Using Machine Learning Technology to Analyze the Transcripts of Chat Reference

Yongming Wang
The College of New Jersey

NJALC 2024
January 5th, 2024, Middlesex County College, NJ

…This research utilizes Topic Modeling, an AI and machine learning technology, to extract the major topics of chat reference Q & A… - Proposal abstract

Agenda

- Fundamentals of Topic Modeling and Machine Learning
  1) Basic concepts
  2) General procedures

- The Research Project
  1) Project Description
  2) Data gathering and preparation
  3) Data preprocessing
  4) Model building
  5) Result Explanation

Basic Concepts

- Topic Modeling is one of methods of Natural Language Processing (NLP)

- NLP can find out from a collection of text documents (partial list below)
  - Word frequency: what words are most popular in the text (will be included in our project)
  - Sentiment: positive, negative, neutral (will not be included in our project)
  - Topics: what are some common topics in the text (the major task of our project)

- Topic Modeling usually involves using statistical and mathematical modeling (That is where the machine learning get involved) techniques to extract main topics, themes, or concepts from the corpus of text documents

General procedures of Topic Modeling

1. Gather the text data (In our case, it's the chat transcript)
2. Normalize the text data (a.k.a. Data preprocessing, or data cleaning)

    Text normalization is a key step in natural language processing (NLP) aimed at improving the quality of the text and making it suitable for machines to process.. It involves cleaning and preprocessing text data to make it consistent and usable for different NLP tasks. The process includes a variety of techniques, such as case normalization, punctuation removal, stop word removal, stemming, and lemmatization, etc. Basically, to remove those symbols and words in the text that don't have much meaning, or are not of much value to our project at hand.

