Internship Details – XL8

1. Goals and Objectives

☞ Description of specific goals and objectives for this program

- Data is the heart of ML systems. We'd like to use this internship program to seek a talented intern who can quickly learn and contribute to our data pipelines. Meanwhile, it'd be also a great opportunity for both the intern and the company to potentially pursue a long-term relationship even after the intern's graduation from their school, e.g. converting a full-time position either in Silicon Valley or in Korea .

2. Working Conditions and Environment

a. Working Conditions

☞ Description of compensation and benefits for Interns

- Experience with senior engineers from Google, Apple, and Qualcomm
- Learning opportunities from production ML data workflows and pipelines
- Monthly Compensation of $2,700
- Free lunches and snacks

b. Working Environment

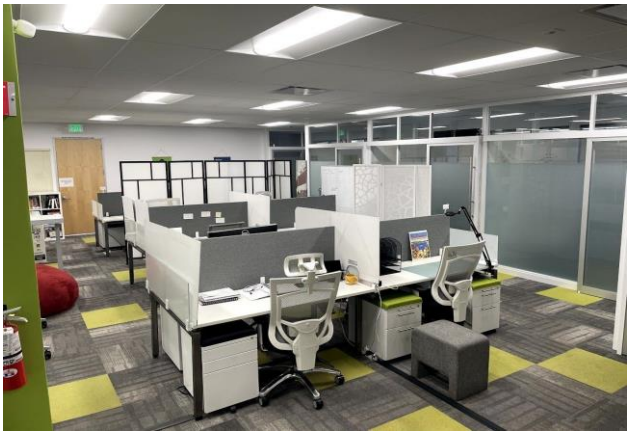
☞ Description of the site where Interns will work

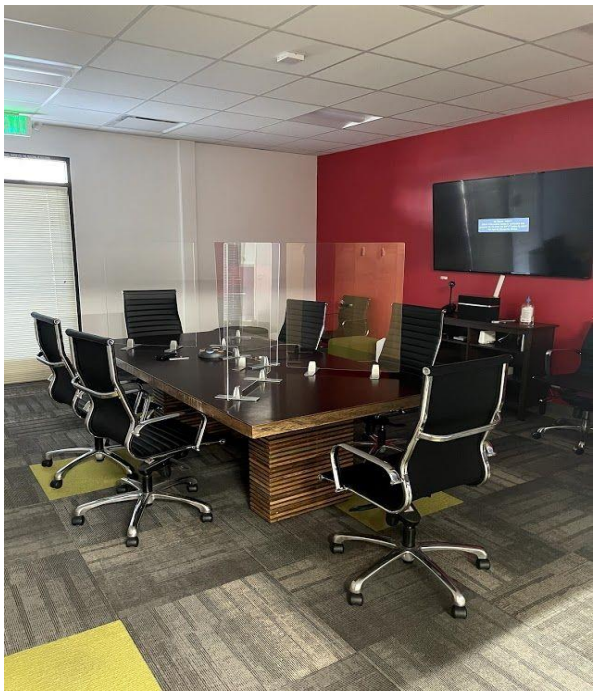
- Located at 690 Saratoga Ave., San Jose, CA 95129
- A private room office in a coworking office space, ZED
- Walking distance to many cafes and restaurants around

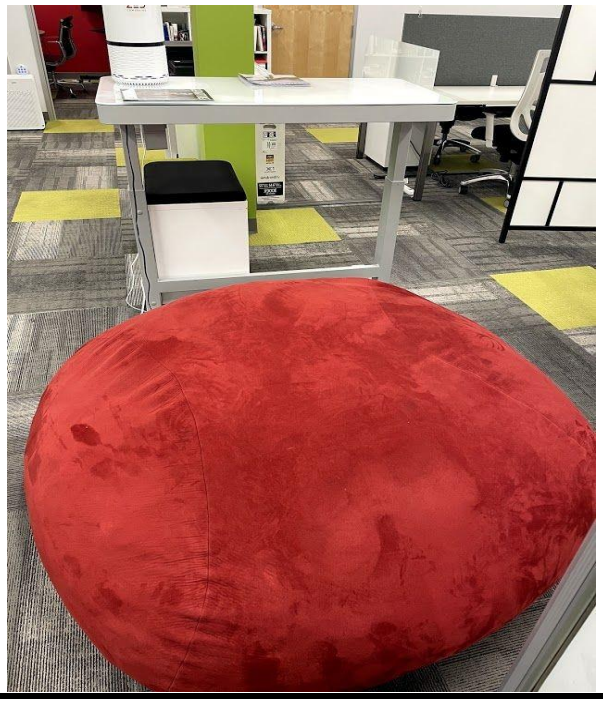
☞ Description of the Korea branch office where Interns will work for the time being(If applicable)

