Dániel Varga

daniel@mokk.bme.hu

http://hu.linkedin.com/in/danielxvarga

+36-30-2125-888

Summary

Veteran applied mathematician / computer scientist with vast experience in working with large datasets, from text corpora through online friendship networks to geospatial data. A generalist: knows which tool to pick for the job, and picks it up fast. (The ecosystem of choice is Python's, but has written or touched code in 13 different programming languages at current job.)

Selected projects

- Image- and frame indexers at prezi.com, 2014, Core developer, code written in Haxe. Backend service heuristically associating objects and surrounding textual information on a zoomable 2D canvas for information retrieval purposes. Indexing throughput tens of thousands of images per hour.
- eovk.mokk.bme.hu gerrymandering web application, 2008, AI developer, code written in Python. A service demonstrating the adverse political effects of gerrymandering in Hungary by allowing users to build gerrymandered district maps online, according to their political preferences. Employs combinatorial optimization on geospatial data to create personalized district maps, and visualizes them on Google Maps.
- hunalign bilingual sentence aligner tool, 2005, one-person project, code written in C++ and Python. After 10 years, hunalign still dominates its small niche of fast-but-reliable sentence alignment: it was used to create five of the six largest published multilingual parallel text corpora, namely JRC-Acquis, DCEP, OPUS, Parasol, and InterCorp. 193 citations. I used it to help create the influential JRC-Acquis multilingual parallel corpus, 369 citations.
- wiwdaemon in-memory graph database and friend recommendation service, 2002, one-person project, code written in C++. Created for iwiw.hu, Hungary's dominant social networking service between 2002 and 2010. This was the first online friend recommendation service in the world that I am aware of. At its peak, it served 300k requests a day, on a friendship network of 5M nodes and 400M edges.

Programming skills

- Fluent: Python, Unix command line tools
- Conversational: SQL, C++, Pig, Haxe

Selected publications

- Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón. Parallel corpora for medium density languages. Recent Advances in Natural Language

Processing IV. Selected papers from RANLP 2005. John Benjamins 2007.
- Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In the Proceedings of the LREC 2006.
- Béla Janky, Dániel Varga. The poverty-assistance paradox. Economics Letters 120 (3). 2013.

Education

- PhD in Information Science and Technology, Eötvös Loránd University, Budapest, 2013. Summa cum laude. Thesis: Natural Language Processing of Large Parallel Corpora. Advisor: András Kornai
- MSc in Mathematics, Eötvös Loránd University, Budapest, 1996. Passed with high distinction (red diploma). Thesis: Algebraic Methods in the Theory of Boolean Circuit Complexity. Advisor: Katalin Friedl

Work history

- 2012-present: Senior Software Engineer at Prezi. Information retrieval, machine learning, and modeling-heavy data analysis.
- 2004-present: Research Fellow (previously Research Assistant) at the Budapest University of Technology, Department of Sociology, Media Research Centre. Currently on sabbatical. Selected projects: hunalign sentence aligner, hunglish.hu searchable English-Hungarian parallel corpus (10k monthly unique visitors, 200k monthly page hits), hunner named entity recognizer, EOVK gerrymandering web application.
- 2003: Researcher at Applied Logic Laboratory Ltd., Budapest. Created SoundStore, a speech indexing system.
- 1997-2002: Researcher at Mindmaker Ltd. Selected project: Backend for VoiceStudio, an IDE for semi-automatic creation of acoustic and intonation models for TTS from raw speech data.

Miscellaneous

- Ranked 510 at Google Code Jam 2011.