

NGUYEN NGOC THANH SANG

DATA ENGINEER INTERN

+84 385 350 187

nguyensang2032003@gmail.com

NguyenNgocThanhSang

nguyenngocthanhsang

CAREER OBJECTIVE

As a student majoring in Data Science and Big Data student with a basic level of Japanese, I am seeking an internship to apply my knowledge and gain practical experience. I am particularly interested in working on data-related projects, especially those connected to the Japanese market.

EDUCATION

University of Information Technology - VNUHCM

Oct. 2021 - Present

Bachelor of Information Technology (Japan orientation)

GPA: 8.17 / 10.0

Expected Graduation Time: Jun. 2025

PROJECTS

Classifying Vietnamese Social Media Text ([Github](#)).

Apr. 2024 - Jun. 2024

- **Description:** Sentiment analysis of Vietnamese text. The project aims to classify Vietnamese text into sentiment categories using pre-trained machine learning models.
- **Techniques:** CNN, PhoBert, VisoBert, CafeBERT, Word2Vec, tokenization
- **Responsibilities:** Vectorizing pre-processed Vietnamese text data by Word2Vec, applying pre-trained models to train the encoded input, coordinating with another team member to tune and evaluate the performance of the models.
- **Skills gained:** Understanding the process and utilization of various Vietnamese text preprocessing methods, possessing knowledge of pre-trained languages models for Vietnamese.

PDF Malware Classification ([Github](#)).

Sep. 2023 - Nov. 2023

- **Description:** This project involves classifying PDF files as malicious or benign based on extracted features from the CIC-PDF Malware 2022 dataset.
- **Techniques:** AdaBoost, XGBoost, RandomizedSearchCV, scikit-learn.
- **Responsibilities:** Analysing the features of the data set, building the XGBoost algorithm model using the scikit-learn library, performing predictions and evaluating that model with pre-splitted test data.
- **Skills gained:** Analysing datasets for training, applying boosting algorithms to capture the complexity of data, selecting appropriate metrics to evaluate models, data visualization.

SKILLS

Programming languages: Python, SQL, HTML/CSS, C++

Machine Learning / Python libraries: Scikit-learn, Pytorch, Django, Numpy, Pandas, Matplotlib

Database: MySQL, Microsoft SQL Server, PostgreSQL

Developer Tools: Git, Jupyter Notebook, VS Code

Languages: English, Japanese (Equivalent to N4)

ACTIVITIES

Book and Action UIT Club - Vice President

Jun. 2022 - Jun.2023

Japanese: 125/180 ASK JLPT N4 Online Practice Half-Test

Apr. 2023