

Amanpreet Singh

✉ amanpreet.singh@stonybrook.edu

☎ +1 631-312-2565

in /amanpreet-singh-k

🔗 /amanpreet692

🎓 EDUCATION

- **Stony Brook University** Stony Brook, NY
Master's in Computer Science; GPA: 3.97 2019 – 2021
 - **Thesis:** Sequence Labeling for Network File System Specifications
Advised by: Prof. Niranjan Balasubramanian
 - **Courses:** Natural Language Processing, Machine Learning, Data Science, Probability & Statistics
- **University of Mumbai** Mumbai, India
Bachelor of Engineering in Information Technology; First Class with Distinction (72.9%) 2011 – 2015
 - **Courses:** Data Structures & Algorithms, Artificial Intelligence, Discrete Mathematics, Databases

💡 RESEARCH INTERESTS

Common Sense Reasoning, Natural Language Generation, and Explainability of models

📖 PUBLICATIONS

- **Singh, A., & Balasubramanian, N.** (2020), “Open4Business (O4B): An Open Access Dataset for Summarizing Business Documents”, *Workshop on Dataset Curation and Security - NeurIPS 2020*
- **Nayak, A., Acharya, N., Singh, A., Sakthapara, A., & Geleda, B.**(2015), “Visualization of Mechanics Problems based on Natural Language Processing”, *International Journal of Computer Applications*, 116(14)

🏗️ PROJECTS

- **NER for system specifications:** Dataset annotation using Brat and code sequence classification in network file system specification documents by fine-tuning pre-trained language models. The language model itself is first fine-tuned on a manually scraped corpus for better domain adaptability. The lack of a large training set for the end task is a challenge.
- **Startup Acquisition Prediction:** Implementation and evaluation of different ensemble methods including anomaly detection, Naive Bayes and random forest on highly imbalanced data to predict whether a startup will be acquired.
- **Toxic Online Comments:** Multi-label toxicity detection in Wikipedia Comments and transfer learning effectiveness of the classifier on Twitter dataset. Analyzed the results of stacked LSTM/GRU against BERT and distilBERT models.
- **Long Documents Classification:** Parsing and multi-class categorization of documents with over 10k tokens using Bag of words, Tf-Idf, Doc2Vec and Attention based Neural models.
- **Chess Player Ratings:** Predicting the Elo rating of a chess player from the moves sequence. Efforts involved EDA and feature engineering using Pandas and Matplotlib; as well as modeling with Linear Regression and Random Forest.
- **Physual:** A text to scene generation system to visualize Physics problems with Stanford NLP, Java3D and Blender.

💼 EXPERIENCE

- **SS&C Intralinks** Waltham, MA
Machine Learning Engineer Intern (NLP) May 2020 – Dec 2020
 - **Abstractive Summarization:** Deep learning and REST service based business document summarizer:
 1. Curated and published a dataset of 18k open access business articles with their abstracts as summaries.
 2. Improved ROUGE score of SOTA models like BART and T5 by more than 10 points via fine-tuning.
 3. Built a custom encoder-decoder for T5 model to compress large inputs and avoid memory constraints.
 4. Adapted existing seq2seq model to ONNX quantization format reducing size by 75% and inference time by 30%.
 5. Flask based service to return raw abstractive summary with highlighted essential parts of a PDF.
- **J.P. Morgan Chase & Co.** Mumbai, India
Senior Software Development Engineer Feb 2018 – Aug 2019
 - **NLP Query Service:** An interactive system to resolve user queries that uses a model trained on the CRF classifier from StanfordCore NLP and returns the nearest possible solution from an existing knowledge base.
 - **Trader Analytics:** Introduced statistical enhancements in the core application such as absolute and percent variance, market share and standard deviation of historical stock prices to aid in trading decisions.
 - **Real-Time Pricing:** Developed a component using Spring, JMS and TDD principles that approximates real-time market risk using live prices; and publishes out the result. It helped retire a legacy system saving the firm ~\$250k.

- **Risk Management System:** Worked extensively on the core app used by traders for visualizing and hedging risk;
 1. Optimized the data feed using LMax Disruptor, a low latency Java queue for upto 20% faster trades processing.
 2. Framework to validate critical live market data results which reduced manual testing effort by 90%.
 3. Mechanism to switch from a MongoDB replica set to standalone instance in the event of a data center failure.

⚙️ TECHNICAL SKILLS

- **Languages:** Python, Java, Unix Shell Scripting, SQL, MATLAB
- **Frameworks:** PyTorch, TensorFlow, HuggingFace(Contributor), Pandas, NumPy, Scikit-Learn, Spring, Swagger
- **Databases:** Sybase ASE, MongoDB, MySQL

📖 TEACHING

- **Stony Brook University** Stony Brook, NY
Teaching Assistant Aug 2019 – May 2020
CSE 214 Data Structures in Java
CSE 354 Natural Language Processing
CSE 416 Software Engineering
- **K.J. Somaiya College of Engineering** Mumbai, India
Guest Lecturer Mar 2019
Seminar on benefits of using Spring and SpringBoot frameworks for client-server apps for undergrad juniors.
- **CrossNibble Technologies** Mumbai, India
Technical Instructor Mar 2015 – July 2015
Course Designer and instructor for hands-on Java programming for a class of 25 undergrads.

🔗 REFERENCES

- **Niranjan Balasubramanian, Ph.D.**
Assistant Professor
Stony Brook University, NY
niranjan@cs.stonybrook.edu
- **Prakash Kanchinadam**
Sr. Director, Engineering
SS&C Intralinks, MA
pkanchinadam@intralinks.com