Tien Nguyen

Email: tttienguyen@gmail.com | Tel: 925-319-7484
Github | Linkedin |Website

Experience

Data Quality Control, Mosaix.ai, Palo Alto, CA *September 2018 - February 2019*

- Researched and verified the quality of the translation output from the NLP model.
- Documented and wrote reports on the quality of the output of the NLP model.
- Generated more training and testing data for specific use cases, such as interacting with a smart TV.

Quality Control Inspector, Flex Electronics, Milpitas, CA *June 2017 - October 2017*

- Inspected and found defects of board circuits using microscopic devices.
- Ensured the production operations meet the specifications.

Projects

Book Rating Prediction (predicted differences in Goodreads and Amazon books ratings) Github

- Collaborated with another person in the project, created and assigned the tasks for each member.
- Obtained and cleaned large and messy Amazon and Goodreads datasets.
- Implemented multiprocessing to maximize CPU utilization to drastically speed up the cleaning process by 100 times.
- Performed exploratory data analysis to explore the characteristics of the datasets and to provided insights about the datasets.
- Applied machine learning algorithms such as random forest, xgboost, k-nearest-neighbors, neural networks, and weighted ensemble to lower MAE, MSE, and RMSE by approximately 10% from the Zero Rule Algorithm baseline.
- Implemented Natural Language Processing (NLP) using Natural Language Toolkit (NLTK) and Linear Discriminant Analysis (LDA) using Gensim Mallet wrapper to improve the accuracy of the predictions.
- Made a presentation and wrote a report communicating the results to non-technical and technical audiences.

Web Crawler and Search Engine

- Built a crawler crawling UCI web pages.
- Wrote scripts for preprocessing text data: extracting text from HTML, parsing text, filtering out common words and punctuations, tokenizing the words in the text, and inverted indexing the words to web pages.
- Built a search engine: providing top-k relevant results from the inverted indexed words using tf-idf scoring, using tkinter library to build a graphical user interface (GUI).
- Utilized multiprocessing & multithreading to optimize the speed of preprocessing data and of the search engine.

The Battle of Neighborhoods (setting up a new coffee shop in Los Angeles) Github

- Scrapped data from Wikipedia websites and Foursquare API.
- Performed exploratory data analysis to explore the characteristics of the datasets and to choose a suitable algorithm (K-means) for the datasets.
- Provided insights and recommended the good locations based on the results.
- Made a presentation and wrote a report communicating the results to non-technical and technical audiences.

Skills

- **Programming Languages**: Python, R, Dask.
- **Big Data & Machine learning**: Hadoop, MongoDB, SQL, MySQL, PostgreSQL, Python (eg. scikit-learn, NumPy, pandas, matplotlib, seaborn)
- **Data Science & Miscellaneous Technologies**: Data science pipeline (cleansing, wrangling, visualization, modeling, interpretation), Statistics, Jupyter Notebook, APIs, Git, Data Management.

Education

University of California, Irvine - B.S in Data Science Graduated July 2020

Relevant Coursework: Algorithms, Database Systems (MongoDB), Machine Learning and Data-Mining, Information Visualization, Information Retrieval, Applications of Probability in Computer Science, Statistical Methods for Data Analysis, Statistical Computing and Exploratory Data Analysis.

De Anza College - A.S in Computer Science Graduated July 2017

Certifications

IBM Data Science by IBM on Coursera

Data Analyst with SQL Server on Datacamp