Quoc-Anh Le

anhlq28102002@gmail.com | (+84) 978 207 750 | linkedin.com/in/davidle2810/ | davidle2810.github.io

EDUCATION

HO CHI MINH UNIVERSITY OF SCIENCE - VNUHCM

Bachelor of Science in Computer Science; Nov 2024

Coursework: Core Statistics, R Programming, Machine Learning, Artificial Intelligence, Natural Language processing, Mining on Big Data

THESIS : Automatic Translation from Nom-scripted poems into contemporary English

Building Nom scripts - English parallel corpus

- Collected and cleaned peoms that are both crawled from websites and taken from pdf files.
- Normalized and tokenize the collected data.
- Performed word alignment on the parallel corpus.
- Performed statistical analysing on the final database.

FEATURED PROJECTS (more details at https://davidle2810.github.io/)

Regression Techniques in House Price Prediction

- Performed comprehensive data analysis, data cleaning and feature extraction.
- Applied Stochatic Gradient Descent to build a regression model to predict house prices.

Recommendation System: Spotify Datasets

- Preprocessed a 100,000 entries datasets by handling missing values, duplicate values, unformatted text, unuseful columns.
- Analyzed the dataset and visualized interesting information.
- Applied PCA to reduce the numeric dimensions.
- Performed BFR clustering based on the mahalanobis distance.

Electrical Vehicle Charging Data

- Preprocessed dataset of 7,000 records.
- Conduct statistical analysis.

Frequent Itemsets with Association Rules Mining

- Performed product recommendation on the itemsets dataset filled with 100,000 entries using A-priori algorithm.
- Built the association rules for the item triples.

Lbl2Vec: Document Classification

- Utilized semantic similarities to retrieve documents related to predefined topics.
- Classified documents (including unrelated documents) across multiple topics by using cosine similarity.
- Experimented with the AG's news topic classification dataset containing 30,000 training samples and 1,900 testing samples among 4 topics (world, sports, busisness, science/technology) and achieve the F1 score of 0.81.

Sentence Alignment

- Predicted cross language by calculating the probability using softmax function with the application of BERT architecture and optimized the spans by using Integer Linear Programming.
- Built the model to align sentences in the Korean-English bi-text crawled from CNN and Yahoo.

Extracting Keywords using TF-IDF

- Performed keywords extraction in BBC news documents.
- Extract the top 5 keywords for each article and topic.

Guns in the USA, by year for 1977–1999.

- Described statistically the data and performed hypothesis testing.
- Performed a regression model of a variable according to the quantitative variables and performed hypothesis testing about their distribution.

SKILLS

Programming: Python (NumPy, Pandas , Scikit-learn, TensorFlow, nltk), R, SQL, Hadoop, Spark
Visualization and Statistical Software: Tableau, Python (Matplotlib, Seaborn)
Project Management: Trello, Jira, Google Calendar
Machine Learning: Regressions, Classification (Decision Tree, PLANET), SVM, Unsupervised Learning (Clustering, PCA), Deep Learning