Chatbot Spring 2021

Pat Tran, Elias Reta, Mark Wu, Andrew Chen

Work Division

- Andrew: LDA
- Mark: Mallet LDA
- Elias: Tomotopy LDA
- Phat: CorEx

Overview

- Provided relevance-based labeling to data
  - Experimented with different models
- Improved on preprocessing
  - Identified and preserved course-specific keywords
  - Identified and preserved course-specific formulas

Problem

- Cosine similarity/distance is great but:
  - For document A, the "closest" document B may not actually be relevant enough:
  - Closeness is relative and the closest point may still be far from the point of interest
- So, can we have a better way to relate the data we have for the chatbot?
  - Speed up the chatbot program by only looking at data known to be related to the same topic/concept
  - Could be helpful for RL?

Solution: Topic Modeling

- Associate documents to one or more topics
- Provide a new kind of labeling for the data that's different from the indexing of chapters and sections

Old Data: dspfirst_paras.json

- Indexed paragraph content from Ch.1 to Ch.4 with keywords

	id	meta	name	content	chapter	section	paragraph	tag_id	keywords
0	3	paragraph	NaN	<h2 class="ITS_BOOK">CHAPTER 1<p>Introduction<...	1	1	1	41,273,332,10,5,7,71,415	audio digital multimedia signal signals system...
1	5	paragraph	NaN	It is likely that your usage and understanding...	1	1	2	41,273,274,333,94,262,10,5,209,27,415	audio digital image information numbers point ...
2	7	paragraph	NaN	The term system ...	1	1	3	41,76,421,74,396,438,291,94,256,281,10,5,7,71,143	audio cd change definition disk general interp...
3	8	paragraph	NaN	Our goal in this text is to develop a framewor...	1	1	4	453,439,256,5,71	equations mathematical representation signals ...
4	11	paragraph	NaN	<h3 class="ITS_BOOK">1-0 Mathematical Represe...	1	2	1	439,256,5	mathematical representation signals

Old Data: espfirst.json

- Indexed paragraph content from Ch.1 to Ch.10 with no keywords
- Including different types of content:
  - Section headers, math expressions, named equations, paragraph contents, etc.

	id	meta	name	content	chapter	section	subsection	paragraph	flag
0	1	section	sec:Ahead-ZZZZ-311	Introduction	-2	1	1	1	NaN
1	2	math		\sqrt{-1}	-2	1	1	1	NaN
2	3	math		z = \pm j	-2	1	1	1	NaN
3	4	math		(b^2 -4 a c)	-2	1	1	1	NaN
4	5	math		z = -3 \pm j 4	-2	1	1	1	NaN

New data: book_1.json

- Indexed DSP-specific terms with other unknown info

	id	meta	name	chapter	section	book_id	index_id	rating	epochtime	duration	event
0	1	toc	Euler's Formula	-2	3	108	1	0	1457501835	0	toc
1	2	toc	Euler's Formula	-2	3	108	1	0	1457501840	0	index
2	3	toc	Square wave â€“ synthesis	3	6	1204	1	0	1457501936	0	toc
3	4	toc	Inverse Euler Formulas	-2	3	117	1	0	1457501942	0	index
4	5	toc	Cyclic\nfrequency	2	3	729	1	0	1457501949	0	toc

New data: index_1.json

- Another indexed DSP-specific terms

	id	name	pages	chapter_id	tag_id
0	1	A-to-D converter			1,2
1	2	Absolutely integral signals	313	11	3,4,5
2	3	Accumulator system	255	9	6,7
3	4	Adder	112	5	8
4	5	Advanced signal	14	2	9,10

