Matrix Computations & Scientific Computing Seminar

Organizer: James Demmel & Ming Gu

Wednesday, 12:10–1:00pm, 380 Soda

May 2 **Brian Gawalt**, UCB

Convex Approaches to Text Mining

Many text mining tasks – ranking, classification, clustering – can be posed as the solving of a convex optimization problem over a vector-space model of the document set. However, the models fit by these methods are often dense, with one parameter fit per ngram token, creating models unsuitable for human understanding. When these text mining models are made sparse, via stringent l1 regularization, they yield valid, fully interpretable summaries of the underlying documents. The preserved convexity of the underlying problem suggests opportunity to scale to large corpora of documents.

The presentation will begin with an example vector-space text retrieval task, proceed to an introduction of the predictive (semi-to-fully-supervised learning) framework for summarization, and present results from a human validation experiment showing the merit of the sparse and convex approaches. It includes an extension of a simple predictive algorithm to a MapReduce framework for large-scale summarization and concludes with an exploration of a sparse topic modeling (unsupervised) approach to summarization.