Senqi Zhang

585-503-4035 | sz3016@columbia.edu | https://www.linkedin.com/in/senqizhang/

EDUCATION

Columbia University New York, NY
MS in Data Science, GPA 3.83/4.0 Anticipated Dec 2022
Courses: Algorithms for Data Science, Applied / Statistical Machine Learning, Data Visualization, Computer Systems

University of Rochester
BS in Data Science, GPA 3.88/4.0 Rochester, NY
Courses: Data Mining, Time-series Analysis, Database System, Data Science at Scale, Statistical Methods May 2021

ACADEMIC EXPERIENCE

Recommender system for Amazon users New York, NY
Applied Machine Learning Project (Python, Surprise, Machine Learning) Sep 2021 – Dec 2021

- Developed a recommender system using Amazon review data in the category of video game.
- Transformed review and product metadata into a ten-core dataset to reduce sparsity.
- Applied content-based filtering using review texts and product descriptions. Evaluated Bag-of-words and TFIDF models' performance using a set of products that were bought and viewed together in addition to human evaluations.
- Applied collaborative filtering using KNN, SVD, NMF models from the Surprise library. Evaluated performance for each model by calculating the RMSE metric and SVD achieved the lowest RMSE of 0.9467.

Elon Musk on Twitter New York, NY
Data Visualization Project (rtweet, R, GitHub) Sep 2021 - Dec 2021

- Collected Tweets related to Elon Musk for seven days using rtweet.
- Selected and cleaned up seven features from raw tweet objects and converted them into tidy data set.
- Created visualization on volume, geological distribution, and sentiment of tweets and discussed their implications.
- Applied topic modeling and identified two major topics of 'Crypto-currency' and 'Tesla'. Classified tweets into each topic and carried out sentiment analysis. Clearly observed more positive sentiment in the topic of 'Crypto-currency'.
- Created an interactive plot using D3 library to present distributions, contents, and magnitudes of positive and negative tweets. Allowed filtering for specific time periods, keywords, and number of followers.

The COVID-19 Pandemic and Mental Health Concerns on Twitter in the US Rochester, NY
Data Science Capstone Project (Python, NLTK, Face++, Ethnicolr) Feb 2021 - May 2021

- Filtered out mental health tweets from 7TB of COVID-related tweets dataset.
- Applied topic modeling (Latent Dirichlet Allocation) to monitor the trend of discussions surrounding mental health concerns. Extracted ten topics and concluded underlying causes with example tweets.
- Applied demographic analysis (Age/Gender/Race) using machine learning algorithms (Face++/Ethnicolr) on over one hundred thousand users with a human profile picture and name.
- Conducted statistical analysis to compare derived user demographics with the general Twitter user population. Concluded that White and Male have statistically significant more mental health concerns.

WORK EXPERIENCE

Columbia University New York, NY
Course Preparation Assistant for GR5293 Jan 2021 - Current

- Utilized data from R libraries, CDC, Census, and NHGIS to design biweekly assignments.
- Researched new visualization techniques and integrated them into assignments as an addition to lectures.
- Complied resources and guidelines about data visualizations for edav.info along with the progress of the course.

SKILLS

- **Programming language:** Python, R, Java, SQL, HTML, CSS; Basic: Spark, JavaScript
- **Tools & Frameworks:** Scikit-Learn, NLTK, Pandas, Tidyverse, BigQuery, GitHub
- **Product skills:** A/B testing, Data Visualization/Modeling

PUBLICATION

- **Senqi Zhang**, Li Sun, Daiwei Zhang, Pin Li, Yue Liu, Ajay Anand, Zidian Xie, and Dongmei Li, "The COVID-19 Pandemic and Mental Health Concerns on Twitter in the United States", Health Data Science, 2022.