New Procedures in Preprocessing

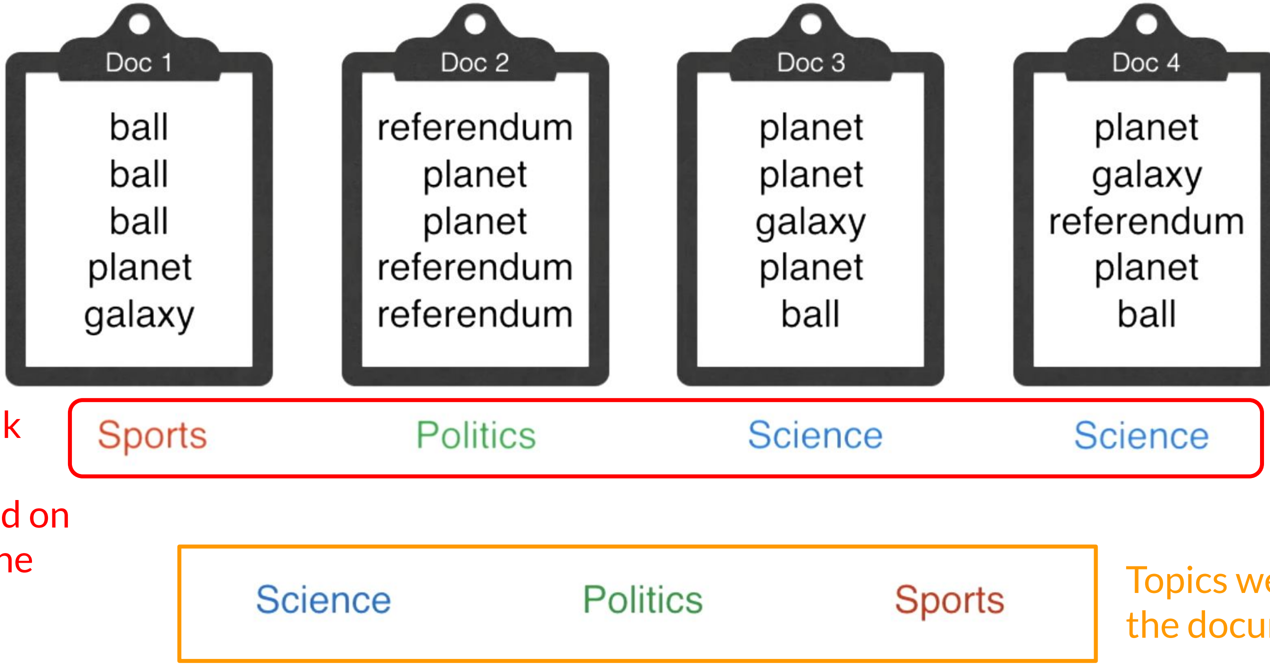
- Mapping of known named equations to their names
  - Preserves as much LaTeX semantics as possible
- Mapping of multi-token course-specific terminologies to single token
  - e.g. *c-to-d conversion* → *c_to_d_conversion*
  - e.g. *finite impulse response* → *finite_impulse_response*
  - Preserves as much key course terms as possible

Approach 1: LDA

- A Generative Model
- Represents a document in terms of Bag-of-Words
  - e.g. *“a bird on a tree”* → *[(“a”, 2), (“bird”, 1), (“on”, 1), (“tree”, 1)]*
- Assumes that there are two probability distributions from which the documents are formed

Approach 1: LDA

- At high level:

