Summer Internship at the Cancer Institute of New Jersey

Prerna Shukla

May 2024-October 2024

Summary of Work

- Worked as a Summer Intern from May 2024 to October 2024.
- Contributed to multiple projects involving data automation and system verification.
- Gained hands-on experience with medical data processing and S3 system management.

Initial Internship Goals

- Gain practical experience in developing software solutions to streamline data workflows.
- Apply programming skills to real-world problems involving data automation and system verification.
- Collaborate with a team to contribute to impactful projects, while enhancing technical and problem-solving abilities.

Project 1: BioBERT based - BIEMPA

- Objective: Automate the extraction and transformation of medical reports for analysis
- Tools Used: Java, JSON, CSV
- Contribution:
  - Developed Java programs that parsed and converted medical reports into JSON format to feed deep learning pipeline
  - Further transformed JSON data into CSV format, enhancing usability for analysis and facilitating data engineering between modules in the workflow
- Challenges:
  - Ensuring accurate transformation of complex data
- Outcome: Reduced manual data extraction effort and improved efficiency of data processing

Project 2: OnBase

- Objective: Coded a module that will be regularly executed in daily medical data transfer which assesses transfer integrity and feeds data ETL for data linkage
- Tools Used: Python, S3, MD5 checksum, logging
- Contribution:
  - Developed Python code to connect to and operates on S3-compatible buckets
  - Verified image existence
  - Established processing milestones by appending to a central log
  - Generated logs and set up email alerts for process monitoring
- Challenges:
  - Ensuring accurate verification of images and smooth file transfer to server
- Outcome: Streamlined bucket processing, significantly reducing manual checks and improved data handling efficiency

Key Learnings

- Technical Skills:
  - Advanced my skills in Java and Python.
  - Learned about data transformation and integrity validation.
  - Gained experience in automating workflows for large-scale data management.
- Soft Skills:
  - Problem-solving and overcoming challenges with data complexity.
  - Improved communication with mentors and team members.

Impact and Achievements

Personal Impact: Developed stronger technical capabilities, especially in automating data workflows.

Team Impact: The automated systems I built improved data processing efficiency and significantly reduced manual efforts.

Acknowledgements

Thank you to my team and especially to Dr. Wenjin Chen for the opportunity to learn and to work at the Cancer Institute of New Jersey this summer.

Future Goals

Short-term: Continue to develop my technical skills, focusing on data automation and integrity.

Long-term: Apply these skills in future projects and internships, particularly in fields like machine learning and AI.