PRATEEK GUPTA

+1-(914) 525-8510 ♦ 155 Claremont Ave, NY, 10027 ♦ pg2455@columbia.edu ♦ GitHub: prateekpg2455

Research Interests: Machine Learning, Natural Language Processing, Deep Learning, Bayesian Statistics, Graphical Models, Semi-Supervised Learning, Scalable Machine Learning, Data Visualization, Application in different domains

EDUCATION

Columbia University **New York, NY**
Master of Science, Operations Research, GPA: 3.97 / 4.00 ***Feb '15***

- *Relevant Coursework*: Statistical Machine Learning, Probabilistic Graphical Models, Applied Data Mining, Analysis of Algorithms, Simulation, Stochastic Modeling, Deterministic Modeling, Quantitative Finance
- Winner, Cornell Tech Data Hackathon, New York
  - We considered the problem of estimating profit/loss from solar panel installation at a given geographic coordinates.
  - Mined through APIs from NASA and US Government for data of solar intensity and solar panel installation cost across US
  - Used K-means clustering on satellite images obtained from Google Maps API to approximate sunlight exposed area

Indian Institute of Technology (IIT), Delhi **New Delhi, India**
Bachelor of Technology, Industrial Engineering, GPA: 8.69 / 10.00 ***May '13***

- *Relevant Coursework*: Probability and Statistics, Numerical Methods in Optimization, Econometric Methods, Calculus I and Matrix Analysis, Vector Calculus and Complex Analysis
- Ranked 2nd in the department
- Awarded Roll of Honor for academic excellence and community service
- Awarded IIT Delhi Director's Merit Award for academic excellence, 2010-11

Online Course Certifications: Machine Learning (Stanford University), Machine Learning with Spark(UC Berkeley), Javascript, D3, AJAX, Jquery (Udacity), Competitive Strategy (Ludwig-Maximilians University)

RESEARCH WORK

Body-Headline Latent Dirichilet Allocation to explore the relationship between body topics and headline topics under the guidance of Prof. Garud Iyengar, Prof. Martin Haugh and Prof. David Blei. Final report can be read at https://goo.gl/0mlcfO

Gautham, B. P., Gupta, P., Kulkarni, N. H., Panchal, J. H., Allen, J. K., & Mistree, F. (2013, August). Robust Design of Gears With Material and Load Uncertainties. In *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. V03BT03A046-V03BT03A046). American Society of Mechanical Engineers.

RESEARCH EXPERIENCE

Body-Headline Latent Dirichilet Allocation, *NLP Using Graphical Models* **New York, NY**
Prof. Garud Iyengar, Prof. Martin Haugh, Prof. David Blei (Columbia University) ***Aug' 14 –Dec' 14***

- We considered the problem of finding correlation between topics at different levels in an article i.e. title and paragraphs
- Designed the graphical model similar to LDA to include different levels and derived the Collapsed Gibbs Sampling update equations for the same; Implemented the algorithm using C++ and tested the results on NYTimes Academic Corpus.
- Described measures like KL Divergence and Hamming distance to assess article quality from correlated word distributions

Entire report can be read at https://goo.gl/0mlcfO

Counting Number of Trees using Satellite Images, *Computer Vision using Deep Learning* **New York, NY**
Prof. Bud Mishra (New York University) ***July' 15 -Present***

- We discussed, with emphasis on scientific rationale behind Feedforward, Convolutional, Recurrent Networks and regularization methods as discussed in Yoshua Bengio's yet unpublished book. Lecture notes on regularization methods as prepared by me can be found at: https://cs.nyu.edu/mishra/COURSES/15.Summer/L4DNN.pdf.
- We hope to be able to employ convolutional neural networks to count number of trees in NYC Manhattan area using satellite images. Ongoing efforts are being made using Caffe and Deep Visualization Toolbox using Python.

Project is well documented and is open sourced. It can be found here: https://github.com/buddeep

Smart Text Editor, *Natural Language Processing Tools For Journalists* **New York, NY**

The New York Times, R&D Data Scientist ***May '14 – Aug '14***

- We built the smart text editor to aid journalists in creating high quality meta features right from the beginning by designing a framework of micro-services based on NLP APIs to handle thousands of request per second
- Trained Decision Tree classifier with 97% accuracy on Treebank dataset to identify end of sentences in an article
- Trained word vectors using Google's Word2Vec recurrent neural net model on NYTimes Corpus and built a K-NN classifier for classifying news articles into multiple categories to aid in tagging of articles
- Built an entity resolution algorithm for matching noun phrases and its possible literary variants. Identification of Noun Phrases was implemented via parts-of-speech tagging and use of contextual free grammar for chunking.
- We studied the ways to capture topic trends in discussion among group of people through applying LDA on websites visited and speech recorded. We hope to be able to predict when the group is nearing its project, needs directions or needs a break. This component involved the use of Redis, beanstalkd, Python and D3 for visualization.
- Trained a decision tree classifier with accuracy of 70% to remove boilerplate from a website by classifying each HTML tag as useful or not. We built a manual tagging system and tagged around 600 HTML tags using the same.