3. Build model
4. Explain result (human interpretation)

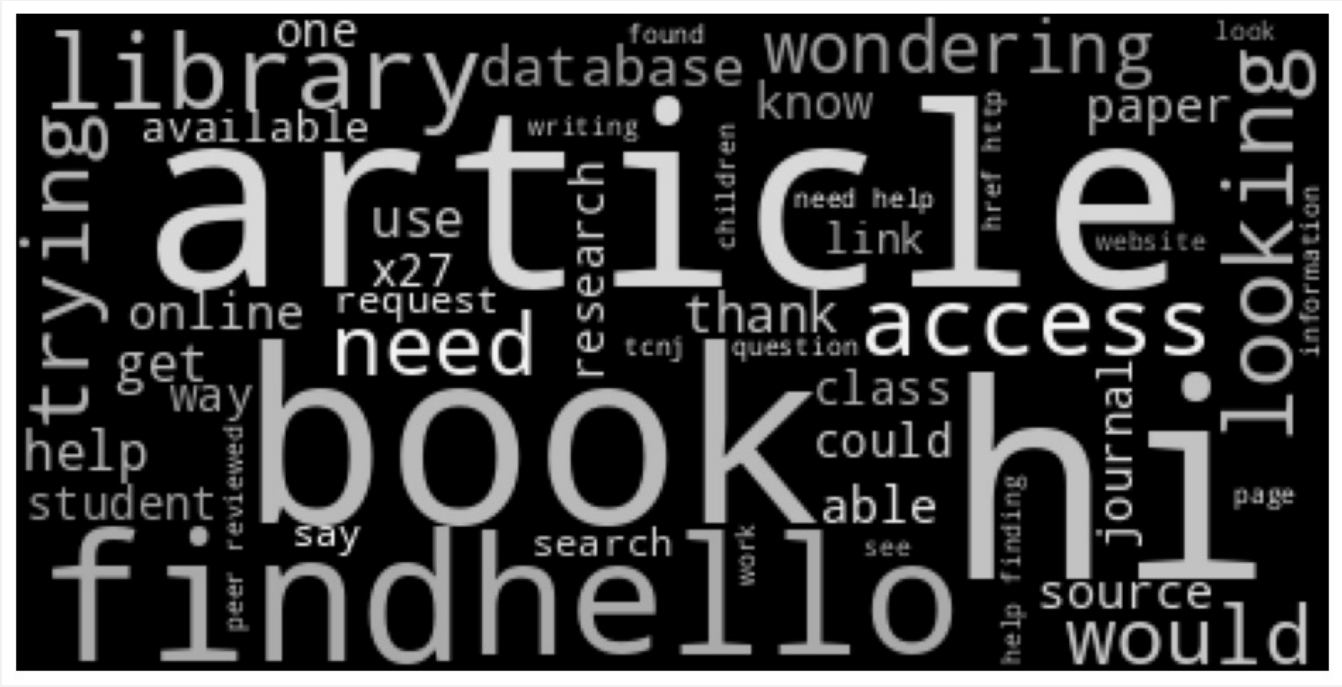
Project Description

- Data --- Transcripts of 10 years chat reference transactions (2014 – 2023), from LibAnswers
- Research Question --- What are the common topics in the questions the chat users ask, and what are the common topics in the chat transcripts?
- Method and Goal --- Using the text data available, we are going to use one of the most popular Topic Modeling models, Latent Dirichlet Allocation (LDA), to find out some common topics.
- Computer Language used--- Python
- Development Tool used --- Jupyter Notebook

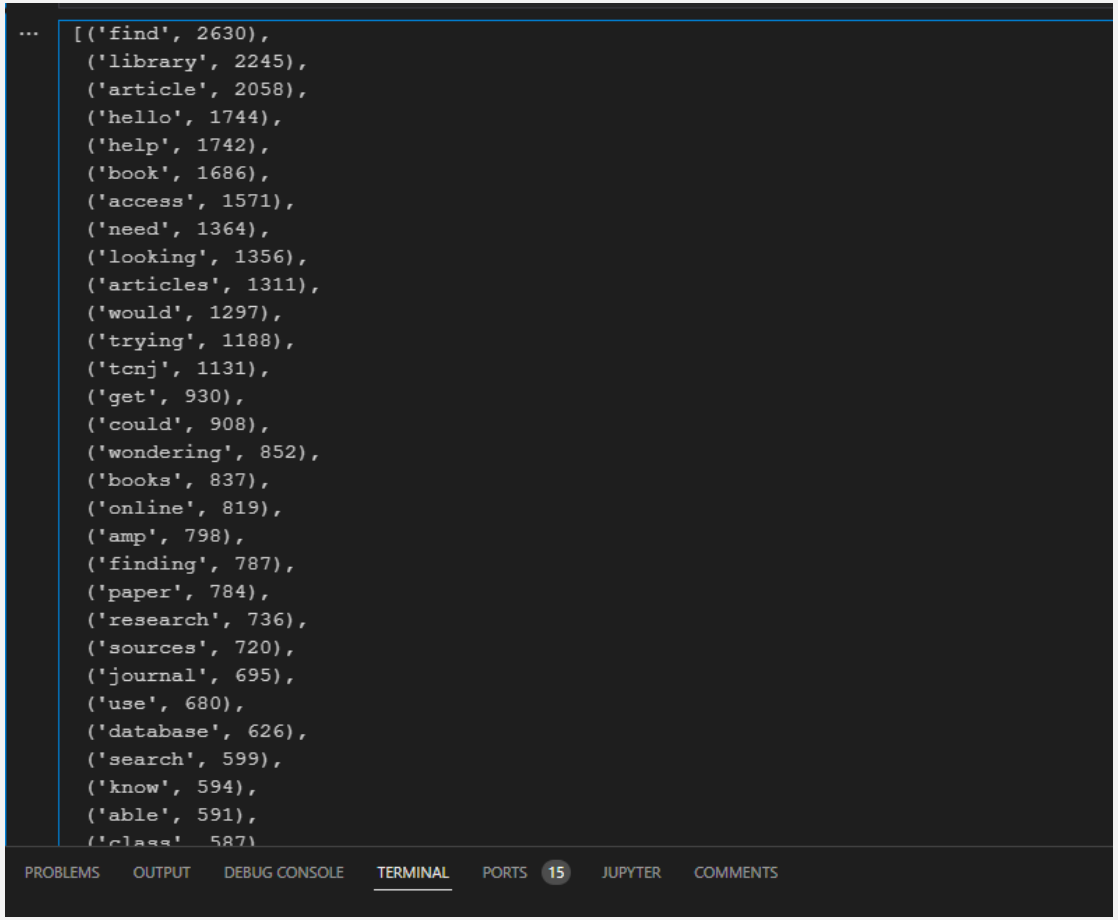
Gather data

- Download the initial questions (File one) and transcripts (File two) for the time period of 2014 to 2023 from SpringShare. (Can only download one year each time.)
- Copy all ten years initial questions into File one
- Copy all ten years transcripts into File two, then copy the content of file one (initial questions) into file two.
- Before going to text normalization, let’s do the word frequency analysis (next two slides show the result)