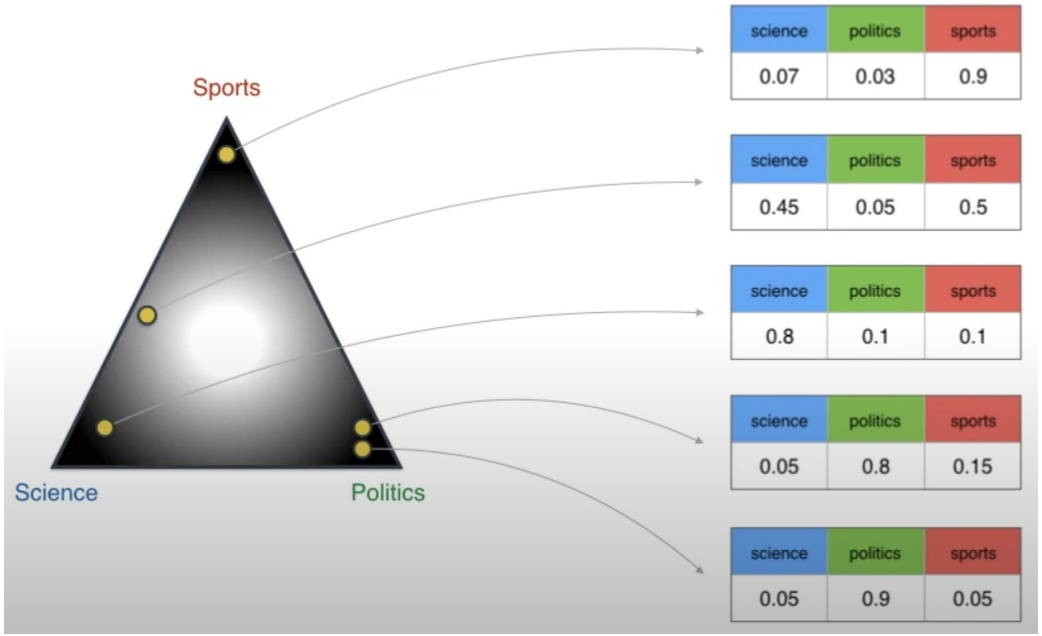


Topics we think that's most probable based on the words in the document

Topics we want to assign the documents to

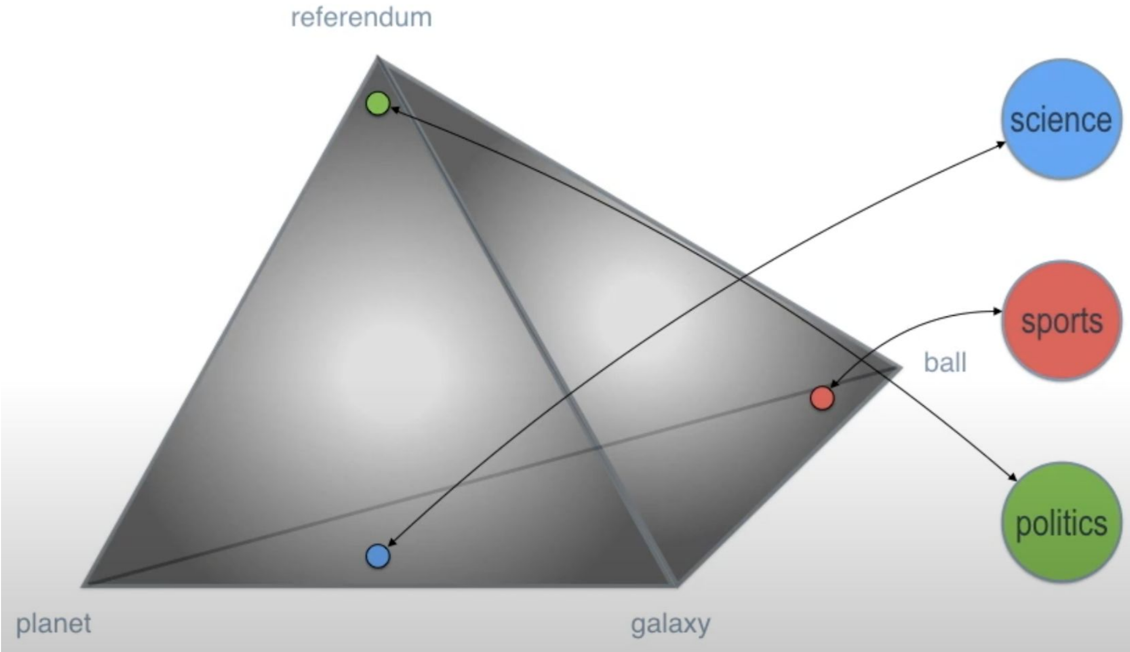
Approach 1: LDA

- Assumption 1: There is a document-topic distribution
  - *"What's the probability of a given document being of a certain topic?"*



Approach 1: LDA

- Assumption 2: There is a **topics-word distribution**
  - *"What is the probability for each word given a topic?"*



