Semi-automatic literature reviews

Sam Dornan

What exactly is a semi-automatic literature review?

Using machine learning to infer the relationship between words in a corpus

- Can review larger volumes of text
- Find relationships between disparate disciplines and fields of study
- Already used across material sciences
- Can it be applied to the social sciences?

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder & Anubhav Jain

Nature **571**, 95–98 (2019) | Cite this article

69k Accesses | **381** Citations | **1899** Altmetric | Metrics

How would it be done?

Social sciences is a part of the world as much as the world is a part of the social sciences

Large language models such as GPT and T-5 can “understand” the world.

Open-source programmes allow models to be tailored to a specific task.

Tailor a language model to the social sciences

The dataset

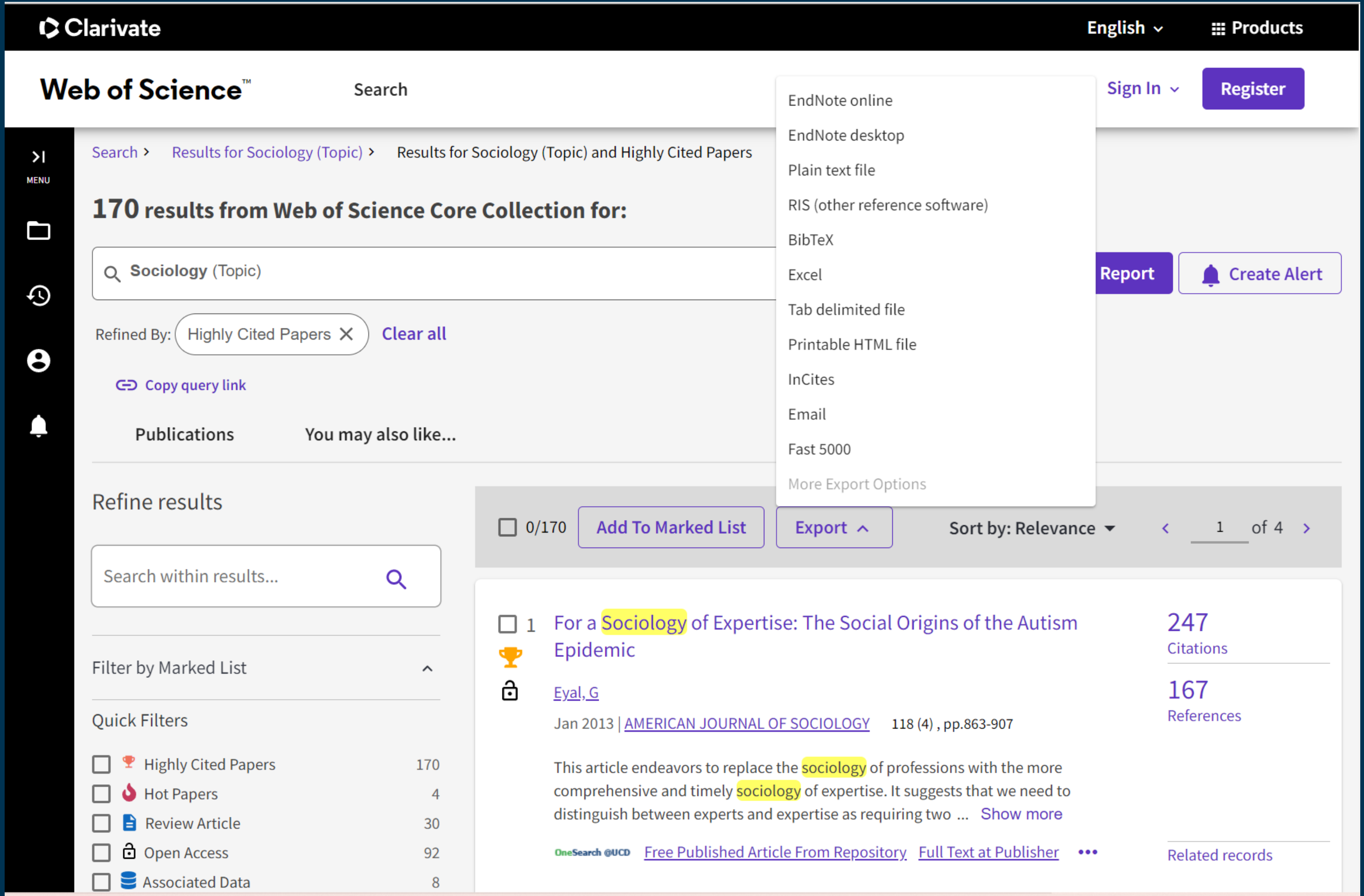
Easy to access

Easy to work with

Small-fits in memory

Tailored

Wisdom of the crowds



The manual part?

