Automated Chat Transcript Analysis Using Topic Modeling for Library Reference Services

Xiaoju Chen

Huajin Wang

Carnegie Mellon University Libraries

asis&t

19 - 23 OCTOBER 2019

CROWN MELBOURNE, AUSTRALIA

82ND ANNUAL MEETING

Motivation

- Web-based reference services are valuable
  - Chat - ideal for internet Savvy generation, busy faculty members who never come to the library

- How to provide better chat service while optimizing staff time?
  - Need to know what questions are asked

- We can make a data-driven decision based on chat transcripts

- Traditionally – open coding or mixed methods
  - Slow and not scalable

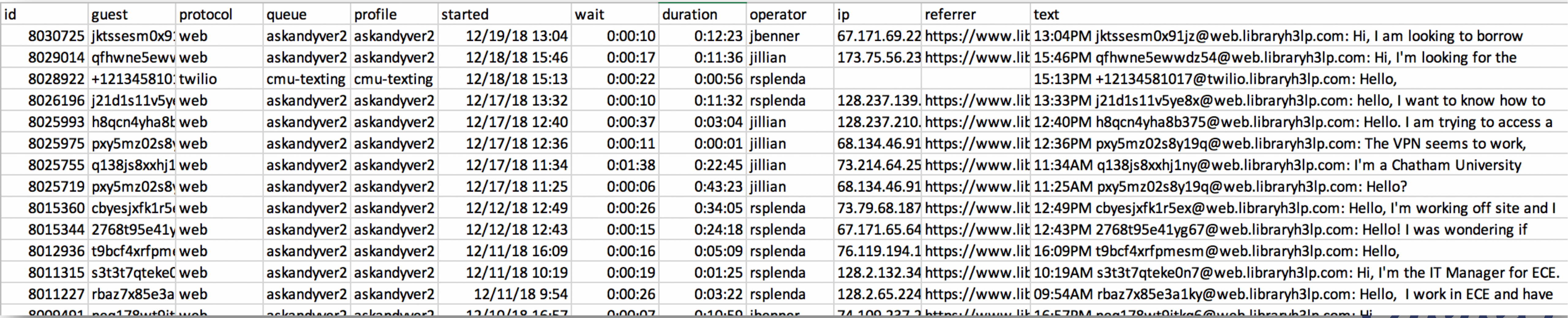
- Let try some machine learning!

Research Questions

- What type of questions are being asked by patrons?
- How frequent each type of question are being asked?
- How do we use this information to optimize chat services?

Dataset

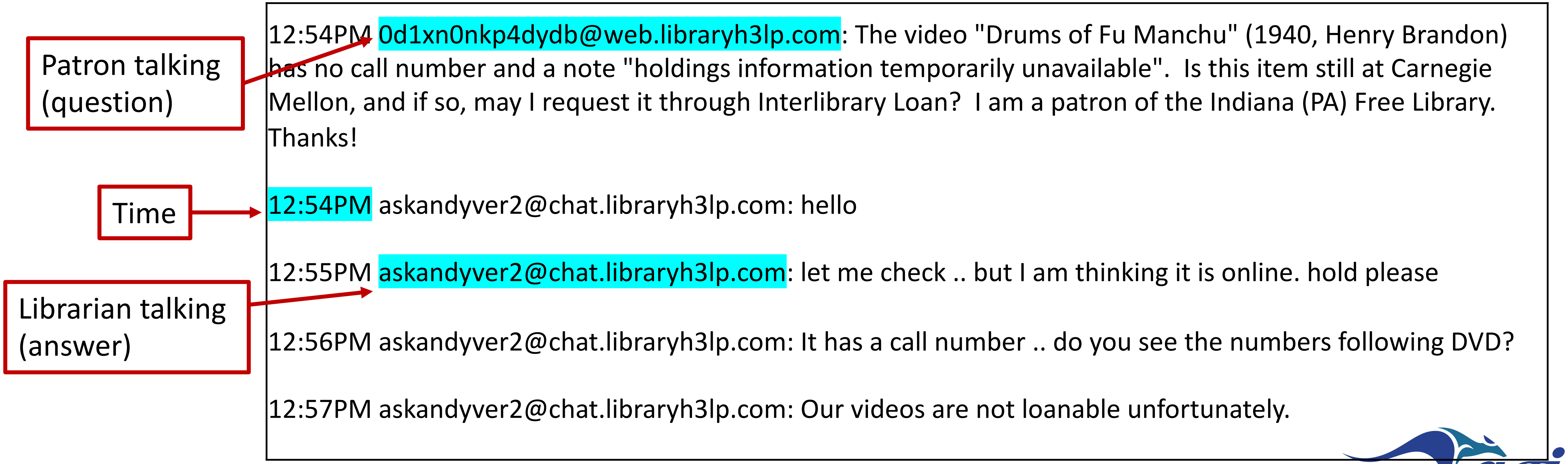
- Data downloaded from *LibraryH3lp* as .csv file
- Date range: January 1, 2013 to December 31, 2018
- Total number of records: 5609
- Total number of words: ~1.3 million



