Title:	Shrinking Knowledge Base Size: Dimension Reduction, Splitting & Filtering
Author:	Vilém Zouhar
Supervisors:	Dietrich Klakow (Saarland University) Gosse Bouma (Groningen University) Marius Mosbach (Saarland University, PhD student advisor)
Defense date:	April 2022
Repository:	https://github.com/zouharvi/kb-shrink
Conference paper:	https://aclanthology.org/2022.spanlp-1.5/

Abstract:
Recently neural network based approaches to knowledge-intensive NLP tasks, such as question answering, started to rely heavily on the combination of neural retrievers and readers. Retrieval is typically performed over a large textual knowledge base which requires significant memory and compute resources, especially when scaled up. On HotpotQA we explore various filtering & splitting criteria. Primarily, we systematically investigate reducing the size of the KB index by means of dimensionality (sparse random projections, PCA, autoencoders) and numerical precision reduction.
Our results show that PCA is an easy solution that requires very little data and is only slightly worse than autoencoders, which are less stable. All methods are sensitive to preand post-processing and data should always be centered and normalized both before and after dimension reduction. Finally, we show that it is possible to combine PCA with using 1bit per dimension. Overall we achieve (1) 100× compression with 75%, and (2) 24× compression with 92% original retrieval performance.