ĐỖ TẤN SANG

Linkedin: https://www.linkedin.com/in/dotansang/
Github: https://github.com/dotansang

Email : dotansang1@gmail.com
Mobile : +84 97 383 0290

EDUCATION

- **National Economics University** — Hanoi, VN
  B.Sc. in Computer Science — *Aug 2018 - Jun 2023*
  Courses: Calculus, Linear Algebra, Probability Theory, Artificial Intelligence, Data Structures, and Algorithms

EXPERIENCE

- **NLP/AI ENGINEER** — Hanoi, VN
  FTECH CO., LTD — *Nov 2021 - Current*
  - **Vietnamese Spelling Correction**:
    - Employed a Hierarchical Transformer model with both word and character levels to enhance the quality of the spelling correction.
    - Analyzed spelling mistakes cases in a 10 million Facebook comment dataset to build a mistakes dictionary.
    - Implemented an optimized rule-based approach using N-gram + Trie, regex for fast, accurate correction.
    - Deployed using FastAPI, created a playground for easy comparison with other spelling correction APIs (Zalo, Viettel, etc..)

    Skill Stack: *PyTorch, Hugging Face, Gradio,transformers, wandb, matplotlib, FastAPI, scikit-learn, marisa-trie, regex, pandas, and Docker.*
  - **Customer Support Chatbot using ChatGPT**:
    - Develop a conversational chatbot using the ChatGPT API, based on predefined bot scripts and documents provided by the client.
    - Enhance performance by optimizing prompts and generating "fake" document embeddings derived from historical queries, linking to actual documents via HyDE pattern.
    - Integrated the bot with Facebook Messenger, Telegram, and CRM systems.

    Skill Stack: *Langchain, LlamaIndex, ChatGPT, Rasa Framework, Huggingface.*
  - **Build Customer Profile from Chatlog**:
    - Build and optimize a baseline PhoBERT model and test alternative models (Envibert, PhoBERT+CRF, etc.) and different evaluation metrics.
    - Implemented an annotation tool (PyQT) and surveyed labeling tools (Label Studio, Docano, Techolic, etc.).

    Skill Stack:*PyTorch, Huggingface, transformers, wandb, NumPy, pandas, FastAPI, Matplotlib, scikit-learn, gensim, regex, PyQT.*
  - **Build NLP Corpus**:
    - Built and preprocessed a diverse corpus of ˜25 NLP datasets, including sentiment analysis, social comments, news articles, recommendations, chatlogs, movie subtitles, and Wikipedia.
    - Efficiently handled large datasets, including processing 10 million Facebook comments in JSON format, 49GB CSV news dataset.

    Skill Stack: *PySpark, pandas, MinIO, regex, Facebook Graph API, Hugging Face datasets, and Kaggle datasets.*
  - **Vietnamese Paraphrase Generation**:
    - Implemented fine-tuning on T5 and BART models, including prompt-tuning and fine-tuning of the top two layers.
    - Utilized data augmentation methods: translation, back translation, and collection of available data, ensuring high-quality paraphrased sentence pairs by employing various filters.
    - Deployed the paraphrase generation system using FastAPI.

    Skill Stack: *Pytorch, pandas, scikit-learn, Hugging Face, transformers, FastAPI, nltk, sentence-transformer.*
  - **TikTok Reply Recommendation Bot**:
    - Design bot script, built intent-based dataset, and applied data augmentation techniques (rule-based, paraphrasing generation, Chatette).
    - Automated identification of trending songs (NER task), analysis user sentiment, and mapping trending songs with high accuracy using fuzzy search.

    Skill Stack: *Rasa Framework, PyTorch, Huggingface, MongoDB, fuzzysearch, and FastAPI.*
- **DS RESEARCHER** — Hanoi, VN
  DS-LAB NEU — *Aug 2021 - Dec 2021*
  - **Extractive Summarization**:
    - Conduct analysis on 8 English and Vietnamese news articles to develop extractive summarization techniques.
    - Implement extractive multi-document Summarization using K-means, Centroid-based Method, MMR, and Sentence Position

SKILL SUMMARY

- **Programming Languages**: Python, C#, C++, SQL, Pascal.
- **AI/NLP**: PyTorch, Hugging Face, transformers, Langchain, LlamaIndex, Rasa Framework, nltk, sentence-transformer, scikit-learn, regex, gensim.
- **Data**: NumPy, pandas, PySpark, MinIO, wandb,matplotlib, seaborn, BeautifulSoup, Selenium.
- **Other**: FastAPI, PyQT, Docker, ChatGPT, Gradio, PyMongo, Facebook Graph API, Hugging Face datasets, Kaggle datasets.
- **English**: General Aptis ESOL International Certificate: C band (Score: 181/200)

COMPETITION AND AWARD

- VLSP 2022 - Vietnamese Abstractive multi-document summarization
- Zalo AI Challenge 2022 - E2E Question Answering
- 2nd prize for excellent students in Informatics, awarded at the city-level competition in Hanoi, 2017

PUBLICATIONS

- **Vietnamese Paraphrase Generation using Pre-trained Language Model**
  Nguyen Doan Dong, Do Tan Sang, Nguyen Dinh Thien, Vu Le Huy, and Tran Duc Quynh
  The 3rd International Conference on Human-centered Artificial Intelligence (Computing4Human-2022)
  December 16th, 2022