Word Cloud from File One (Initial Questions)



50 most frequent words from File one (Initial Questions)



Normalize text data

- Remove numbers (they don’t contribute to the value of our topic modeling project)
- Lowercase each word
- Tokenize (Split the text into words and sentences for topic modeling)
- Stem each word (to return the varieties of words to its base form. For example: after stemming, “articles” will become “article”)
- Remove stop words (such as a, an, the, … both standard list and our customized list)
- Remove short words less than 3 letters in length (because they usually don’t have much meanings, for example: hi)

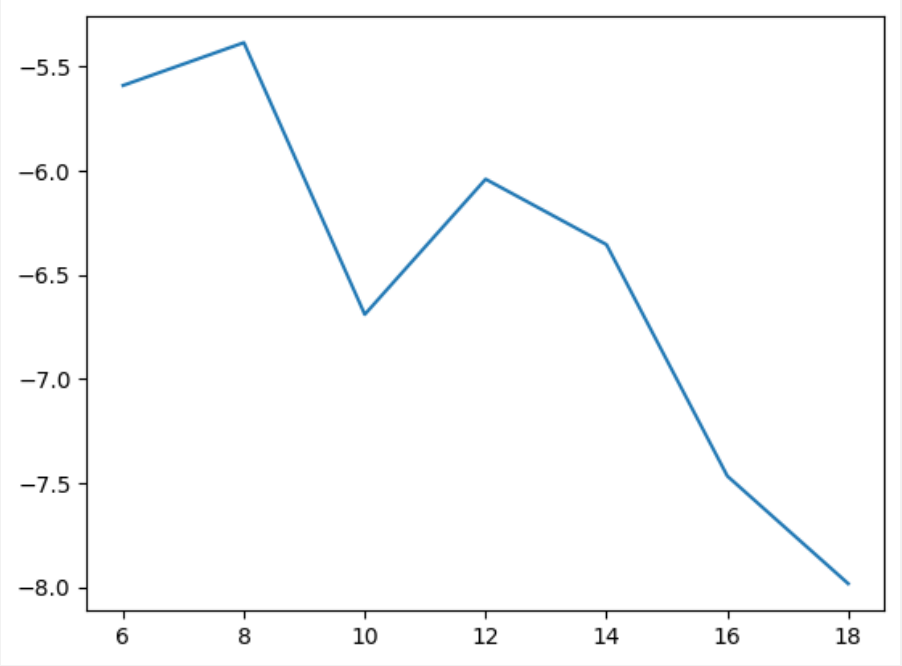
Build the Model

- Create dictionary and corpus to be fed to LDA model
  - Dictionary contains unique words in a file
  - Corpus is a list of documents (sentences in our case)

- Find optimal value for the number of topics (range 6 – 18) by calculating the coherence value. The optimal number of topics has the highest coherence

- Run the LDA model

Optimal value of topics for File one (initial questions(8) (Y axis is the coherence value)



Run LDA model and Print the results (Keywords that are relevant to certain topic)

- Result from File one (Initial questions)
- [(0,
- '0.115*"book" + 0.066*"library" + 0.023*"request" + 0.018*"loan" + 0.014*"check" + 0.013*"get" + 0.011*"number" + 0.011*"pick" + 0.011*"hold" + 0.011*"wondering"'),
- (1,
- '0.078*"find" + 0.063*"article" + 0.041*"source" + 0.034*"help" + 0.031*"looking" + 0.028*"finding" + 0.018*"trying" + 0.017*"database" + 0.014*"information" + 0.013*"trouble"'),
- (2,
- '0.027*"floor" + 0.021*"study" + 0.014*"room" + 0.013*"doi" + 0.013*"com" + 0.012*"film" + 0.011*"talking" + 0.011*"talk" + 0.010*"loud" + 0.009*"table"'),
- (3,
- '0.025*"tcnj" + 0.018*"see" + 0.017*"http" + 0.016*"health" + 0.015*"blank" + 0.014*"target" + 0.013*"rel" + 0.013*"noopener" + 0.013*"noreferrer" + 0.013*"edu"'),
- (4,
- '0.046*"http" + 0.027*"paper" + 0.020*"href" + 0.020*"writing" + 0.018*"good" + 0.018*"www" + 0.015*"org" + 0.014*"research" + 0.013*"working" + 0.012*"morning"'),
- (5,
- '0.064*"access" + 0.052*"article" + 0.033*"library" + 0.030*"get" + 0.027*"trying" + 0.026*"online" + 0.023*"tcnj" + 0.020*"journal" + 0.018*"say" + 0.016*"text"'),
- (6,
- '0.029*"new" + 0.021*"com" + 0.020*"disconnected" + 0.014*"search" + 0.012*"time" + 0.011*"think" + 0.010*"state" + 0.010*"tcnj" + 0.009*"keep" + 0.009*"primo"'),
- (7,
- '0.050*"help" + 0.026*"peer" + 0.024*"reviewed" + 0.021*"question" + 0.020*"child" + 0.020*"book" + 0.017*"article" + 0.016*"looking" + 0.012*"student" + 0.012*"wondering"')]