id	guest	protocol	queue	profile	started	wait	duration	operator	ip	referrer	text
8030725	jktssesm0x9	web	askandyver2	askandyver2	12/19/18 13:04	0:00:10	0:12:23	jbenner	67.171.69.22	https://www.lib	13:04PM jktssesm0x91jz@web.libraryh3lp.com: Hi, I am looking to borrow
8029014	qfhwne5ewv	web	askandyver2	askandyver2	12/18/18 15:46	0:00:17	0:11:36	jillian	173.75.56.23	https://www.lib	15:46PM qfhwne5ewwdz54@web.libraryh3lp.com: Hi, I'm looking for the
8028922	+121345810	twilio	cmu-texting	cmu-texting	12/18/18 15:13	0:00:22	0:00:56	rsplenda			15:13PM +12134581017@twilio.libraryh3lp.com: Hello,
8026196	j21d1s11v5y	web	askandyver2	askandyver2	12/17/18 13:32	0:00:10	0:11:32	rsplenda	128.237.139.	https://www.lib	13:33PM j21d1s11v5ye8x@web.libraryh3lp.com: hello, I want to know how to
8025993	h8qcn4yha8b	web	askandyver2	askandyver2	12/17/18 12:40	0:00:37	0:03:04	jillian	128.237.210.	https://www.lib	12:40PM h8qcn4yha8b375@web.libraryh3lp.com: Hello. I am trying to access a
8025975	pxy5mz02s8	web	askandyver2	askandyver2	12/17/18 12:36	0:00:11	0:00:01	jillian	68.134.46.91	https://www.lib	12:36PM pxy5mz02s8y19q@web.libraryh3lp.com: The VPN seems to work,
8025755	q138js8xxhj1	web	askandyver2	askandyver2	12/17/18 11:34	0:01:38	0:22:45	jillian	73.214.64.25	https://www.lib	11:34AM q138js8xxhj1ny@web.libraryh3lp.com: I'm a Chatham University
8025719	pxy5mz02s8	web	askandyver2	askandyver2	12/17/18 11:25	0:00:06	0:43:23	jillian	68.134.46.91	https://www.lib	11:25AM pxy5mz02s8y19q@web.libraryh3lp.com: Hello?
8015360	cbyesjxfk1r5	web	askandyver2	askandyver2	12/12/18 12:49	0:00:26	0:34:05	rsplenda	73.79.68.187	https://www.lib	12:49PM cbyesjxfk1r5ex@web.libraryh3lp.com: Hello, I'm working off site and I
8015344	2768t95e41y	web	askandyver2	askandyver2	12/12/18 12:43	0:00:15	0:24:18	rsplenda	67.171.65.64	https://www.lib	12:43PM 2768t95e41yg67@web.libraryh3lp.com: Hello! I was wondering if
8012936	t9bcf4xrfpm	web	askandyver2	askandyver2	12/11/18 16:09	0:00:16	0:05:09	rsplenda	76.119.194.1	https://www.lib	16:09PM t9bcf4xrfpmesm@web.libraryh3lp.com: Hello,
8011315	s3t3t7qteke0	web	askandyver2	askandyver2	12/11/18 10:19	0:00:19	0:01:25	rsplenda	128.2.132.34	https://www.lib	10:19AM s3t3t7qteke0n7@web.libraryh3lp.com: Hi, I'm the IT Manager for ECE.
8011227	rbaz7x85e3a	web	askandyver2	askandyver2	12/11/18 9:54	0:00:26	0:03:22	rsplenda	128.2.65.224	https://www.lib	09:54AM rbaz7x85e3a1ky@web.libraryh3lp.com: Hello, I work in ECE and have
8009491	nog178wt9it	web	askandyver2	askandyver2	12/10/18 16:57	0:00:07	0:10:59	jbenner	74.109.237.2	https://www.lib	16:57PM nog178wt9itkg6@web.libraryh3lp.com: Hi

