Shifan Chen

646-724-3792 | lovekano233@gmail.com | he/him/his | New York City | GitHub: csf233csf

\

Education

CUNY Bernard M. Baruch College 2020/8 – Now

B.B.A in Computer Information System

- **Courses:** Intro to Python Programming, Object-Oriented Programming in C++, Visual Basic for Excel Programs

Technical Skills

Programming Languages： C#, Python, Java, C++, Typescript, Rust, VBA for Excel, HTML, CSS

Tech Stacks: VUE3JS, Django, Flask, Pytorch, Tensorflow, Selenium, Regex Construction, Scikit-learn, BeautifulSoup, Linux tools and commands, Bash Scripts

Professional Experience

Sigma AI (JD.Group Tech Building in Beijing) 2024/02/28 – 2024/06/29

Natural Language Processing / AI Intern

- Collaborated with XuekeWang, China's largest online educational platform, to process billions of educational documents, converting them into PowerPoint presentations for real world classroom use.
- Constructed and trained three BERT models (NER, Sequence Tagging, Text Classification) with over 10 million data units, solved issues with missing content and special paddings with BERT tokenizer.
- Wrote algorithms for header parsing and leveling, fine-tuned GPT for semantic tag verification, and built a Flask RESTful API integrating the BERT models, API task queuing for correct GPU memory allocation.
- Fine-tuned the LLaMA3 model from META for generating English reading articles and questions, constructed datasets, prompts, and workflows. Minimized the hallucination problems of Large Language Model.
- Participated in training a Computer Vision model for PPT pagination. Wrote the train script and augment Parser.

The Oil Technology / The Oil Cop (Remote) 2018/03 – 2022/09

Admin / Tech Maintaining

- Managed a team of 8 people and a community over 500 members.
- Maintained the backend of the services, hosting the backends to platforms like GCP, AWS, and Heroku.
- Solved the rate limit problems with discord api and implemented into the code.
- Coded a selenium-based Cookie Generator that is still being used among communities.
- Wrote the webpage for the company and its code maintenance.

Projects (Some Codes are open sourced on GitHub)

Nike Account Generator **C#**

An automatic script based on selenium and chrome driver, a fully functional script that supports SMS api to pass the mobile verification. Using lxml xpath to locate the web elements and stimulate human actions, the script can generate accounts while passing through Nike's Akamai anti-bot Detection. By generating random user-agents, each trial can build its own user-agent database that is reusable.

Repository: https://github.com/csf233csf/Nike-Account-Generator-written-in-Csharp

Technology Used: C#, Selenium, lxml, HTML parsing, User-Agent headers, Bogus

The Light House Project **Typescript**

This project is an interactive web platform that gathers and encourages users to create, upload, and share their artworks and stories about aliens to become a part of the website. Utilizing technologies like TypeScript, Vue 3, and Firebase, the site transforms user submissions into dynamic visual experiences. It features animations via Gsap, 3D models through Three.JS, and a sophisticated interface using Vuetify. **This project provides a creative space where art meets technology, allowing for real-time interaction and artistic collaboration.**

Website: https://www.thelighthouseproject.life/

Repository: https://github.com/csf233csf/UFO-project

Technology Used: Vue.JS, Typescript, Gsap, ThreeJS, Sass, Firebase database, Vuetify, Nginx, Cloudflare

Wordtoppt (Sigma AI) **Python/C++**

The Wordtoppt project converts Word documents into HTML for semantic tag parsing. It employs three different BERT models and GPT to ensure precise semantic segmentation. The processed HTML documents are then transformed into PowerPoint presentations using pre-designed templates that have sophisticated animations and styles. This technology enhances the educational content, making it more interactive and visually appealing for teaching purposes. The PowerPoints generated have millions of downloads and usages among teachers.

This project consists of 4 key modules:

- **Pre-cleaning**: Split complex tables and remove useless elements, make good alignment and style attribute, ensure that the html documents were clean for after-process.
- **Semantic tagging**: Tag the elements with AI, the models were trained in massive datasets that were labelled by humans.
- **Answer Insertion**: The baseline algorithm is called DNA sequence insertion that inserts answers from answer sheet to questions. The algorithm is improved so that even if some semantic tags were wrong, most of the insertion are checked to be correct.
- **PPT Generation**: Utilize NodeJS and AIGC (Computer Vision model) to generate PowerPoints.

Technology Used: Python, Pytorch, HuggingFace Transformers, AsposeWords Microsoft, Pyquery, BeautifulSoup, Regular Expression, LLaMA3 Meta, Git, OpenCV.

Lottery System Bot on Discord (The Oil Technology) **Python**

It is a discord bot that interacts with users by discord webhooks and reactions. The bot uses Dhooks and discord.py to communicate with discord's api. MongoDB is used to store user data including user id and lottery points.

Technology Used: Python, discord.py, MongoDB, Dhooks