

Kenneth Heafield

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave NSH 4502
Pittsburgh, PA 15213

<r at kheafield.com>
<http://kheafield.com>

Interests

Natural language processing, machine learning, theoretical computer science, distributed systems

Education

PhD program, Carnegie Mellon August 2008-
Language Technologies Institute in the School of Computer Science

Bachelor of Science, Caltech September 2003-March 2007
Double major in Mathematics and Computer Science
3.8/4.0 GPA, graduation with honors

Skills

Languages

Written extensively in C++, C, Ruby, SQL, UNIX Shell, \LaTeX , and HTML

Software

Linux, Hadoop, PostgreSQL, MySQL, Apache, Octave, Matlab, Gnuplot, and GTK

Hobbies

Volunteer system administrator for 1900 user Linux cluster

Experience

Google <http://books.google.com> March 2007-August 2008
As a Software Engineer with Google Book Search, I worked on a team that uses machine learning to compile card catalogs from multiple sources into a single coherent catalog of books. Previously, I created the scoring system behind a search function in Picasa Web Albums. To share Google's approach to distributed systems, I lectured at MIT on the Hadoop MapReduce framework.

Infosys Technologies <http://www.infosys.com> July-September 2006
I traveled to Bangalore, India to intern with the research division of Infosys, India's second largest software outsourcing company. The goal was to automatically organize source code files into a meaningful directory structure. I investigated a technique based on semantic information from the names of functions, local variables, and files. To derive topics from this information, I elected to use Latent Dirichlet Allocation and tweaked it to the domain of source code. As an example, it was able to identify both SSL and logging topics in Apache and correctly label a file covering both topics. Our results were presented at the 2008 India Software Engineering Conference.
Reference: Dr. Girish Rama <Girish_Rama at infosys.com>

Fastsoft <http://www.fastsoft.com> January-April 2006
Netlab spun off a startup and I worked for them as a part-time contractor. Using FAST TCP, the Netlab algorithm responsible for breaking Internet speed records, their Aria product accelerates connections passing through it. This allows senders to use high performance networks more efficiently without custom operating systems. I setup experiments and worked on the performance monitoring and configuration interface.
Reference: Prof. Steven Low <slow at caltech.edu>

Netlab <http://netlab.caltech.edu>

June 2005-June 2006

As a Richard and Dena Krown Summer Undergraduate Research Fellow, I developed an error model for kernel Principal Component Analysis (kPCA). Professor Low hired me to continue with implementation during the school year. I applied it to identify possible attacks in network traffic, which appear as points with unusually high distance from the manifold learned by kPCA.

Reference: Prof. Steven Low <slow@caltech.edu>

Galaxy Evolution Explorer <http://www.galex.caltech.edu>

June 2004-March 2007

I started working for the Galaxy Evolution Explorer (GALEX) project as a Summer Undergraduate Research Fellow. My goal was finding variable stars and asteroids in observations made by their satellite. To do so, I created a database of all 193 million source measurements and used it to find and analyze over ninety variable objects. The findings were reported in two posters and one journal article. After the summer, they hired me to continue working on the database and to help scientists find interesting data.

References: Dr. Mark Seibert <mseibert@srl.caltech.edu> and Prof. Chris Martin <cmartin@srl.caltech.edu>

Awards

National Science Foundation Graduate Research Fellowship	2008-
\$121,500 in stipend and tuition over three years	
International Collegiate Programming Contest Regional	2006-07
Third of fifty places as a team of two instead of three	
Carnation Scholarship	2005-06
Caltech full tuition academic merit scholarship, 38 awarded per year	
Richard and Dena Krown Summer Undergraduate Research Fellowship	2005
\$5,000 for ten weeks of summer research	
Summer Undergraduate Research Fellowship	2004
\$5,000 for ten weeks of summer research	

Publications

G. Rama, S. Sarkar, K. Heafield. Mining Business Topics in Source Code using Latent Dirichlet Allocation. Proceedings of the 1st *India Software Engineering Conference*, Hyderabad, India (Feb 19-22 2008), 113-120.

S. Browne, J. Wheatley, B. Welsh, M. Seibert, K. Heafield, R. Rich, and the GALEX Science Team. RR Lyrae Stars in the Far Ultraviolet: GALEX Observations Compared with Theoretical Predictions. *Bulletin of the American Astronomical Society, Poster Sessions* **37** (2006).

K. Heafield, S. Low. An Error Model For Kernel Principal Component Analysis. For *Summer Undergraduate Research Fellowship* (2005).

B. Welsh, J. Wheatley, K. Heafield, M. Seibert, et al. The GALEX Ultraviolet Variability Catalog. *The Astronomical Journal* **130** (2005), 825-831.

B. Welsh, J. Wheatley, K. Heafield, M. Seibert, S. Browne, and the GALEX Science Team. The Flaring UV Sky. *Bulletin of the American Astronomical Society, Poster Sessions* **36** (2005).

Publications and unofficial transcript are available at <http://kheafield.com/professional/>.