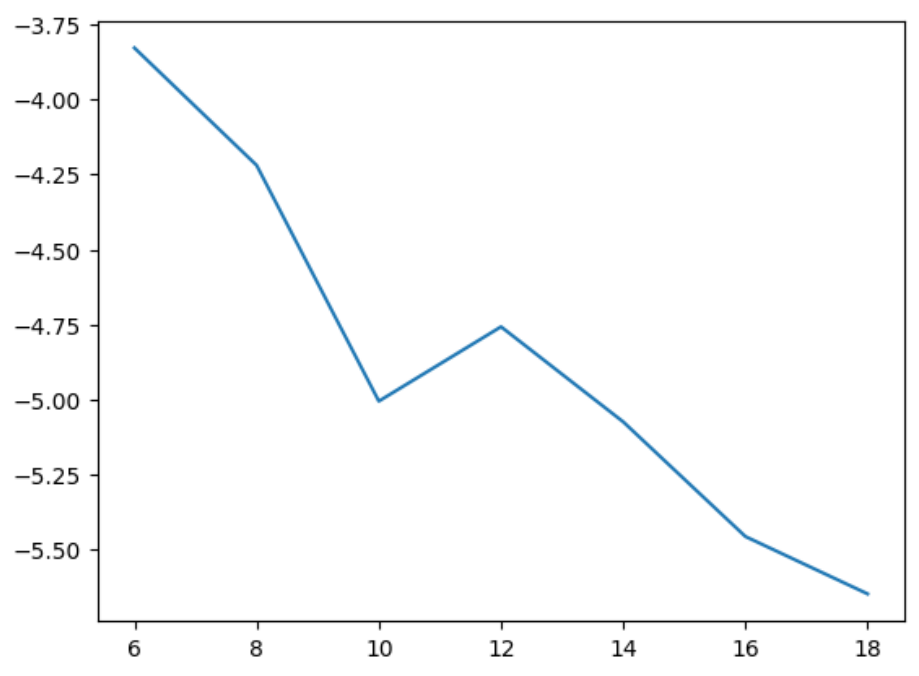
Explain the result

Topics from File one (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	book, library, request, loan, check, get, number, pick, hold, wondering	physical book access
T2	find, article, source, help, looking, finding, trying, database, information, trouble	article access
T3	floor, study, room, doi, com, film, talking, talk, loud, table	noise complain
T4	tcnj, see, http, rel, blank, target, rel, noopener, noreferrer, edu	access by http link*
T5	http, paper, href, writing, good, www, org, research, working, morning	general reference
T6	access, login, library, get, trying, online, tcnj, journal, say, article	off-campus access
T7	new, com, disconnected, search, time, think, state, tcnj, keep, sorry	chat disconnected(?)
T8	help, peer, reviewed, question, child, book, article, looking, student, wondering	peer reviewed articles

*I'm trying to access this article: https://tcnj.primo.exlibrisgroup.com/permalink/01COLLNJ_INST/12od3b9/proquest237310839

Optimal value for file two (transcripts + initial questions) (6)



Run the LDA model and print the results (2)

