Mining ETDs for Trends in Graduate Research

CNI Fall 2020 Virtual Membership Meeting
Bill Ingram, Virginia Tech
November 12, 2020

© 2020 William A. Ingram, licensed under CC BY-SA 4.0
Opening Books and the National Corpus of Graduate Research

IMLS National Leadership Grants for Libraries


Investigating innovative ways machine learning and natural language processing can be applied to the national corpus of electronic theses and dissertations in order to extract knowledge, bibliographic and scientific data, and facilitate its identification, discovery, and reuse.

Research Areas:

1. Document analysis, information extraction
2. Adding value through automatic classification and summarization
3. User services – building better digital libraries
ETD Research Team at Virginia Tech and Old Dominion University

Bill Ingram
*Principle Investigator*
Assistant Dean and Director of IT, University Libraries, Virginia Tech

Dr. Edward A. Fox
*Co-PI*
Professor, Computer Science, Virginia Tech

Dr. Jian Wu
*Co-PI*
Professor, Computer Science, Old Dominion University

Bipasha Banerjee
*Graduate Assistant*
Ph.D. Candidate, Computer Science, Virginia Tech

Muntabir Choudhury
*Graduate Research Assistant*
Ph.D. Student, Computer Science, Old Dominion University

And several students past and present:
Thanks to CNI, my talk led to an introduction and conversation with the chief strategy officer at ProQuest, which led to a collaboration and the opportunity for our team to pilot the new ProQuest TDM Studio.
Outline

- Overview of the TDM Studio
- Research Question
- Describing the Data
- Methodology
- Results
- Discussion
Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any organization.
TDM Studio and the Research Workflow

TDM Studio streamlines the process—making research more efficient & productive.

Select Content
PQ Content, OA Content, Patents, User Uploaded Content

Perform Analysis

PQ Methods

User-Generated Methods

Manage Results & Export Findings
Sample export options (Graph, Table, Feature Set)
- Web interface for selecting data
- Access to material from your library’s subscriptions
- Content rights cleared for TDM
- Build datasets up to 2,000,000 documents
Choose Databases (2 of 118)

- ProQuest Dissertations & Theses Global
  
  ProQuest Dissertations & Theses (PQDT) Global is the world’s most comprehensive collection of dissertations and theses from around the world, offering millions of works from thousands of universities...

- Dissertations & Theses @ Virginia Polytechnic Institute and State University
  
  This database gives access to the dissertations and theses produced by students at your institution.

Database selected
Sample Analysis: Topic Modeling NYTs September 11, 2001

Method: This sample notebook provides an example of topic modeling using non-negative matrix factorization. Topic modeling can be valuable for discovering topics within a dataset of text documents. It can also be valuable for tracking the changes in trends and topics over time. Another popular topic modeling algorithm is LSA or Latent Dirichlet allocation.

Dataset: For this example, we use 30 days of newspaper coverage from the New York Times. This covers the dates of Sept. 1, 2001 to Sept. 30th 2001.

```python
In [1]: # Import Operating System modules
import boto3
s3 = boto3.resource('s3')
from zipfile import ZipFile
from os import listdir
from os.path import isfile, join
import os
from time import time
import pprint as pp
import datetime

# Import BeautifulSoup for XML parsing
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
from collections import Counter, defaultdict
import re

# Import Machine Learning Libraries
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import NMF

# Items for Visualisation
```
Research Question

What can we learn through text and data mining of the ETD corpus about how graduate research topics have evolved, how different topics and disciplines overlap, and how has interdisciplinarity evolved in graduate research?
Data and Feature Set

• Roughly 1.3 million ETDs from 2000-2018
• Full-text XML files with metadata
• Only about 600,000 with department metadata
• Extract features
  • Title
  • Abstract
  • Department
  • Year of publication
• Organize them in batches by years and majors (departments)
• Intuition: top terms in title and abstract will indicate research topic

Distribution of top 20 depts
## Data Sources – Top 20 Institutions

<table>
<thead>
<tr>
<th>School Name</th>
<th>counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Toronto (Canada)</td>
<td>10886</td>
</tr>
<tr>
<td>University of London, University College London, London</td>
<td>10870</td>
</tr>
<tr>
<td>Purdue University</td>
<td>10855</td>
</tr>
<tr>
<td>Walden University</td>
<td>8954</td>
</tr>
<tr>
<td>Capella University</td>
<td>8702</td>
</tr>
<tr>
<td>University of Washington</td>
<td>8391</td>
</tr>
<tr>
<td>University of Maryland, College Park</td>
<td>7948</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>7767</td>
</tr>
<tr>
<td>The University of North Carolina at Chapel Hill</td>
<td>7205</td>
</tr>
<tr>
<td>The University of Manchester (United Kingdom)</td>
<td>6880</td>
</tr>
<tr>
<td>University of California, Los Angeles</td>
<td>6669</td>
</tr>
<tr>
<td>Iowa State University</td>
<td>6477</td>
</tr>
<tr>
<td>Arizona State University</td>
<td>6322</td>
</tr>
<tr>
<td>University of California, San Diego</td>
<td>6227</td>
</tr>
<tr>
<td>University of California, Davis</td>
<td>6048</td>
</tr>
<tr>
<td>Veer Bahadur Singh Purvanchal University, Jaunpur</td>
<td>5777</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>5708</td>
</tr>
<tr>
<td>The University of Arizona</td>
<td>5684</td>
</tr>
<tr>
<td>University of Southern California</td>
<td>5532</td>
</tr>
<tr>
<td>University of Kansas</td>
<td>5321</td>
</tr>
</tbody>
</table>
Distribution of Data Sources
Methodology

- Determine research focus using TF-IDF (term frequency–inverse document frequency) to calculate the most important two- and three-word phrases in the corpus.
- Initially this looked promising, but the method yielded too many irrelevant phrases.