Source code of standalone services can be found at https://github.com/nytlabs/linguo. Demo of the final Text Editor incorporating above services can be found at NYT R&D Labs as shown in this video: http://nytlabs.com/projects/editor.html.

Designing Optimized Gears For Reliability, Robustness & Cost, *Design Optimization* **Pune, India**

Prof. F. Mistree, Dr. B P Gautham (Tata Research, Design & Development Center (TRDDC)) ***May '12 – Aug '12***

- We worked on designing gears optimized for reliability, robustness and cost using available data as opposed to relying on prescribed safety factors by AGMA
- Formulated the Compromise Decision Support Problem (cDSP) in form of bounds, constraints and multiple goals on mechanical and physical properties of gear and its material such as uncertainty in material strength and force on gear obtained from data
- Optimized Adaptive Linear Programming (ALP) code as implemented in DSIDES (Decision Support In Design of Engineering Systems) and used it for solving above formulation; Algorithm uses non linear approximation of the decision function and optimizes the linear formulation of the problem iteratively
- We presented the comparative study of design factors in optimized methodology and empirical method in different scenarios determined by weights on goal which convinced the utility of the new methodology

Work was a close collaboration of TRDDC with Prof. Farrokh Mistree, Prof. Janet K. Allen, Prof. Jitesh H. Panchal and eventually was presented in August, 2013 at ASME –IDETC conference in Portland, Oregon by a scientist from TRDDC

PROFESSIONAL EXPERIENCE

GenesisMedia LLC, *Startup in Advertisement Technology Space* **New York, NY**

Data Scientist ***July '15 -Present***

I wear a lot of hats ranging from product innovation, product design, developing data visualizations, software engineering, data engineering but machine learning engineer and researching under guidance of Prof. Bud Mishra is my main role.

- NLP Researcher – Benchmarked POS taggers like Conditional Random Field, Maximum Entropy and Perceptron and employed the one with optimized time and F-Score. Built a Noun Phrase extractor based on Contextual Free Grammar Rules that gave fastest run time thereby employing the algorithm on millions of pages every hour. Thus, enabled Keyword extraction from websites using the above method and Collocations for n-grams.
- Built an ensemble of classifiers to define similarity score of a website with classes such as Food, Health, etc. Ensemble was designed optimally after grid searching hyper-parameters in logistic regression, SVM and passive-aggressive algorithms. Scores were calibrated using Isotonic Regression on output to prevent extreme prediction on probabilities.
- Designed interactive data visualizations using D3 and Javascript to display control and effectiveness that the above tools can give to customers. Currently, classifier based on Effron's FDR is being built to improve upon present models.
- Initiated data collection effort by building a modular crowd sourcing platform from scratch that can fit into various use cases
- Currently building a large scale experimental/optimizing capability that can run on millions of websites simultaneously and can consider multiple scenarios. Algorithm uses a mix of exploration/exploitation strategy and Thompson Sampling technique as a solution to multi-armed bandit problem.
- User Behavior Research – Developing an online learning model to classify users based on their response to content on a webpage. Currently, experimenting with graphical and neural network models to classify the users.

American Express **New York, NY**
Data Scientist Manager ***Feb '15 - Jun '15***

- Improved the data quality by implementing zip level variables from billions of transactions by consumers using credit cards
- Built a binary classification model pipeline from data cleaning to optimization that predicts if the prospect can be acquired

GetWiser, *Startup in News Aggregation Space* **New York, NY**
Machine Learning Engineer ***Jan '14 – May '14***

- We considered the problem of designing news recommendation engine to increase user engagement with everyday news
- Used PCA to reduce dimension and eliminate noise from data and built a one-class SVM with 70% accuracy to predict user engagement with the article. One-class SVM formulation was scalable and matched the behavior of user reading behavior.

Skills & Leadership Experience

Programming Skills: Python, R, C, Java, Javascript, Bash, JQuery, HTML, D3, SQL**,** Hadoop, Redis, MongoDB, MATLAB

Societies: NYU Deep Learning, Columbia University Machine Learning Group, Columbia Data Science Society, Data Science Meetups NY, IIT Sports Society (General Secretary), IIT Academic Council (Convener), IIT Delhi Table Tennis (Captain/Vice-Captain with 15 years at competition level), CrossFit

Interests: MOOCs, Data Visualization, Hacking, Hardware Hacking, Programming**,** Crossfit**,** Table Tennis, Yoga, Skating, Fencing, Music, Cubing, Running

References

Prof. Garud Iyengar
IEOR Department
School of Engineering & Applied Sciences
Columbia University
New York, NY, USA
Email: garud@ieor.columbia.edu

Prof. Bud Mishra
Computer Science & Mathematics,
Courant Institute of Mathematical Sciences
New York University,
New York, NY, USA
Email : mishra@nyu.edu

Prof. Martin Haugh
IEOR Department
School of Engineering & Applied Sciences
Columbia University
New York, NY, USA
Email: mh2078@columbia.edu

Dr. BP Gautham
Principal Scientist
Tata Research Development and Design Centre (TRDDC)
Pune, India
Email: bp.gautham@tcs.com