Data Cleaning and Preprocessing

- Extract questions and answers
- Each interaction = 1 document



Patron talking (question)

Time

Librarian talking (answer)

12:54PM 0d1xn0nkp4dydb@web.libraryh3lp.com: The video "Drums of Fu Manchu" (1940, Henry Brandon) has no call number and a note "holdings information temporarily unavailable". Is this item still at Carnegie Mellon, and if so, may I request it through Interlibrary Loan? I am a patron of the Indiana (PA) Free Library. Thanks!

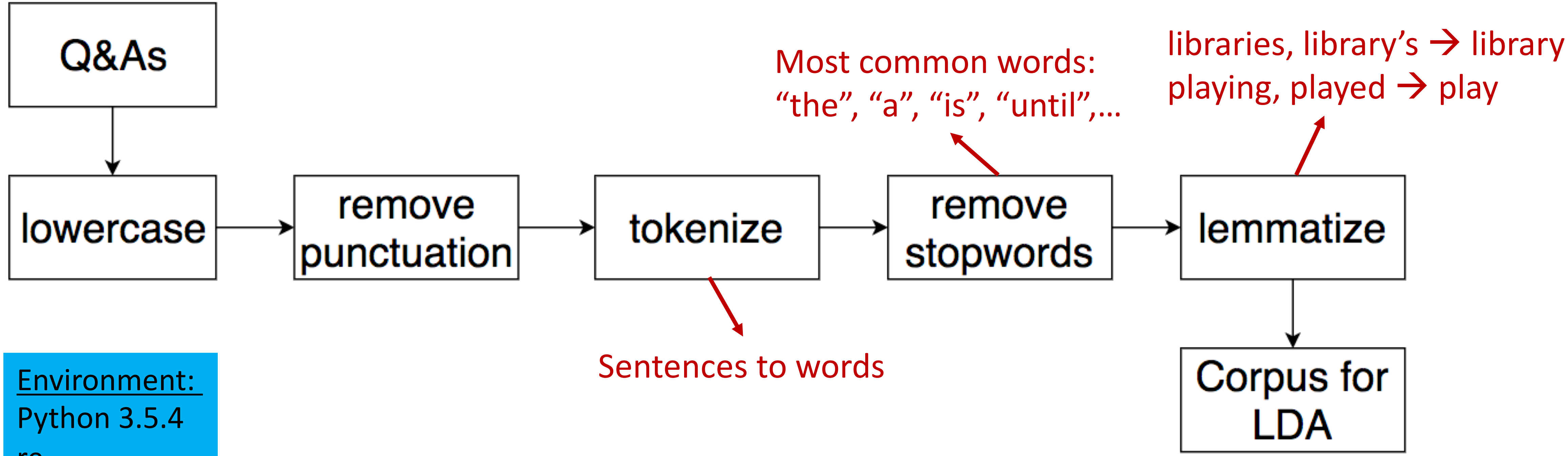
12:54PM askandyver2@chat.libraryh3lp.com: hello

12:55PM askandyver2@chat.libraryh3lp.com: let me check .. but I am thinking it is online. hold please

12:56PM askandyver2@chat.libraryh3lp.com: It has a call number .. do you see the numbers following DVD?