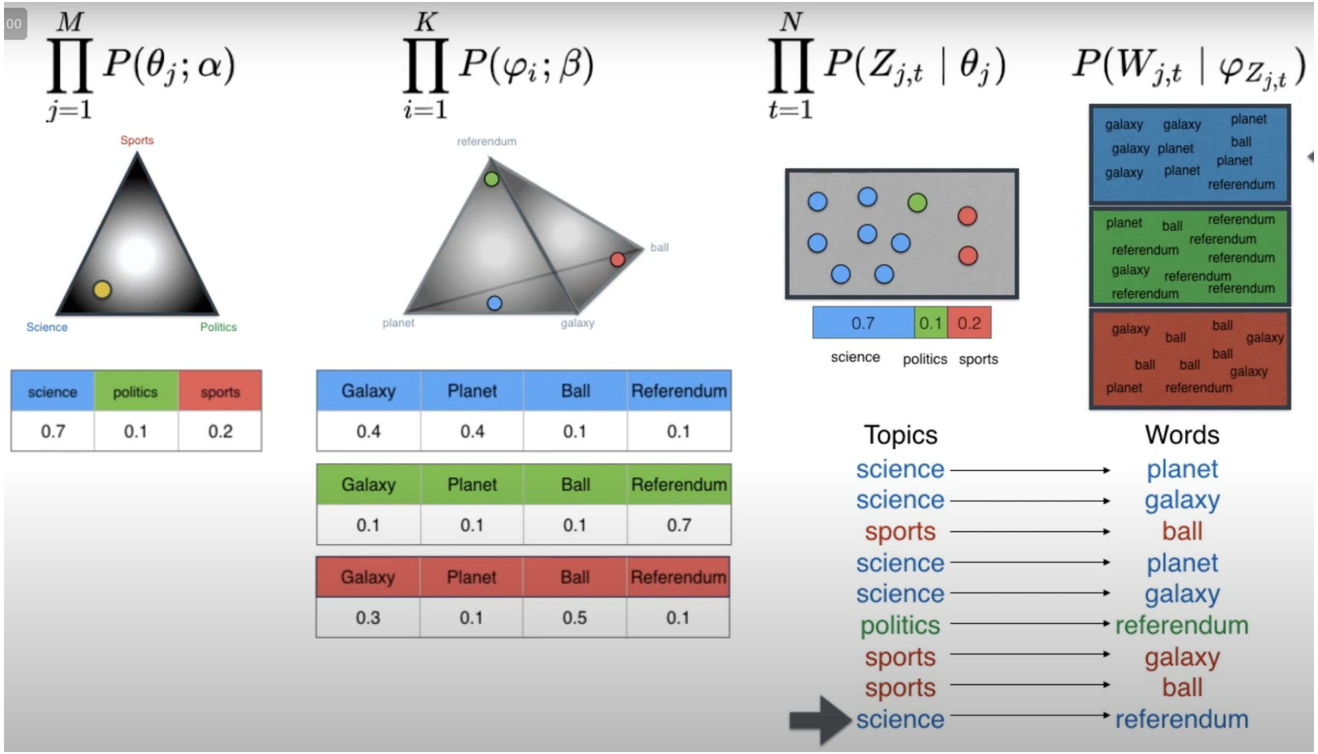
Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

