AbderRahman Sobh

Data Scientist

169 Charles St
Sunnyvale, Ca 94086
(217) 979-4127
abbysobh@gmail.com

EXPERIENCE

Data Scientist – AJ+ / Al Jazeera International

Dec-2016 – Present

Data Science Tasks:

Developed Data Science tools for use from end-to-end:
Product conceptualization, database/schema design, pipeline creation, Dockerized deployment, insight via easily consumable deliverables (dashboards, Slackbot, etc.), data-driven weight corrections, refinement of deliverables based on user feedback.

Developed a tool for live scoring of Social Media Engagement (i.e. Likes, Shares, Comments) for media produced by the company on Facebook, YouTube, Instagram, Twitter.

Published an article under the AJ+ brand, documenting the scoring tool's creation process for the benefit of the public community.
<https://medium.com/aj-platforms/re-thinking-engagement-at-aj-69a35e0a38c>

Developed tools for text-feature extraction including analysis of grammar used, content topic modeling, word vector embedding, sentiment analysis.

Proved the need for, and successfully implemented, an overhaul of legacy data science tools and inferences that were being used by the company.

Developed and presented a comprehensive, high-level Roadmap of potential Data Science projects available for development in the Social Media context.

Met with vendors and reviewed their 3rd party platforms for data solutions.

Data Engineering Tasks:

Worked with Data Engineers on the team to resolve various issues.
Essentially, providing service as an additional Data Engineer when needed.

Planned, reviewed, implemented, and populated database schemas and views.

Built, modified, and repaired ETLs for various data collection pipelines.

Provided re-usable tools to supplement Data Engineering tasks
(i.e. a script that copies data from S3 storage into a Redshift database)

SKILLS

Python, SQL, NoSQL, MySQL, Docker, AWS(EC2, Lambda, S3, EMR, Redshift), Spark, Tableau, R, SAS, Matlab, Mathematica, C, C++, TCL, VBA/Excel, Fortran90, HTML,PHP, JavaScript, CSS, XML, OSX, Linux, Windows

Groups

Data Scientist Engineers Project - Burlingame, Ca

New York Open Statistical Programming - Manhattan, NY

LANGUAGES

English, Arabic, Japanese

Software Engineer - University of Illinois at Urbana-Champaign

Jan-2012 - Dec-2016

Data Science Platform:

Designed data science pipelines using Jupyter Notebooks, Python ML libraries, and Spark to provide single-click analysis tools for educating future data scientists.

Built distributed machine learning tools using Python, PySpark, and Pandas which simplifies the use of NLP and Text feature extraction.

Optimized AWS Elastic MapReduce by leveraging Spark and increasing the distributed performance.

Collaborated with Full Stack developers to integrate data science pipelines using Docker containers to encapsulate various data science software stacks providing end-users with quick usage of data-science tools and visualizations.

Simulation Workflow Development:

Collaborate and integrate software applications within openVZ containers providing rapid deployment and accessibility of custom built GUI within Nanobio portal webapp.

Develop graphical user interfaces using TCL and RAPPTURE (XML framework) for the containerized applications to ease the usage of applications for portal end-users.

Optimize simulation memory-handling and runtime on millions of data points using Python, TCL- specifically using associative arrays/hash mapping.

Connected developed workflows to remote cluster systems and optimized parallel threading on serial coding procedures to accommodate scaling.

Integrated visualized workflow results using the VTK toolkit for robust manipulation on custom displays.

Online Database for Yeast Strain Mutations and Replications:

Developed web application in PHP and MySQL that provides end users with an interface for creating, uploading, curating, and visualizing data for gene expression analysis.

EDUCATION

University of Illinois at Urbana-Champaign

Bachelors of Science in Statistics

August 2007 - May 2012

Dean's Honor List (2008), (2012)

PROJECTS

GDAX Automated Trader — Developer (2017)

With the recent hype in cryptocurrency markets, a lot of progress has been made in terms of making publicly available APIs for sending trades. I took advantage of the GDAX API to make my own automated trader, watching for trends and sending out orders to buy/sell accordingly based on the trend detection algorithm's advice.

This project is necessarily closed source.

Kaggle Data Science — Competitor (2016)

During the Kaggle competition I cleaned data sets, normalized data, and predicted missing values. I leveraged the use of NLP and clustering to investigate text features. In addition, I performed feature selection, created models to predict new data values using Python/Pandas, GraphLab, scikit-learn, XGBoost, and Amazon Web Services for parallel data processing workflows. My efforts received public recognition and a key part of the script is open source (aka a "Kaggle Kernel").

https://github.com/itg-abby/KaggleScripts/

Automated Level Creator — Developer (2015)

Development of level files for rhythm-based video games is something that is typically done by hand. The aim of this open source script is to reduce the amount of human effort spent on analyzing songs for relevant sound features. Specifically, relevant features are mapped according to a given time signature and presented in an output format readable by programs such as Stepmania. Time series analysis is powered by the AUBIO library for Python.

https://github.com/itg-abby/StepGen

Portfolio

https://itg-abby.github.io/portfolio/