12:57PM askandyver2@chat.libraryh3lp.com: Our videos are not loanable unfortunately.

Data Cleaning and Preprocessing



Q&As

lowercase

remove punctuation

tokenize

remove stopwords

lemmatize

Corpus for LDA

Sentences to words

Most common words:
“the”, “a”, “is”, “until”,…

libraries, library’s → library
playing, played → play

<u>Environment:</u>
Python 3.5.4
re
pandas
NLTK Toolkit

Customization of stop words

Standard English stop words in NLTK: 174
Added customized stop words: 86

['actually', 'alright', 'also', 'and', 'appreciate', 'awesome', 'believe', 'best', 'bye', 'can', 'cmu', 'cool', 'done', 'dont', 'else', 'even', 'fairly', 'glad', 'good', 'great', 'haha', 'happy', 'hello', 'helpful', 'hey', 'hi', 'hmm', 'hopefully', 'however', 'im', 'just', 'let', 'lol', 'lot', 'luck', 'many', 'may', 'maybe', 'might', 'minute', 'moment', 'much', 'nice', 'nope', 'no', 'now', 'often', 'ok', 'okay', 'one', 'ones', 'or', 'perhaps', 'please', 'pleasure', 'plz', 'possibly', 'probably', 'really', 'right', 'ryan', 'seems', 'sorry', 'sure', 'thank', 'thanks', 'thats', 'think', 'though', 'thought', 'thx', 'today', 'tried', 'trying', 'unfortunately', 'welcome', 'well', 'will', 'wonderful', 'wondering', 'worries', 'yea', 'yeah', 'yep', 'yes', 'yet']

