Electronic Theses and Dissertations: A Research Corpus of Scholarly Big Data

- *Curated corpus of over 500,000 full-text documents and metadata*
- *Covering doctoral dissertations, master's and bachelor's theses*
- *From 40+ universities and 2,000+ departments and disciplines*
- *To support teaching, learning, and research in text mining and NLP*

- *Research projects supported:*
  - *Figure and table extraction*
  - *Metadata extraction*
  - *Topic modelling*
  - *Document segmentation*
- *Pretrained language models*
- *Chapter summarization*
- *Reference parsing*
- *Information retrieval*
- *Many more...*

Thanks to the efforts of university libraries, graduate programs, and the open repository movement, millions of electronic theses and dissertations (ETDs) are publicly disseminated online. This enormous volume of scholarship exhibits many interesting characteristics, which make it a valuable corpus for developing new technologies based on computational analysis of academic writing. Digital archives of scholarly publications have been used to support research, but ETDs are unique in that they are much longer than most conference papers and journal articles. ETDs contain novel ideas and findings that contribute significantly to the subject areas of their authors. They often contain useful figures, tables, and equations, as well as extensive literature reviews, bibliographies, and links to other publications. As grey literature, access to ETDs is not controlled by commercial publishers, copyright belongs to the authors, and most are disseminated under permissive copyright licenses.

We have constructed a large document corpus consisting of full-text PDFs and metadata for more than 500,000 ETDs retrieved from university institutional repositories across the United States. The ETD corpus supports research projects conducted by librarians, computer science faculty, undergraduates, master's students, and doctoral students studying natural language processing, information retrieval, bibliometrics, language modeling, and other areas of investigation related to scholarly big data. So far, analysis of the ETD corpus has aided the creation of new models for extracting figures and tables from academic papers, segmenting long documents into chapters and sections, topic modeling algorithms, document classification, summarization algorithms, and improved digital library user interfaces.

William A. Ingram
University Libraries
Virginia Tech

Jian Wu
Dept. of Computer Science
Old Dominion University

Edward A. Fox
Dept. of Computer Science
Virginia Tech

This project was made possible in part by the Institute of Museum and Library Services LG-37-19-0078-19.