- **Supply Chain Networks** - Proof of concept commercial product to derive company dependence and concentration risk from global trade data
  - Created SQL queries to merge shipping data with CapitalIQ industry data, providing a holistic view into product-industry distributions and relations
  - Constructed matrix-models to impute probable product inputs for a given company, & derive hidden disruptions in a company's direct supply chain.
- **LIBOR Exposure & Risk Mitigation** - Application to highlight risk in hundreds of trillions of dollars of global assets in preparation for the LIBOR phaseout
  - Designed a LIBOR extraction model based on the cosine similarity of analyst-reviewed feedback, achieving over 90% in similarity metrics
  - Led the development of a neural network to automatically tag the final LIBOR fallback category, achieving over 98% accuracy on validation data
- **Table Extraction in Image Documents** - Proof of concept pipeline to identify, extract, and organize tabular content from unstructured financial documents
  - Trained an Inception-ResNet & U-Net segmentation model to extract a table from an image, and enhance OCR results with OpenCV and TesseractOCR
  - Led the creation of several sequence models based on spatial & lexical metadata to provide approximate column boundaries that allows the pipeline to organize content into a structured format (CSV, LaTeX)
  - Accepted as a full paper at the 1st ACM International Conference on AI in Finance (ICAIF'20)
- **Criteria Validation Search & Citation Recommendation** - Platform for multi-phrasal search and citation reccomendation on ratings criteria
  - Managed the development of a full-stack application with an Angular UI, Flask RESTful API, & ElasticSearch engine to provide meanigful ranked results & connections between criteria, improving the efficiency of validating framework changes
  - Trained a transformer-based graph neural network to recommend missing citations, helping reduce inconsistencies & liabilities in the current framework
  - Accepted as an extended abstract at the 1st ACM International Conference on AI in Finance (ICAIF'20)
- **Ratings Data Science Reading Group** - Led several reading groups instructing coworkers on advanced concepts in machine learning & data science
  - Introduction to Differentiable Probabilistic Models (July 25th, 2019) - Reformulates deep learning frameworks as differentiable probabilistic models
  - Introduction to Random Numbers, Sampling, and MCMC Methods (August 22nd, 2019) - When deterministic approaches fail to yield approximations
  - Seq2Seq in Action: Column Segmentation (January 23rd & 30th, 2020) - Primer to encoder-decoder models with a use-case for aligning table columns

Publications

Leadership In Turbulent Times: Women CEOs During COVID-19 — *S&P Global*

Cross-divisional work to highlight women's communication style during the COVID-19 pandemic from earning call transcripts with NLP techniques

Financial Table Extraction in Image Documents — *ICAIF 2020*

Presents an end-to-end pipeline for identifying, extracting and transcribing tabular content from with segmentation, OCR, and sequence modeling

Directed Criteria Citation Recommendation & Ranking Through Link Prediction — *ICAIF 2020: Extended Abstract*

Details how a transformer graph network can be used to recommend & rank missing citations

Modeling Color Terminology Across Thousands of Languages — *EMNLP-IJCNLP 2019*

Outlines several measures on massively cross-linguistic data to operationalize and critique the Berlin & Kay color term hypotheses for basic color terms