Latent Dirichlet allocation (LDA) Topic Modeling

LDA:
A generative statistical model

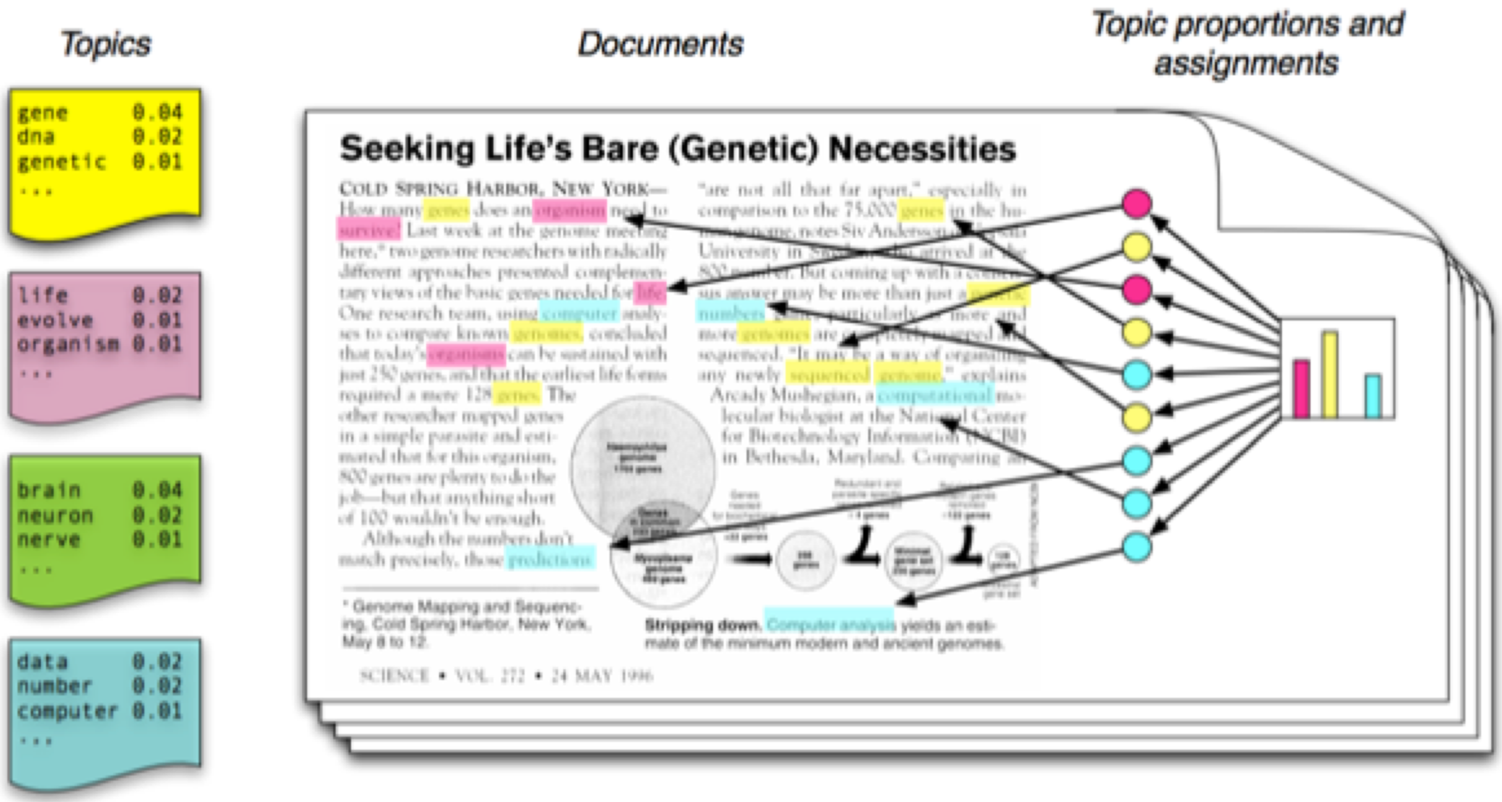

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

LDA modeling with Gensim

- Training: 20 passes; random_state: 50
- Ran model for k = [2, 20]
- Decide best k
  - Evaluate perplexity scores and coherence scores for each k to decide best k
  - Evaluate human interpretability for each k
  - Topic stability through multiple runs

Environment:
Python Gensim library
(version 3.4.0)

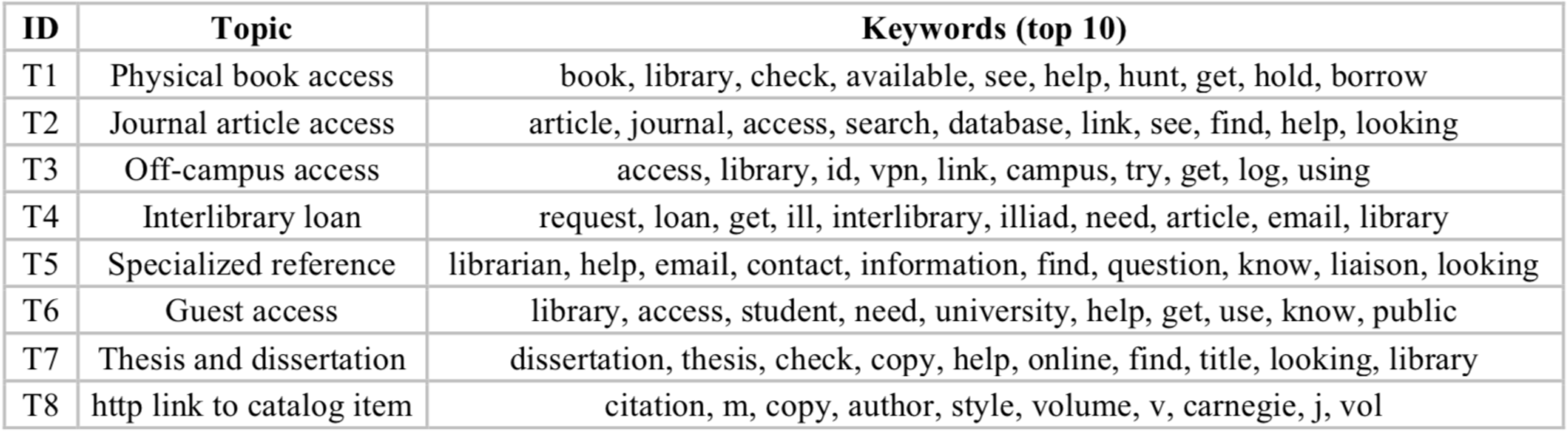
➔ k = 8, perplexity score = -7.38, coherence score = 0.44

