DATA TITLE: Term Document Matrices for Higher Education Studies Topic Modeling

PROJECT TITLE: Far Beyond Post-Secondary: Longitudinal analyses of topic and citation networks in the field of Higher Education Studies.

DATA ABSTRACT: Project contains five term document matrices to reproduce the analyses featured in the Review of Higher Education article “Far Beyond Post-Secondary: Longitudinal analyses of topic and citation networks in the field of Higher Education Studies”. Each matrix represents a different corpus (Journal of College Student Development, Journal of Higher Education, Research in Higher Education, Journal of Diversity in Higher Education, Review of Higher Education, and the Association for Higher Education annual meeting) to facilitate topic modeling within a corpus and between venues.

AUTHORS:

Author: Brown, Michael
ORCID: https://orcid.org/0000-0002-2561-7037
Institution: Iowa State University
Email: brownm@iastate.edu

Author: Smith, Rachel
ORCID: https://orcid.org/0000-0003-0392-3735
Institution: Iowa State University
Email: rsmith2@iastate.edu

Corresponding author: Michael Brown

ASSOCIATED PUBLICATIONS:

Smith, R. & **Brown, M.** Far Beyond Post-Secondary: Longitudinal analyses of topic and citation networks in the field of Higher Education Studies. *Review of Higher Education*

COLLECTION INFORMATION:

Time period(s): 1998-2017
Location(s): NA

----- FILE DIRECTORY -----

----- FILE LIST-----

1. DataShare-M.Brown
    1a. ashe tdm.csv (Term document matrix for Association for Higher Education annual meeting presentation titles)
    1b. jcsd tdm.csv (Term document matrix for Journal of College Student Development abstracts)
    1c. jhe tdm.csv (Term document matrix for Journal of Higher Education abstracts)
    1d. rihe tdm.csv (Term document matrix for Research in Higher Education abstracts)
    1e. rhe tdm.csv (Term document matrix for Review of Higher Education abstracts)

----- CODEBOOK -----

Number Of Variables/Columns: 6

Number Of Cases/Rows: Varies

Missing Data Codes: “NA”

----- VARIABLES -----

Name *[variable name]*	Description *[what is it?]*	Values *[units, valid data types, etc.]*
rn	Word Topic pair	Character string
wave1	Frequency of observation of word topic pair in period 1	Numeric whole number
wave2	Frequency of observation of word topic pair in period 2	Numeric whole number
wave3	Frequency of observation of word topic pair in period 3	Numeric whole number
wave4	Frequency of observation of word topic pair in period 4	Numeric whole number
sum	Total sum of observations across all periods in the study	Numeric whole number

----- METHODS AND MATERIALS -----

The data contains word pairs from abstracts of journal articles and titles of conference presentation. Each observation is a two word pair connected by their concurrent appearance in either an abstract or a title.

----- DATA COLLECTION METHODS -----

Data were collected from the text of journal manuscript abstracts and conference title presentations provided by the respective publisher and professional association through their publications.

----- DATA PROCESSING METHODS -----

Each title or abstract was processed using a word tokenization feature in the 'tm' (text mining) package in R.

------- SOFTWARE -------

Name: tm: Text mining Package

Version: 7

System Requirements: C++11

URL: http://tm.r-forge.r-project.org/

Developer: Ingo Feinerer

Additional Notes: Available through CRAN: https://cran.r-project.org/web/packages/tm/index.html

------- LICENSING -------