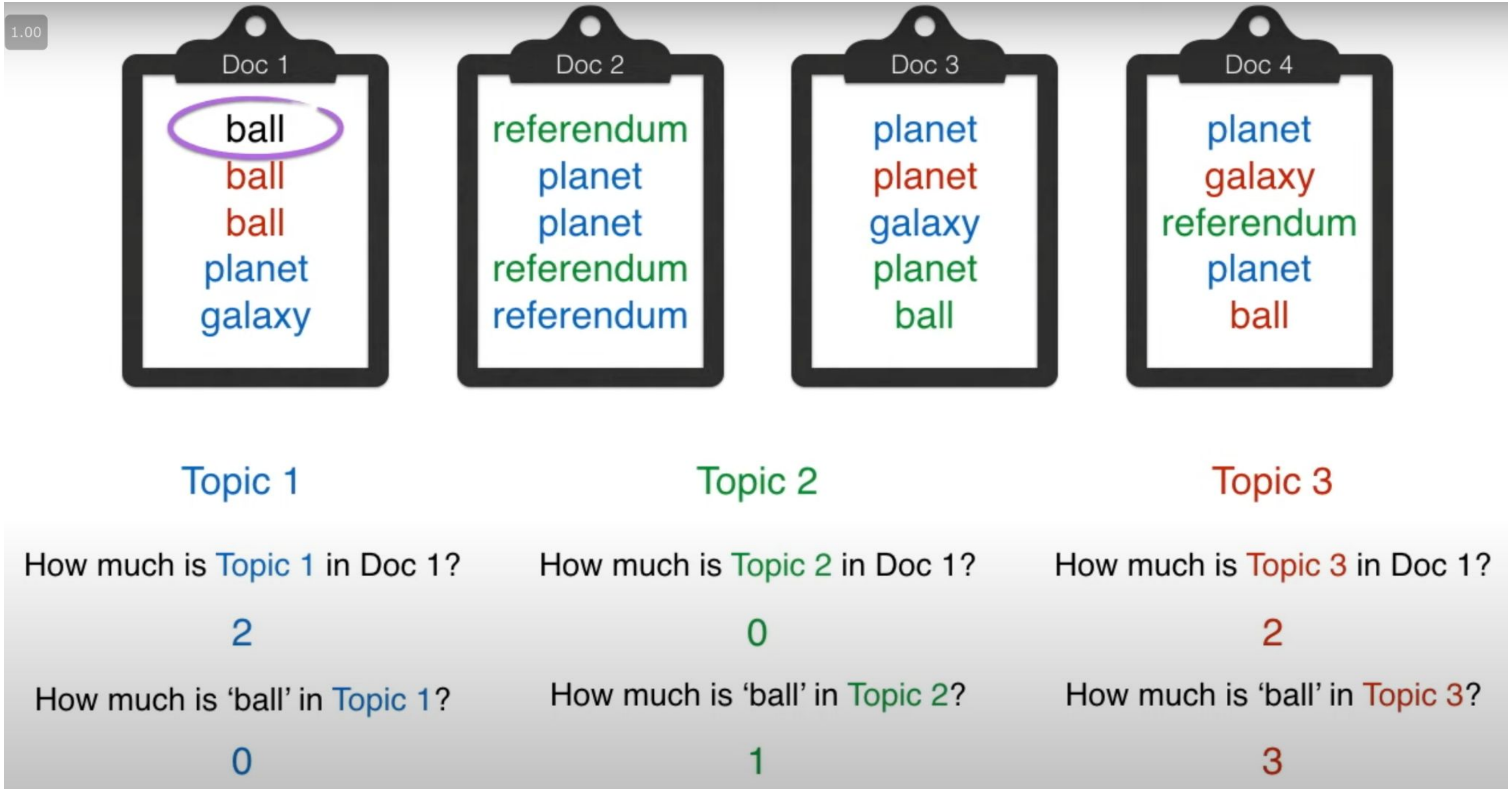
Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Approach 1: LDA

