Sean W Evans

☎ +1·516·330·2517
✉ Sean.W.Evans@gmail.com
github.com/SeanWEvans

Skill

Languages

- Python (6yrs)
- LaTeX (6)
- C# (5)
- C (5)
- Perl (4)
- Batch/CMD (4)
- Bash (4)
- PowerShell (4)
- C++/CUDA (2)
- JavaScript/CSS (1)
- Java (1)

Tools

- OpenCV
- PyTorch
- Tensorflow
- CMake
- Make
- git
- MathJax
- PostGresSQL
- T-SQL
- PostScript
- GhostScript
- Microsoft Office SDK
- Adobe SDK
- React
- libXML
- PCRE
- iTextSharp
- Linux CoreUtils
- Android Studio
- Gradle
- MathType
- ImageMagick

Experience

Data Conversion Laboratory Queens, NY
Machine Learning / Software Engineer *May 2018 – Current*

- **PDF Extraction**: *Implemented machine-learning model for automated document analysis and extraction*
  - Used detectron2 (PyTorch) to train a statistical model to recognize equations, tables and figures from image data.
  - Comparable accuracy to industry leaders, including Microsoft Word and Adobe Acrobat
  - Extremely general, can estimate document chunk types across several different document domains
- **Document Auto-Styling**: *Used natural language processing techniques to automate a manual styling process*
  - Used SpaCy and Microsoft Office SDK to automatically style Word documents
  - Reduction of human touch-time on documents by 50%
  - Potential to completely eliminate manual process
- **Equation OCR**: *Implemented machine-learning model to estimate LaTeX markup from raw image data*
  - Implemented CNN-RNN, Encoder-Decoder in Torch7 / Lua to convert images into LaTeX markup
  - Reproduced state-of-the-art results from academic paper
  - Used perl to create large, high-quality training dataset from business XML
- **Checkbox Detector**: *Designed and implemented an optical mark recognition engine*
  - Used OpenCV to achieve greater than 96% accuracy on arbitrary checkboxes
  - Multiple checkbox geometries and fills supported
  - Prototype in python, production version to be built in C++
- **PDF Cleaning**: *Designed and implemented an PDF cleaning regiment based on common computer vision techniques*
  - Batch PDF processing system designed to deskew, despeckle, rotate and OCR all pages
  - Real world throughput on the order of 100,000 pages per week per 8-core server
  - Used as a back end in a dynamic template generation application to quickly view and classify examples across documents
- **JATS Quick Fixes**: *Designed and implemented automated LaTeX fixes inside of JATS xml documents*
  - Used libxml to parse JATS xml and fix the contained LaTeX equations based on commonly made errors
  - Decreased manual intervention from 40 hours per week (dedicated position) to less than 3 hours per week
  - Intimate knowledge of LaTeX and mathjax required
- **General .docx Converter**: *Used Microsoft Office SDK to convert any format Microsoft Word can handle to .docx*
  - Microsoft Office SDK and Microsoft InterOp libraries used to handle conversion
  - Production level options built-in including accepting tracked changes automatically and batch processing
- **General**: *Miscellaneous programming*
  - XML parsing using DOM & SAX parsers
  - Unit, regression & integration testing
  - Bash and Batch scripting
  - Test driven development
  - Agile development