Model Output

```
[(0, '0.018*"dissertation" + 0.018*"thesis" + 0.012*"check" + 0.011*"copy" + 0.010*"help" + 0.009*"online"
+ 0.008*"find" + 0.007*"title" + 0.006*"looking" + 0.006*"library"'),
(1, '0.068*"book" + 0.040*"library" + 0.017*"check" + 0.013*"available" + 0.012*"see" + 0.011*"help" +
0.011*"hunt" + 0.011*"get" + 0.010*"hold" + 0.010*"borrow”’),
(2, '0.026*"request" + 0.023*"loan" + 0.022*"get" + 0.022*"ill" + 0.022*"interlibrary" + 0.015*"illiad" +
0.014*"need" + 0.014*"article" + 0.014*"email" + 0.011*"library”’),
(3, '0.030*"article" + 0.023*"journal" + 0.020*"access" + 0.019*"search" + 0.017*"database" + 0.015*"link"
+ 0.014*"see" + 0.013*"find" + 0.012*"help" + 0.011*"looking”’),
(4, '0.039*"library" + 0.020*"access" + 0.013*"student" + 0.009*"need" + 0.009*"university" + 0.009*"help"
+ 0.008*"get" + 0.008*"use" + 0.007*"know" + 0.007*"public”’),
(5, '0.006*"citation" + 0.004*"m" + 0.003*"copy" + 0.003*"author" + 0.003*"style" + 0.003*"volume" +
0.003*"v" + 0.003*"carnegie" + 0.003*"j" + 0.003*"vol”’),
(6, '0.023*"access" + 0.020*"library" + 0.018*"id" + 0.016*"vpn" + 0.016*"link" + 0.016*"campus" +
0.014*"try" + 0.012*"get" + 0.011*"log" + 0.011*"using”’),
(7, '0.022*"librarian" + 0.017*"help" + 0.017*"email" + 0.014*"contact" + 0.012*"information" +
0.009*"find" + 0.009*"question" + 0.008*"know" + 0.008*"liaison" + 0.008*"looking"')]
```