- Result from File two (transcripts + initial questions)
- [(0,
- '0.054*"patron" + 0.039*"article" + 0.037*"book" + 0.037*"find" + 0.021*"looking" + 0.017*"library" + 0.014*"paper" + 0.014*"trying" + 0.010*"way" + 0.009*"peer"'),
- (1,
- '0.119*"tcnj" + 0.119*"http" + 0.053*"href" + 0.052*"edu" + 0.049*"blank" + 0.049*"target" + 0.040*"library" + 0.038*"database" + 0.030*"search" + 0.026*"inst"'),
- (2,
- '0.040*"search" + 0.037*"source" + 0.023*"article" + 0.021*"class" + 0.017*"database" + 0.013*"text" + 0.012*"ebsco" + 0.012*"page" + 0.012*"citation" + 0.011*"full"'),
- (3,
- '0.145*"help" + 0.034*"good" + 0.029*"finding" + 0.029*"question" + 0.020*"trouble" + 0.018*“class" + 0.014*"anything" + 0.014*"doi" + 0.013*"student" + 0.013*"professor"'),
- (4,
- '0.063*"access" + 0.038*"library" + 0.029*"wondering" + 0.029*"article" + 0.022*"book" + 0.020*"request" + 0.019*"online" + 0.017*"journal" + 0.016*"loan" + 0.015*"org"'),
- (5,
- '0.057*"com" + 0.027*"primo" + 0.026*"exlibrisgroup" + 0.022*"login" + 0.020*"check" + 0.019*"let" + 0.018*"chat" + 0.018*"new" + 0.017*"get" + 0.016*"take"')]

Topics from File two (Transcripts + Initial Question)

ID	Keywords (Machine generated)	Topic (human interpretation)
T1	patron, article, book, find, looking, library, paper, trying, way, peer	general reference
T2	tcnj, http, href, edu, blank, target, library, database, search, inst	access by http link
T3	search, source, article, class, database, text, ebsco, page, citation, full	citation help
T4	help, good, finding, class, question, trouble, anything, doi, student, professor	class assignment
T5	access, library, wondering, article, book, request, online, journal, loan, org	loan request
T6	com, primo, exlibrisgroup, login, check, let, chat, new, get, take	access to our Discovery Tool

- Are there any trends in the questions people asked?

- Let’s find out by running each year’s initial questions for the last ten years.

Topics – 2014 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	find, information, wondering, search, topic, database, issue, tell, music, people	database usage
T2	article, access, database, question, online, available, use, full, text, find	article and database access
T3	book, find, paper, research, way, library, chat, look, tcnj, see	find physical book
T4	library, article, find, finding, journal, open, trouble, access, time, copy	open access resources
T5	wondering, article, http, cite, find, research, help, www, source, journal	citation help
T6	find, paper, looking, trying, text, full, library, source, publication, class	find full text article
T7	test, chat, article, journal, tcnj, access, question, way, book, pdf	pdf article
T8	library, article, book, tcnj, looking, use, help, search, email, report	general reference
T9	help, source, book, library, find, question, finding, take, primary, article	primary source
T10	article, book, possible, find, trying, looking, trouble, student, journal, access	general reference

Topics – 2015 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	help, book, looking, library, new, use, suggestion, print, find, time	find physical book
T2	book, library, tcnj, student, article, quot*, looking, journal, access, number	general reference
T3	book, library, loan, request, online, article, access, interlibrary, find, chat	interlibrary loan
T4	find, article, database, source, quot, access, looking, library, tcnj, paper	find article
T5	article, library, find, access, help, database, trying, use, search, get	database usage
T6	find, library, class, looking, help, article, book, wondering, journal, access	general reference

*" = “

Topics – 2016 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	library, article, looking, tcnj, journal, wondering, question, source, access, database	article and database
T2	article, tcnj, find, library, edu, journal, database, available, trying, source	artcle and database availability
T3	find, article, library, book, trying, quot, database, looking, journal, tcnj	general reference
T4	quot, com, book, loan, interlibrary, http, target, blank, find, research	interlibrary loan
T5	book, access, help, library, research, database, get, still, way, class	book and database
T6	article, http, use, find, quot, looking, href, database, library, paper	article and database

Topics – 2017 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	access, help, library, get, find, article, book, trying, disconnected, cannot	chat disconnected
T2	tcnj, floor, journal, library, people, loud, access, article, book, third	noise complain (3rd fl.)
T3	article, someone, floor, room, study, help, tell, quiet, library, student	noise complain (study room)
T4	find, book, article, library, source, online, trying, database, looking, class	book, article, database
T5	article, library, book, find, looking, finding, database, use, source, search	book, article, database
T6	database, paper, http, research, find, writing, something, history, look, href	database usage

