Finding Needles in a Haystack: A Case Study of Text Mining the Corpus of 15 Academic Journals

Tara Baillargeon  
*Marquette University*, tara.baillargeon@marquette.edu

Eric A. Kowalik  
*Marquette University*, eric.kowalik@marquette.edu

Jennifer Cook  
*Marquette University*

Follow this and additional works at: [https://epublications.marquette.edu/lib_fac](https://epublications.marquette.edu/lib_fac)

**Recommended Citation**
Finding Needles in a Haystack: A Case Study of Text Mining the Corpus of 15 Academic Journals

Tara Baillargeon, Eric Kowalik & Jennifer Cook
Marquette University
DLF 2019
Aerial view of Marquette University’s campus
Digital Scholarship Services

- Curriculum support
- Project consultation
- Workshops and programming
- Digital Scholarship Lab – open collaborative space
Project Origins

2016 - Faculty member approached library
  - American Counseling Association (ACA) journal articles
  - Full-text, 17 year span

2017 – Library partners with research team
  - Use of “social class” & “socioeconomic status” in articles
THE LIFE OF A PROJECT*

THIS IS THE BEST IDEA EVER

OK, THIS IS HARDER THAN I THOUGHT

THIS IS GONNA TAKE SOME WORK

THIS SUCKS—AND IT'S BORING

(DARK NIGHT OF THE SOUL)

IT'S DONE AND IT SUCKS, BUT NOT AS BAD AS I THOUGHT

IT WILL BE GOOD TO FINISH BECAUSE I'LL LEARN SOMETHING FOR NEXT TIME

*STOLEN FROM MY FRIEND MAUREEN MCHUGH

From Steal Like an Artist by Austin Kleon
Does Marquette Have It?

Library journal content access inconsistent
- Some through direct subscription
- Some through content aggregators
- Some were open access
Negotiating Access

Access to 6 Wiley Journals

- Negotiated limited access license agreement
- Restricted access for text data mining only
- Content must be deleted after project ended
Project Workflow

Gather Articles

Convert Articles

Generate Reports

Crossref API

PDF Download

Data Janitoring

PDF

XML

Search articles against keywords
Gather Articles

- CrossRef API to download full text PDFs of most journals.
- Bulk download articles from Marquette’s aggregator.
  - Either ProQuest or EBSCOHost.
Convert Articles

- Convert PDF to XML using pdfx.
- Convert text to lower case using dd.
- Use now BBEdit to remove References section.
- Remove stop words, (i.e. class).
Generate Reports

- Python script searched XML files against keyword list
  - If keyword was found, pulled 150 characters before and after keyword
  - Separate report file was created
- Report files uploaded to research team Sharepoint site.
A 17-Year Systematic Content Analysis of Social Class and Socioeconomic Status in Two Counseling Journals

Jennifer M. Cook, Madeline Clark, Katharine Wojcik, Dhanya Nair, Tara Baillargeon & Eric Kowalik

Received 06 Apr 2019, Accepted 06 Jun 2019, Published online: 03 Sep 2019

Download citation https://doi.org/10.1080/21501378.2019.1647409

Abstract

We conducted a qualitative, systematic content analysis of articles from 2 counseling journals (N = 636), Counselor Education and Supervision and Counseling and Values, to understand social class and socioeconomic status (SES) term usage and operationalization. Through PRISMA
Key Takeaways

- Librarians are an important value add.
- Embrace the anxiety, it will turn out okay!
- Need understanding collaborators who #RespectTheProcess.
- Managerial support is vital.
- Document, document, then document some more.
Come Join Us

https://github.com/MarquetteRML/text-mining-script