Gaurav Baruah

*Applying for a **Data Scientist** position*

gauravbaruah@gmail.com https://cs.uwaterloo.ca/~gbaruah +1-519-722-6825

SUMMARY OF EXPERIENCE

- 5+ years of experience in processing, building models, and running analyses over big (multi-terabyte) data.
- Ph.D. in Computer Science focusing on information retrieval: researched quality assessment for information stream filtering systems using user models and simulations.
- 4+ years of industry experience in product and service oriented startups.

SKILLS, TOOLS AND TECHNOLOGIES

- *Data Science*: Convolutional Neural Networks, Information Retrieval, Natural Language Processing, text data mining and analysis, user modeling and simulation.
- *Data processing/analysis*: Python, R, PyTorch, MongoDB, MySQL, sqlite, Elasticsearch, json, XML.
- *Distributed Computing*: Amazon Web Services (EC2/S3), SHARCNET, map-reduce.
- *Programming*: C, C++, Python, Thrift, SQL, Java, bash, git, svn, maven.
- *Web development*: MEAN, bootstrap, angular.js, jquery, LAMP, HTML.

WORK EXPERIENCE

University of Waterloo, ON, Canada **Post-Doctoral Fellow** **Oct 2016 – present**

- Implemented deep learning models using PyTorch for matching questions with candidate answers.
- Researched a user modeling and simulation based evaluation method for push notification systems.
- Developed a web-app using the MEAN stack to gauge users' preferences for timeline summaries.

Sciencescape, Toronto, Canada **Contractor (part-time/remote)** **May 2015 – Jul 2015**

- Converted various metadata in pseudo-XML to a standardized format using Python.

University of Waterloo, ON, Canada **Graduate Research Assistant** **Sep 2011– Aug 2016**

- Collaborated in devising active learning methods to expedite procurement of labeled data required for automatic evaluation of timeline summarization algorithms.
- *PhD Thesis*: Devised a quality assessment framework that incorporates models of user behavior to estimate the utility of news search/filtering systems; researched factors affecting evaluation of such systems.
- Gained experience in processing big streaming data like the KBA Stream Corpus (~16.1 TB compressed, ~1.2 billion documents) and the Clueweb12 dataset (~27 TB, 733 million documents) over computing clusters.
- Developed a news search/filtering system using C++, Python and bash, to retrieve sentences relating to a breaking news event (e.g. earthquake/storm) from a time-ordered stream of ~1 billion documents.
- Assisted in investigations on search engine quality assessment using crowd-sourced judgements.
- Provided assessment and feedback to students as an Instructional Apprentice/Teaching Assistant.

Geodesic Limited, Bangalore, India **Lead Engineer** **Jul 2009 – Jul 2011**

- Led a team of 4 engineers to develop Linux applications using C, C++, GTK, pthreads and PDF libraries for prototype devices with 2 touchscreens designed for students and researchers.
- Collaborated with the design team, management, and my team's engineers fix key deliverables and schedules.
- Worked on on-line curve smoothing for handwriting on low-resolution touch-screens; mentored students for developing geometric shape recognition algorithms.

Intellection Software and Technologies, Pune, India **Software Engineer** **Jun 2005 – Jul 2007**

- Designed and implemented edge-detection algorithms for images and 3D models using C++ for applications in quality control, reverse engineering and computer-aided design of mechanical parts.

Indian Institute of Technology, Bombay, India **Project Engineer** **Jul 2004 – Jan 2005**

- Developed a prototype application using VC++ and OpenGL for the design and analysis of metal castings given a library of metallurgical data in XML format.

EDUCATION

University of Waterloo, Ontario, Canada	**Ph.D.** in Computer Science	**Aug 2016**
Indian **I**nstitute of **T**echnology, Guwahati, India	**M.Tech.** in Computer Science & Engineering	**Jul 2009**
BVCOE, **Pune University**, India	**B.E.** in Computer Engineering	**Jul 2004**

Graduate Level Courses: Information Retrieval, Health Informatics, Computer-mediated Advertising, Probabilistic Inference and Machine Learning, Pattern Recognition, Speech Recognition, Mobile Robotics, Embedded Systems, Computational Geometry, Information and Randomness, Computer Systems, Distributed Systems.

RESEARCH PUBLICATIONS

- Gaurav Baruah, Richard McCreadie, Jimmy Lin, *"A Comparison of Nuggets and Clusters for Evaluating Timeline Summaries"*, to appear at CIKM 2017, Singapore.
- Gaurav Baruah and Jimmy Lin, *"The Pareto Frontier of Utility Models as a Framework for Evaluating Push Notification Systems"*, to appear at ICTIR 2017, Amsterdam, Netherlands.
- Royal Sequeira, Gaurav Baruah, Zhucheng Tu, Salman Mohammed, Jinfeng Rao, Haotian Zhang, Jimmy Lin, *"Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End Question Answering"*, SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR 2017), Tokyo, Japan.
- Luchen Tan, Gaurav Baruah, and Jimmy Lin, *"On the Reusability of "Living Labs" Test Collections: A Case Study of Real-Time Summarization"*, SIGIR 2017, Tokyo, Japan.
- Gaurav Baruah, Haotian Zhang, Rakesh Guttikonda, Jimmy Lin, Mark D. Smucker, and Olga Vechtomova, *"Optimizing Nugget Annotations with Active Learning"*, CIKM 2016, Indianapolis, U.S.A.
- Jimmy Lin, Charles L.A. Clarke, and Gaurav Baruah, *"Searching from Mars"*, IEEE Internet Computing, 2016
- Gaurav Baruah, Adam Roegiest, and Mark D. Smucker, *"Pooling for User-oriented evaluation measures"*, ICTIR 2015, Northampton, U.S.A.
- Gaurav Baruah, Mark D. Smucker, and Charles L. A. Clarke, *"Evaluating Streams of Evolving News Events"*, SIGIR 2015, Santiago, Chile.
- Gaurav Baruah, Adam Roegiest, and Mark D. Smucker, *"The Effect of Expanding Relevance Judgements with Duplicates"*, SIGIR 2014, Gold Coast, Australia.
- Gaurav Baruah, Rakesh Guttikonda, Adam Roegiest and Olga Vechtomova, *"University of Waterloo at the TREC 2013 Temporal Summarization Track"*, TREC 2013, Gaithersburg, U.S.A.

OTHER RELEVANT PROJECTS

- *Castorini.io*: Deep learning and Information Retrieval synthesis projects at the University of Waterloo.
- *Master's thesis*: a meta-search engine using Java, Maven, MySQL and Google/Freebase APIs that re-ranked search results and recommended related queries based on Freebase tuples contained in the resultant documents.
- Developed a search engine for personal collection of research papers; extracted citations and references in order to present a concise view for a given research paper using Elasticsearch, MEAN and Python.
- Developed a web-app to tag papers by topic and make summarizing notes for facilitating literature review.
- Trained logistic regression and neural network classifiers to infer the gender of Twitter users.
- Built a speech-recognition system using HMMs.
- Developed a program to gauge English proficiency of night-school students using a fixed number of questions

SERVICE AND VOLUNTEERING

- Program Committee member for ACM SIGIR 2017, ACM ICTIR 2017 and ACM CIKM 2017 conferences.
- Placement Representative for the class of 2008-09 at IIT Guwahati; assisted Placement-Cell in coordinating students during campus placements; developed a web-app to consolidate student and company information.
- Student head of undergrad college Art-Circle; selected and coordinated teams for various inter-collegiate performing arts competitions; also played lead guitar in college band.