Topics – 2018 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	room, access, study, article, loud, floor, library, find, trying, group	noise complain
T2	article, help, find, paper, trying, finding, database, access, get, full	find article and database
T3	book, computer, use, find, looking, get, technology, way, issue, child	?
T4	library, article, book, tcnj, loan, find, class, see, student, wondering	interlibrary loan
T5	find, floor, article, journal, table, http, quiet, tell, someone, talking	noise complain
T6	database, source, looking, help, use, access, finding, information, tcnj, reference	database usage

Topics – 2019 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	book, article, http, com, access, href, library, copy, looking, www	resource by http link
T2	article, find, access, tcnj, library, student, journal, found, database, looking	find article
T3	book, find, library, online, trying, article, available, get, log, page	resource availability
T4	floor, library, article, book, looking, journal, table, loud, girl, talking	noise complain (girl)
T5	book, library, request, loan, use, look, wondering, check, interlibrary, article	interlibrary loan
T6	help, find, source, database, article, wondering, good, paper, finding, book	find resources in database

Topics – 2020 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	book, library, pick, article, search, hold, request, tcnj, available, student	curbside pickup
T2	access, article, tcnj, trying, online, get, library, journal, database, find	article access
T3	article, book, find, trying, text, online, help, get, cite, looking	citation help
T4	http, href, blank, target, noreferrer, noopener, rel, com, tcnj, org	access by http link
T5	article, loan, looking, interlibrary, find, research, look, library, wondering, gender	interlibrary loan
T6	help, find, source, article, looking, peer, reviewed, finding, paper, trying	peer reviewed article

Topics – 2021 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	article, looking, find, help, trying, question, journal, adult, school, citation	find resource
T2	http, peer, href, reviewed, article, www, help, student, issue, exproxy	peer reviewed article
T3	find, book, article, library, database, trying, finding, topic, source, loan	general reference
T4	access, article, book, library, online, available, trying, way, get, tcnj	resource availability
T5	help, find, source, article, finding, com, looking, paper, disconnected, library	disconnected
T6	http, blank, ref, target, noopener, noreferrer, pdf, com, www, film	access by http link (film)
T7	rft, pick, hold, put, paper, request, working, book, primo, write	hold and pick up
T8	tcnj, utm, library, make, course, dvd, work, page, login, time,	DVD resource

Topics – 2022 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	com, article, argument, idea, book, south, performance, teaching, thought, anything	?
T2	search, looking, good, find, class, article, source, impact, floor, paper	find article
T3	library, book, get, paper, wondering, request, possible, time, com, loan	interlibrary loan
T4	small, primo, business, service, jersey, cnn, new, library, social, current	?
T5	http, article, access, tcnj, trying, www, find, book, college, paper	find resource
T6	help, finding, article, source, find, looking, trouble, medium, people, look	find article
T7	article, access, book, library, help, find, peer, reviewed, online, research	peer reviewed article
T8	book, find, source, trying, looking, hold, help, renew, history, cite	find book

Topics – 2023 (Initial Question)

ID	Keywords (Machine Generated)	Topic (Human Interpretation)
T1	help, book, trenton, paper, writing, apa, find, information, wondering, cancer	citation help
T2	article, help, find, finding, source, peer, reviewed, looking, woman, book	peer reviewed article
T3	find, book, source, topic, search, class, com, paper, get, article	find book
T4	article, http, access, tcnj, www, library, database, trying, get, research	access article and database
T5	library, book, room, get, sociology, possible, study, student, find, looking	book source
T6	library, com, book, cancer, hero, full, utm*, confused, way, access	find resource

*Do we have access to this article? https://www.cnbc.com/2022/12/21/survey-shows-a-costco-membership-hike-would-face-little-resistance-.html?utm_source=ground.news&utm_medium=referral

General Trends in last ten years?

- More article seeking, less book request
- Less individual database question after we switched to Alma/Primo in 2019
- Less noise complains (most happened in 2017, 2018, and 2019)
- Interlibrary loan requests are steady

Thank you

Questions?

- Yongming Wang, Systems Librarian, The College of New Jersey (TCNJ)
- wangyo@tcnj.edu