<table>
<thead>
<tr>
<th>user interface</th>
<th>cell cycle</th>
<th>results show</th>
</tr>
</thead>
<tbody>
<tr>
<td>digital libraries</td>
<td>organic matter</td>
<td>gene expression</td>
</tr>
<tr>
<td>web services</td>
<td>eps production</td>
<td>large scale</td>
</tr>
<tr>
<td>digital library</td>
<td>wild type</td>
<td>recent years</td>
</tr>
<tr>
<td>requirements generation</td>
<td>gene expression</td>
<td>results indicate</td>
</tr>
<tr>
<td>web service</td>
<td>land use</td>
<td>also provides</td>
</tr>
<tr>
<td>data assimilation</td>
<td>leaf breakdown</td>
<td>goal research</td>
</tr>
<tr>
<td>software development</td>
<td>corticosterone levels</td>
<td>response time</td>
</tr>
<tr>
<td>software engineering</td>
<td>quorum sensing</td>
<td>future work</td>
</tr>
<tr>
<td>usability engineering</td>
<td>headwater streams</td>
<td>commonly used</td>
</tr>
<tr>
<td>information visualization</td>
<td>xenopus laevis</td>
<td>novel approach</td>
</tr>
<tr>
<td>requirements engineering</td>
<td>cell wall</td>
<td>experimental data</td>
</tr>
<tr>
<td>usability evaluation</td>
<td>mine drainage</td>
<td>life cycle</td>
</tr>
<tr>
<td>data mining</td>
<td>acid mine drainage</td>
<td>high level</td>
</tr>
<tr>
<td>virtual environments</td>
<td>acid mine</td>
<td>multiple sources</td>
</tr>
</tbody>
</table>

Top phrases from computer science and biology

Common irrelevant phrases
Methodology

- **Wikiifier** [1] is a named entity recognition tool to disambiguate terms using Wikipedia.

Methodology

<table>
<thead>
<tr>
<th>Topic Identification</th>
<th>For every document in each department batch, use Wikifier to identify terms from Wikipedia – these terms will represent research topics for that department or major.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot Results</td>
<td>Plot the terms with highest document frequency for a department and time interval.</td>
</tr>
<tr>
<td>Compare Plots</td>
<td>Compare plots of other time intervals for the same department to see the evolution of research topics within that department or major.</td>
</tr>
<tr>
<td>Repeat</td>
<td>Repeat these steps for other departments.</td>
</tr>
<tr>
<td>Plot Multiple Departments</td>
<td>See how research topics are shared across departments and how this interdisciplinarity evolves over time.</td>
</tr>
</tbody>
</table>
Results

The following graphs show the top terms from different disciplines, mainly **computer science** and **biology**, for different years. We also show the intersection of terms from both disciplines.
Computer Science 2006-2009
Economics and Mathematics 2010-2013

monte carlo simulations

- financial market
- brownian motion
- mathematics
- stochastic volatility
- time series
- markov chain
- maximum likelihood
- data analysis
- economics
Economics and Mathematics 2014-2018

monte carlo simulations

time series

machine learning

lower bound

regression analysis

maximum likelihood

continous time

linear regression

game theory

mathematics

markov process

growth model

economics

high-dimensional

data analysis

markov chain

random walk

high-frequency

markov chain monte carlo
Discussion

Back to our research question:

What can we learn through text and data mining of the ETD corpus about how graduate research topics have evolved, how different topics and disciplines overlap, and how has interdisciplinarity evolved in graduate research?

We’ve shown that it is possible to determine the research focus of ETDs using the Illinois Wikifier method for concept disambiguation. Graphing the document frequency of these research topics allows us to visualize the relative importance of these topics within and across academic disciplines and their evolution over time.
Support was made in part by the Institute of Museum and Library Services for grant LG-37-19-0078-198.

This project was supported in part by ProQuest, which provided access to TDM Studio. The university subscribes to the dataset, ProQuest Dissertations & Theses (PQDT).

Thank you!

https://opening-etds.github.io

Mining ETDs for Trends in Graduate Research by William A. Ingram is licensed under CC BY-SA 4.0
For more information about ProQuest’s TDM Studio or ProQuest Dissertations & Theses Global contact:
John.Dillon@proquest.com or Austin.Mclean@proquest.com

Where can I learn more about TDM Studio?
Visit: https://about.proquest.com/products-services/TDM-Studio.html

Where can I learn more about PQDT Global?
Visit: dissertations.com

How can I trial TDM Studio?
Email: TDMStudio@proquest.com

How can I subscribe to PQDT Global?
Inquire at: https://about.proquest.com/contact/contact-sales/