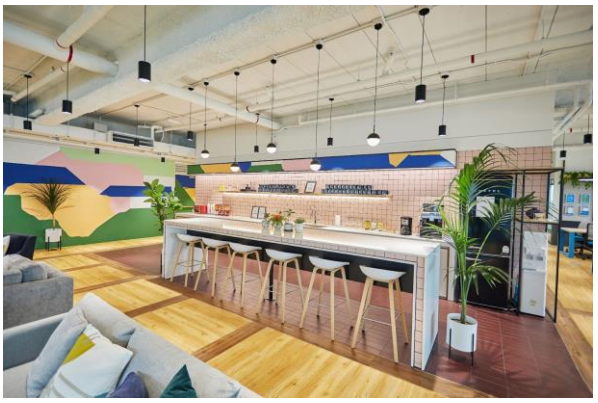
- Located at 431, Teheran-ro, Gangnam-gu, Seoul 06159, South Korea
- A private room office in a coworking office space, JustCo

c. Training Site Picture

U.S.A	Korea(if applicable)
   	 







3. Trainee's Roles and Responsibilities (in detail)

☞ Description of Interns' job description during the program

- Building Pipelines for ML systems. The intern will build a couple pipelines for various ML systems that we train and use for inference in production.
    - Data Preparation Pipeline
        - This pipeline downloads the source data stored in the AWS S3. It splits the data into three groups, 1) a train set, 2) a validation set, and 3) a test set. Then, it applies filtering using alignment algorithms.
        - This pipeline also confirms the consistency of the current and the previous data.
    - Training and Validation Pipeline
        - This pipeline runs data binarization and training scripts with a target sentence pair data.
        - This pipeline keeps the users up-to-date regarding its status and uploads the final model and the metadata to S3 when it's completed.
        - This pipeline runs automatic evaluation of the trained model on validation and/or test sets and keeps track of the evaluation results on the server.
- Visualizing business data
    - Developing dashboards for
        - Data Pipelines
        - Evaluation History
        - Business Intelligence

☞ What specific knowledge, skills, or techniques will be learned?

- Knowledge on ML system infrastructure
- Training techniques of Machine Learning models
- Architectures of state-of-the-art Deep Learning models
- Data Engineering specifically for Analytics and NLP

4. Requirements for Position

☞ Which specific knowledge, skills, or techniques will be required to perform the tasks?

- Python
- Data Processing (such as Hadoop, Spark, etc)
- Apache Airflow (recommended but not required)
- Streamlit.io (recommended but not required)

5. Methodology of Training

☞ Include specific tasks and activities and/or methodology of training.

- Thorough peer code reviews from experts experienced at Apple, Google, and Qualcomm
- An iterative process of ideation, design, development, and feedback
- 1:1 with the mentor, also with other members of the company including CEO

6. Milestone

☞ Time schedule for the program

1. Data Preparation Pipeline (July 2021 - Aug 2021)
    a. Developing basic pipeline (data download, extraction, split, filtering, deploying, etc): July 2021
    b. Developing advanced features (consistency check, duplication removal, periodic data download, etc): Aug 2021

2. Training and Validation Pipeline (Sep 2021 - Oct 2021)
    a. Developing basic pipeline (data binarization, launching training and validation, model upload and notification): Sep 2021
    b. Developing advanced features (Identification of errors/exceptions during training and re-launching, extension of the pipeline to different types of models and tasks): Oct 2021

3. Data Visualization (Nov 2021 - Dec 2021)
    a. Developing a web interface to visualize the status of the data preparation, training and validation pipelines: Nov 2021
    b. Extending the web interface by supporting search data/model history, scheduling a new preparation/training tasks on the web: Dec 2021