Booked for Brilliance: Sweden’s National Library Turns Page to AI to Parse Centuries of Data

The library is training state-of-the-art AI models on a half-millennium of Swedish text to support humanities research in history, linguistics, media studies and more.

Author: Isha Salian

For the past 500 years, the National Library of Sweden has collected virtually every word published in Swedish, from priceless medieval manuscripts to present-day pizza menus.

Thanks to a centuries-old law that requires a copy of everything published in Swedish to be submitted to the library — also known as Kungliga biblioteket, or KB — its collections span from the obvious to the obscure: books, newspapers, radio and TV broadcasts, internet content, Ph.D. dissertations, postcards, menus and video games. It’s a wildly diverse collection of nearly 26 petabytes of data, ideal for training state-of-the-art AI.

“We can build state-of-the-art AI models for the Swedish language since we have the best data,” said Love Börjeson, director of KBLab, the library’s data lab.

Using NVIDIA DGX systems , the group has developed more than two dozen open-source transformer models, available on Hugging Face . The models, downloaded by up to 200,000 developers per month, enable research at the library and other academic institutions.

“Before our lab was created, researchers couldn’t access a dataset at the library — they’d have to look at a single object at a time,” Börjeson said. “There was a need for the library to create datasets that enabled researchers to conduct quantity-oriented research.”

With this, researchers will soon be able to create hyper-specialized datasets — for example, pulling up every Swedish postcard that depicts a church, every text written in a particular style or every mention of a historical figure across books, newspaper articles and TV broadcasts.

The library’s datasets represent the full diversity of the Swedish language — including its formal and informal variations, regional dialects and changes over time.

“Our inflow is continuous and growing — every month, we see more than 50 terabytes of new data,” said Börjeson. “Between the exponential growth of digital data and ongoing work digitizing physical collections that date back hundreds of years, we’ll never be finished adding to our collections.”

Soon after KBLab was established in 2019, Börjeson saw the potential for training transformer language models on the library’s vast archives. He was inspired by an early, multilingual, natural language processing model by Google that included 5GB of Swedish text.

KBLab’s first model used 4x as much — and the team now aims to train its models on at least a terabyte of Swedish text. The lab began experimenting by adding Dutch, German and Norwegian content to its datasets after finding that a multilingual dataset may improve the AI’s performance.

The lab started out using consumer-grade NVIDIA GPUs, but Börjeson soon discovered his team needed data-center-scale compute to train larger models.

“We realized we can’t keep up if we try to do this on small workstations,” said Börjeson. “It was a no-brainer to go for NVIDIA DGX. There’s a lot we wouldn’t be able to do at all without the DGX systems.”

The lab has two NVIDIA DGX systems from Swedish provider AddPro for on-premises AI development. The systems are used to handle sensitive data, conduct large-scale experiments and fine-tune models.

They’re also used to prepare for even larger runs on massive, GPU-based supercomputers across the European Union — including the MeluXina system in Luxembourg .

“Our work on the DGX systems is critically important, because once we’re in a high-performance computing environment, we want to hit the ground running,” said Börjeson. “We have to use the supercomputer to its fullest extent.”

The team has also adopted NVIDIA NeMo Megatron , a PyTorch-based framework for training large language models , with NVIDIA CUDA and the NVIDIA NCCL library under the hood to optimize GPU usage in multi-node systems.

“We rely to a large extent on the NVIDIA frameworks,” Börjeson said. “It’s one of the big advantages of NVIDIA for us, as a small lab that doesn’t have 50 engineers available to optimize AI training for every project.”

In addition to transformer models that understand Swedish text, KBLab has an AI tool that transcribes sound to text, enabling the library to transcribe its vast collection of radio broadcasts so that researchers can search the audio records for specific content.

KBLab is also starting to develop generative text models and is working on an AI model that could process videos and create automatic descriptions of their content.

“We also want to link all the different modalities,” Börjeson said. “When you search the library’s databases for a specific term, we should be able to return results that include text, audio and video.”

KBLab has partnered with researchers at the University of Gothenburg, who are developing downstream apps using the lab’s models to conduct linguistic research — including a project supporting the Swedish Academy’s work to modernize its data-driven techniques for creating Swedish dictionaries.

“The societal benefits of these models are much larger than we initially expected,” Börjeson said.

Images courtesy of Kungliga biblioteket

Original URL: https://blogs.nvidia.com/blog/2023/01/23/sweden-library-ai-open-source/