Topic Interpretation

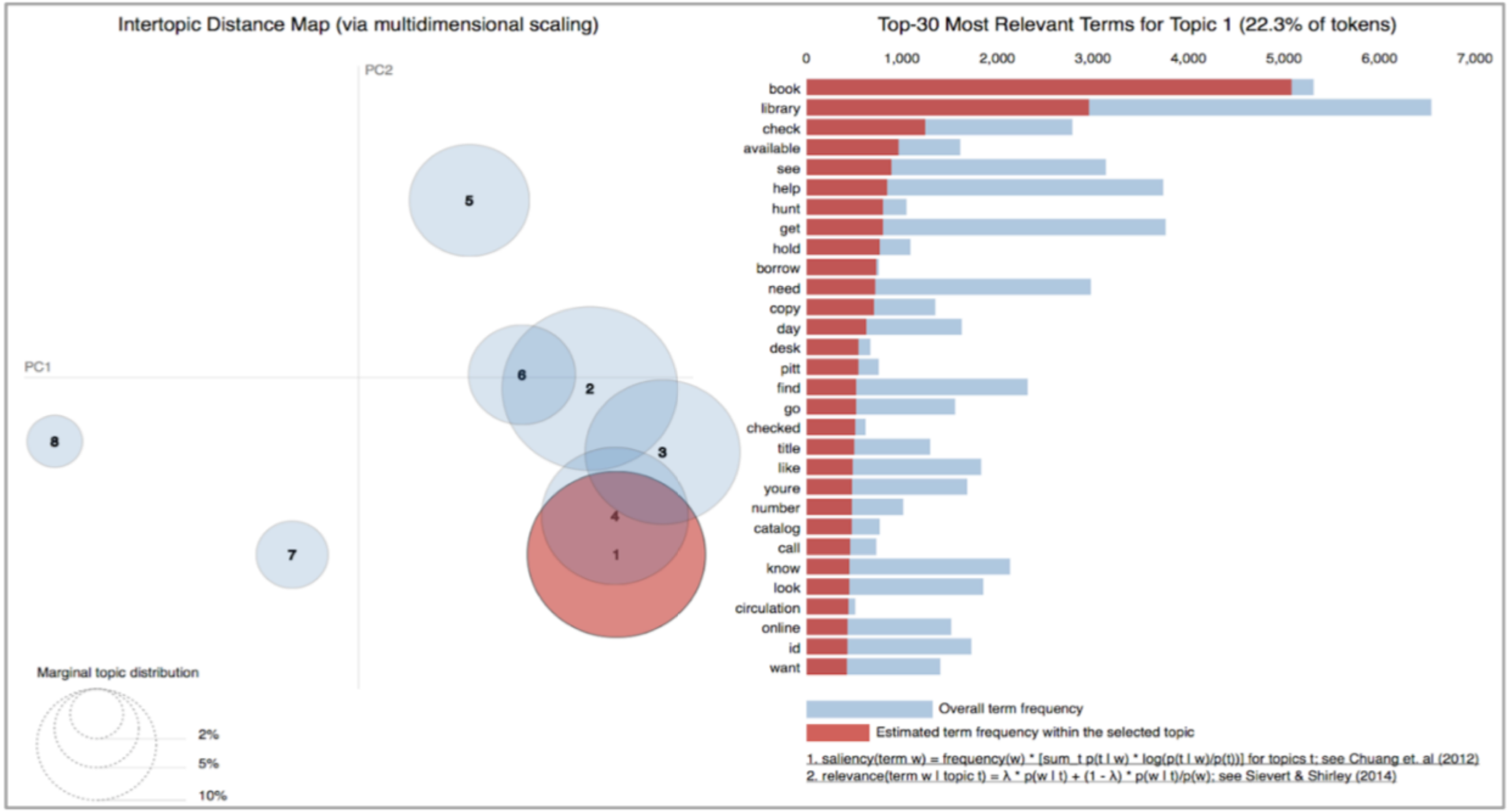
Topic prevalence



ID	Topic	Keywords (top 10)
T1	Physical book access	book, library, check, available, see, help, hunt, get, hold, borrow
T2	Journal article access	article, journal, access, search, database, link, see, find, help, looking
T3	Off-campus access	access, library, id, vpn, link, campus, try, get, log, using
T4	Interlibrary loan	request, loan, get, ill, interlibrary, illiad, need, article, email, library
T5	Specialized reference	librarian, help, email, contact, information, find, question, know, liaison, looking
T6	Guest access	library, access, student, need, university, help, get, use, know, public
T7	Thesis and dissertation	dissertation, thesis, check, copy, help, online, find, title, looking, library
T8	http link to catalog item	citation, m, copy, author, style, volume, v, carnegie, j, vol

Table 1. The 8 topics and top keywords associated with each topic discovered by the LDA model. Names of the topics are generated based on human interpretation.

Visualization with pyLDAvis



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 1. Screenshot of interactive visualization output from pyLDAvis. Left: distance map created based on keywords occurrence. Each cluster represents a topic generated by the LDA model. The index number of each cluster corresponds to the topic ID in Table 1. Right: distribution of the top 30 most relevant terms among topics. Red: the current topic; blue: other topics.

