Using Text Analytics to Identify Shared Themes in Interdisciplinary Research

Jerald Hughes

Information Systems
Robert C. Vackar College of Business and Entrepreneurship

University of Texas Rio Grande Valley

Get a Corpus

- Web of Science Database recommended, or
  - use papers in your literature section
- Fields to include:
  - Web of Science - get ‘Full Record’
  - Journal
  - JOURNAL DISCIPLINE
  - Article Title
  - ABSTRACT TEXT
- Optional fields:
  - Full text
  - Authors
  - Year, vol, #, etc
- Export to Excel

Inspect Corpus

- Delete unneeded fields
- Format as table
- Code Articles for Discipline
  - Simplify Web of Science's lengthy descriptors
  - Create extra column 'Discipline'
- Sort/Filter for your research interests
- Store in convenient location (path)
- Recommend: DON'T save as CSV
  - Commas in full text ruin CSV columns

Load Corpus into Python pandas DataFrame

- Read Excel file
- Inspect header and first row
- Inspect fields you care about
- INSPECT YOUR DISCIPLINE CODE FIELD
- A column:
  myCorpus[‘Article Title’]
- A row:
  myCorpus.iloc[3]
- 2 columns of 3 rows:
  myCorpus[['Article Title','Source Title']].iloc[5:8]

TIP: Ask ChatGPT how to do something you need in Python

Computer Science
Computer Science
Criminology
Criminology
Criminology
Criminology
Criminology
Criminology
Health
Dentistry
Dentistry
Dentistry
Economics
Education
Education
Education

Analysis 1: Search for Terms

- Prepare Abstract text column
  - Remove unneeded terms
    - Present in every abstract
    - Irrelevant to topic
- Change all text to lower case
  - Find both ‘vaccination’ and ‘Vaccination’
- Find all rows with target term in ‘Abstract’ column values

Analysis 2: Prepare results

- Collect relevant columns from all rows
- Identify Disciplines in your subset
- Export results to Excel file

- Iterate: repeat term search on SUBSET of previous search
- Conduct multiple trials to focus on topics

- Create word clouds