Garbage-in; garbage-out

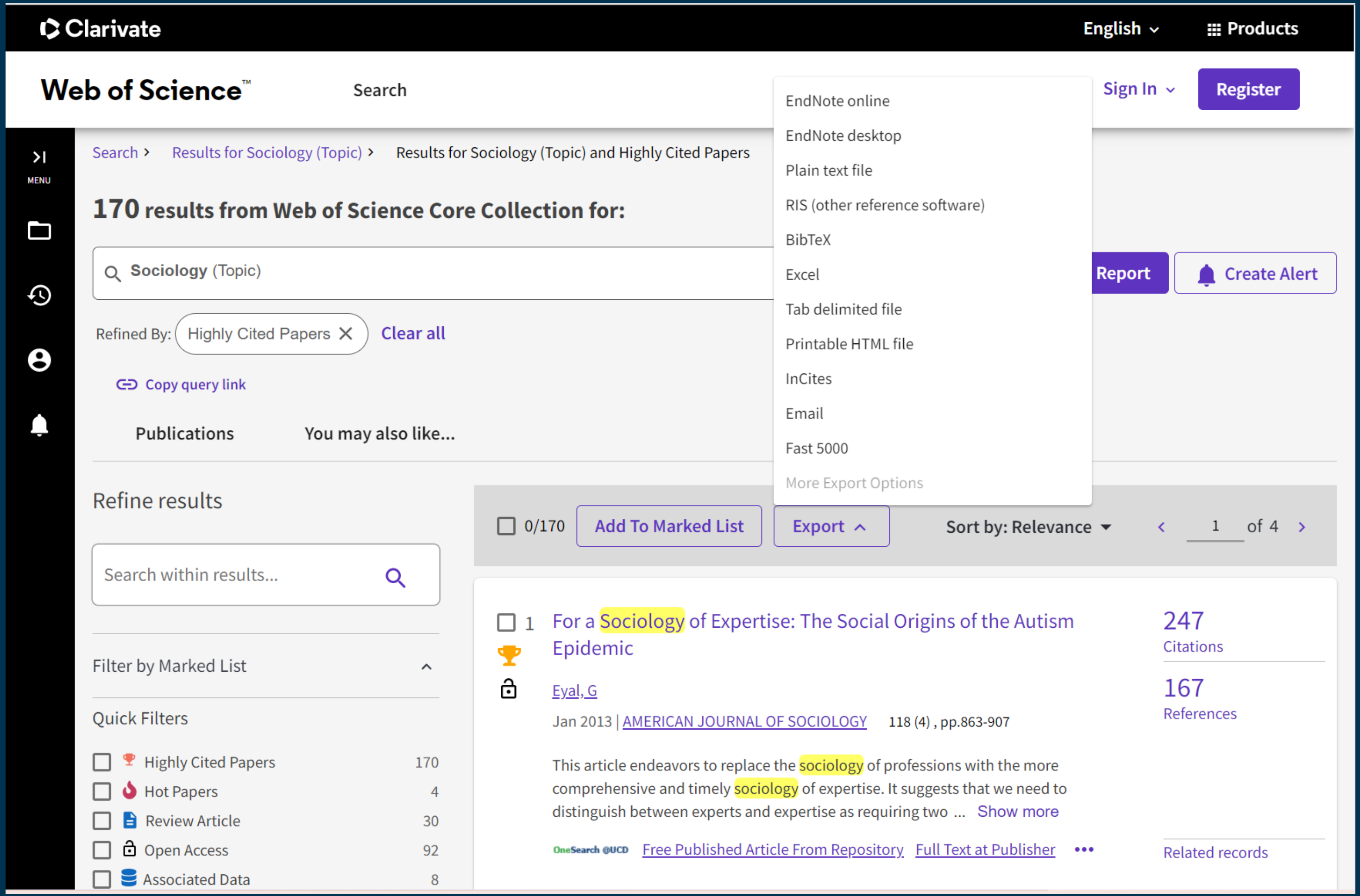
Preprocessing/Post-processing

Where's the data from?

What's in the data?

What language model do we use?

Why?



The Model

Pretrained

Small – fits in memory

Only compatible with T-5 by Raffel et al (2020).

LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale

Tim Dettmers[λ,*] Mike Lewis[†] Younes Belkada[§,∓] Luke Zettlemoyer[†,λ]

University of Washington[λ]
Facebook AI Research[†]
Hugging Face[§]
ENS Paris-Saclay[∓]

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

Table 1: Average and standard deviation of scores achieved by our baseline model and training procedure. For comparison, we also report performance when training on each task from scratch (i.e. without any pre-training) for the same number of steps used to fine-tune the baseline model. All scores in this table (and every table in our paper except Table 14) are reported on the validation sets of each data set.

The inputs and outputs - Further research

Semantic Search Programme

Semantic Network Graph

References

DETTMERS, T., LEWIS, M., BELKADA, Y. & ZETTLEMOYER, L. 2022. LLM. int8 (): 8-bit Matrix Multiplication for Transformers at Scale. *arXiv preprint arXiv:2208.07339*.

RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W. & LIU, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.,* 21**,** 1-67.

TSHITOYAN, V., DAGDELEN, J., WESTON, L., DUNN, A., RONG, Z., KONONOVA, O., PERSSON, K. A., CEDER, G. & JAIN, A. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature,* 571**,** 95-98.

YOSHITAKE, M., SATO, F., KAWANO, H. & TERAOKA, H. 2022. MaterialBERT for natural language processing of materials science texts. *Science and Technology of Advanced Materials: Methods,* 2**,** 372-380.