▷ These two distributions gives us a way to generate documents

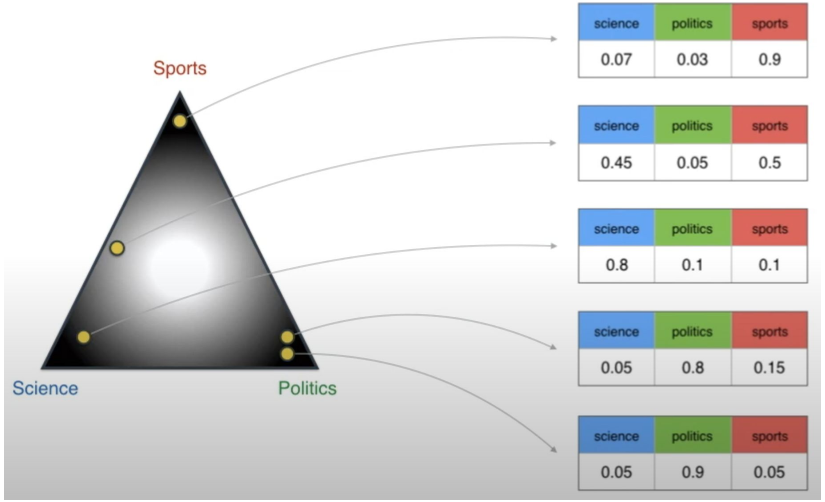


LDA Hyper-parameters

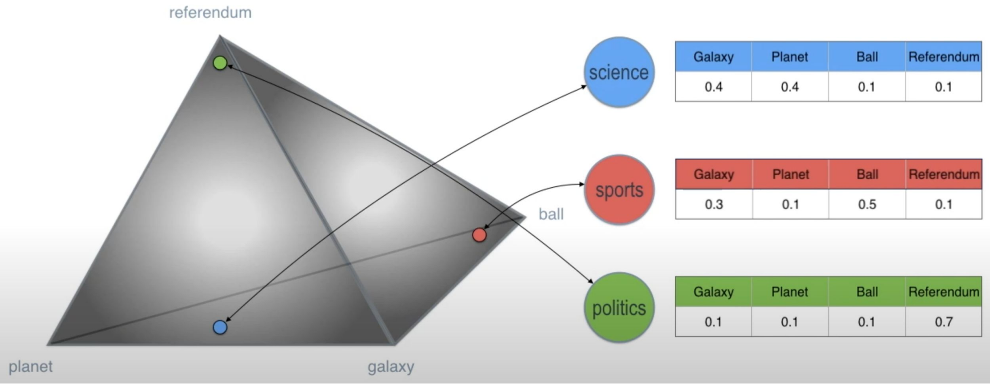


LDA Hyper-parameters

Alpha



Beta



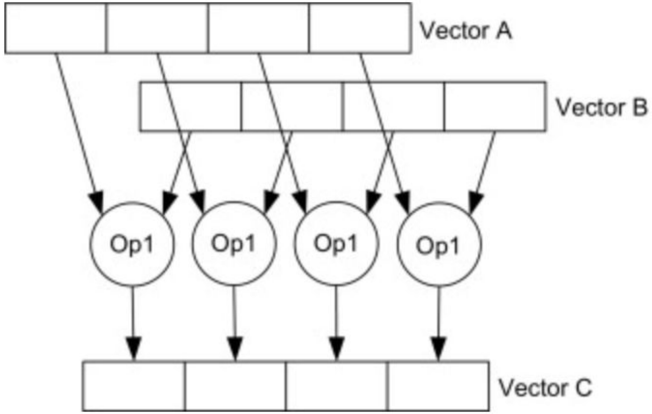
Example: [Plotly Interactive Plot]

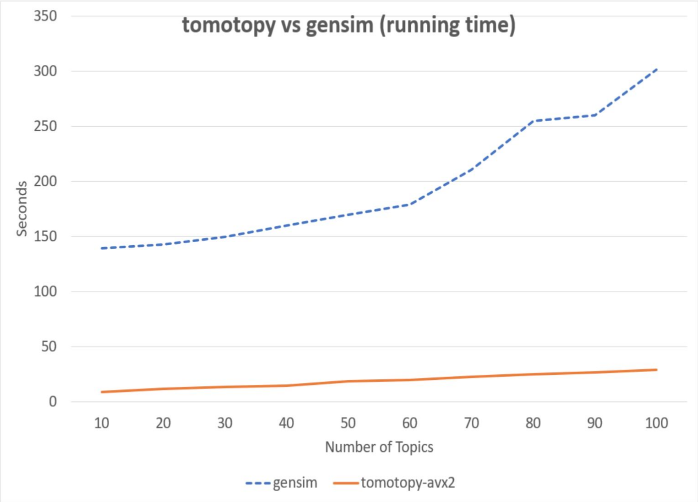
Approach 1.2 and 1.3: Mallet LDA and Tomotopy LDA

	LDA	Mallet LDA	Tomotopy LDA
Methods	Variational Bayes Sampling(VB)	Collapsed Gibbs Sampling(GS)	Collapsed Gibbs-Sampling(CGS)
Precision	Less Precise	More Precise	More Precise
Speed	2	3	1
Library Source	gensim	gensim	tomotopy
Supervision	GuidedLDA	NA	Word Priors
Speed of Computation of Iterations	1	1	2

Tomotopy vs. Gensim

- Faster Iterations
  - SIMD instruction Set
  - 1000 documents: tomotopy trains 200 iterations while gensim trains 10 iterations
- Collapsed Gibbs-Sampling(CGS) as opposed to Variational Bayes(VB)
  - Infers topics and word distribution
  - Converges slower than VB, but computation of iterations