Environment:
pyLDAvis library
(version 2.1.2)

Validation with VOSviewer

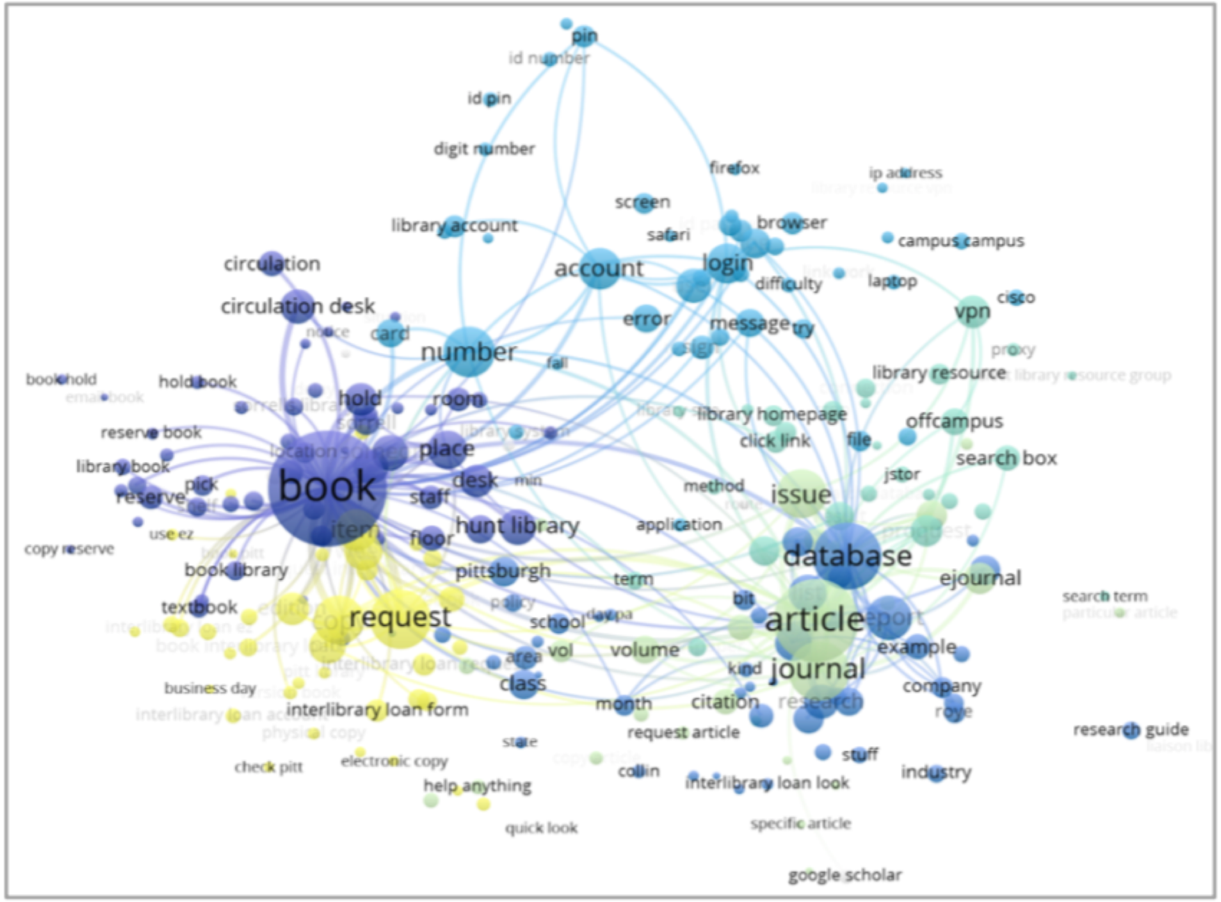


Figure 2. Screenshot of a representative distance map generated by VOSviewer, using the same preprocessed dataset used to build the LDA model. Colors show different clusters.

Software:
VOSviewer
(version 1.6.10)

Conclusions

- Chat is valuable, especially to answer quick, basic questions
- Most chat questions are related to access to resources
- Not many questions are in-depth reference questions
- Reasonable to involve both circulation and specialists in chat