Logout

Confirmation of Application Receipt:

Your proposal was successfully submitted to the Google. No further action on your part is required and you can expect to receive notice of your proposal's status shortly. To print a copy of this completed application go to 'File', then 'Print' on your browser toolbar. Click here to return to the homepage when you are finished.

Contact Information

PI Name (required)	Frank
PI Last Name (required)	Wood
PI E-mail Address	fwood@stat.columbia.edu
PI Website	http://www.stat.columbia.edu/~fwood
Telephone	6467172202
University	Columbia University
Department	Department of Statistics
Department Website	http://www.stat.columbia.edu/
Department Mailing Address (required)	1255 Amsterdam Avenue
Department City (required)	New York
Department State (required)	New York
Department Zip Code (required)	10027
Department Country	United States
Contact Type (required)	Principal Investigator

Contact Information

PI Name (required)	David
PI Last Name (required)	Madigan
PI E-mail Address	madigan@stat.columbia.edu
PI Website	http://www.stat.columbia.edu/~madigan
Telephone	212.851.2132
University	Columbia University
Department	Statistics
Department Website	http://www.stat.columbia.edu/
Department Mailing Address (required)	Department of Statistics, 1255 Amsterdam Avenue
Department City (required)	New York

Department State (required)	New York
Department Zip Code (required)	10027
Department Country	United States
Contact Type (required)	Co-Principal Investigator

Organization Information

Legal Name (required)	Columbia University
Address (required)	Department of Statistics, 1255 Amsterdam Avenue
City (required)	New York
State (required)	New York
Zip (required)	10027
Province	
Country	United States
Telephone	212.851.2132
Fax	212.851.2164
Website Address	http://www.columbia.edu/

Proposal Information

Proposal Title (required)	Scalable sequence memoization for natural language modeling and lossless compression
Requested Cash Amount (required)	$123,947.00
Google Sponsor	
Google Contact	
Primary Topic (required)	Machine Learning and Data Mining
Secondary Topic	Machine Translation Natural Language Processing
Keywords (required)	Bayesian nonparametric inference, n-gram natural language modeling, sequence memoizer
Research Abstract and Goals (required)	We propose to develop and demonstrate scalable inference software for a Bayesian nonparametric (BNP) natural language model called the sequence memoizer (SM). We propose to use this software to train a SM using the trillion word Google text corpus and to empirically study the effect of taking into account long-range textual dependencies on language model performance as the amount of training data grows. We also propose to develop a latent variable extension of the SM and demonstrate scalable inference in

	the same. We propose to demonstrate both models and algorithms in two ways, general purpose lossless compression and n-gram natural language model performance. Goals * Develop a freely available, downloadable software development kit (SDK) that contains a scalable implementation of a constant space, linear time SM language model that demonstrably scales to sequences that are billions to trillions of tokens long. * Empirically explore the impact of being able to probabilistically model and exploit long contextual dependencies given Google-scale corpora on language model perplexity and compressor log-loss.
Application Packet (required)	first_submission.pdf (494.07 K)
Proposed Start Date (required)	06/01/10
Proposed End Date	
Measuring Progress	* Decreasing perplexity of a mutually agreed upon test corpus relative to prior art benchmarks * Increasing training sequence length scalability in terms of numbers of tokens processed per unit memory.
Expected Outcome and Results	Expected Outcomes * The primary result of this project will be the establishment of compelling evidence that supports the practicality and usefulness of Bayesian nonparametric language models for large scale commercial applications including machine translation, automated speech detection, and general purpose lossless compression. * We will develop a sequence memoizer SDK that will be downloaded and used by researchers in a wide-variety of industrial and academic fields. * We will contribute to the state of the art in Bayesian nonparametric modeling by developing an latent variable extension to the sequence memoizer.

Need Support?

Note: Submitting this form will send your information on the form to CyberGrants, Inc., who is managing this form on behalf of Google. The information is being collected only for the purposes served by this form and will not be shared with or sold to anyone else. To learn more, please see Google's Privacy Center.