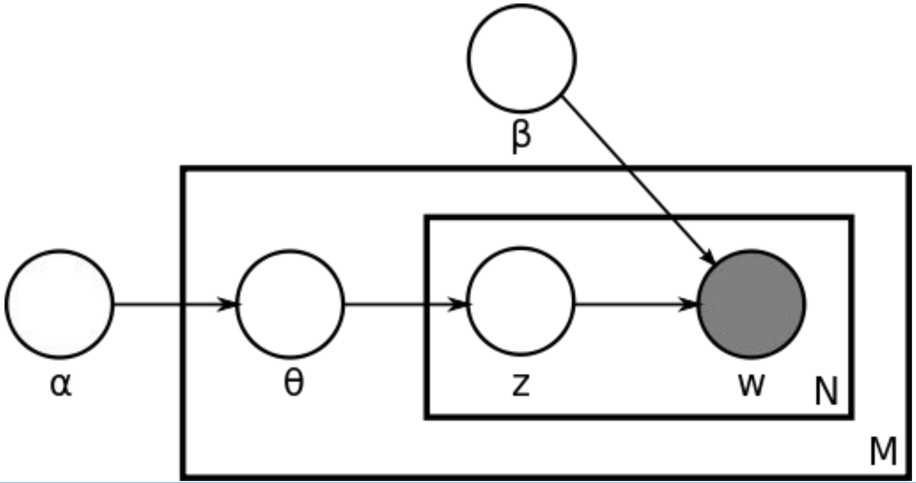




Problem with the LDA Model

1. Need a lot of data to “learn” anything meaningful
   a. LDA = Generative Model
   b. Do not work well with short documents with little text
      i. e.g. “What are finite-impulse-response filters”

         => “finite_impulse_response filter”



Approach 2: CorEx Topic Model (Correlation Explanation)

```
anchored_topic_model.get_topics(topic=0, n_words=15, weighted_rank=True)

[('term', 1.0434725396558882, 1.0),
 ('example', 0.6738308284053873, 1.0),
 ('signal', 0.40485689214985165, 1.0),
 ('audio', 0.2398543846439643, 1.0),
 ('adding', 0.16712040590418778, 1.0),
 ('stored', 0.15065388372814484, 1.0),
 ('including', 0.10059147423503077, 1.0),
 ('assumed', 0.09019139519158798, 1.0),
 ('within', 0.08207001297736025, 1.0),
 ('representation', 0.0646294119884227, 1.0),
 ('formed', 0.061998150265960755, 1.0),
 ('naturally', 0.059970221445522914, 1.0),
 ('measure', 0.050217961242714944, 1.0),
 ('remarkable', 0.04543562902655192, -1.0),
 ('mathematical', 0.032329503091223066, 1.0)]
```

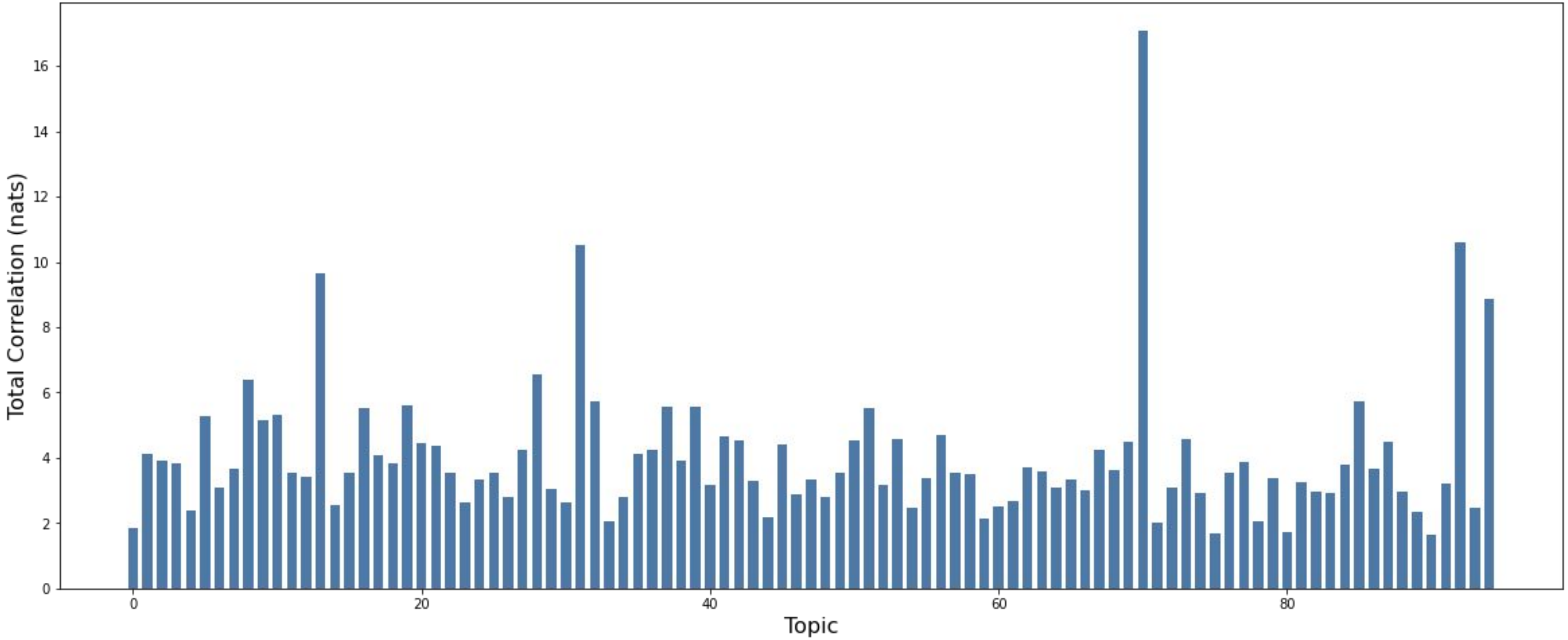
- Mutual Information (Words)
  - The information obtained about one random value given another
    - Example: knowing the month will not reveal the exact temperature, but will make certain temperatures are more or less likely to occur.
    - In CorEx, the higher the MI score, the more representative the word is to the topic.
- Total Correlation (Topics) (a.k.a. Multivariate Mutual Information)
  - Additive contributions from each word (Mutual Information score)
  - Compare the total correlation of each topic to overall to see which topic “contribute” the most to the model (next slide)

“Mutual Information.” *Wikipedia*, Wikimedia Foundation, 10 May 2019, simple.wikipedia.org/wiki/Mutual_information#:~:text=Mutual%20information%20measures%20how%20much,it%20will%20give%20some%20hint.

Approach 2: Total Correlation Score

Overall TC Score: 383.8647

Avg. TC Score: 4.04068 for 95 topics



Approach 2: CorEx Topic Model

	chap_sec	content	book_keywords	topic_0	topic_1	topic_2	topic_3	topic_4	topic_5	topic_6	topic_7	topic_8	topic_9	topic_10	topic_11	topic_12	topic_13	topic_14	topic_15	topic_16	t
0	1.0	introduction book signal system age multimedia...	signal system mathematical representation exam...	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
1	1.1	mathematical representation signal signal patt...	signal time_waveform speech continuous_time di...	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	
2	1.2	mathematical representation system already sug...	system definition one_dimensional continuous_t...	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
3	1.3	thinking system block_diagram useful represent...	cd audio system analog_to_digital a_to_d conve...	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	1.4	next step cd audio_system good example discret...	multimedia information system tuning fork expe...	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
90	10.8	steady_state response stability stable system ...	output pole sequence term response	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	
91	10.9	second_order filter turn attention filter two ...	pole difference_equation difference_equations ...	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	
92	10.10	frequency_response second_order iir filter sin...	frequency_response pole zero unit_circle funct...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
93	10.11	example iir lowpass_filter first_order second_...	filter frequency_response example pole zero	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
94	10.12	summary link class iir filter wa introduced ch...	domain between_domains three_domain demo three...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

95 rows × 98 columns

Each document can be associated to multiple topics

CorEx: Anchor Words

```
print(esp_seed_topic)
print(len(esp_seed_topic))
```

```
[['signal', 'system', 'mathematical', 'representation', 'example', 'term', 'audio', 'stored', 'store'], ['signal', 'ti
95
```

```
anchored_topic_model = ct.Corex(n_hidden=95, max_iter=500, seed=1)
# Anchor the main keywords
anchored_topic_model.fit(doc_word, anchors=esp_seed_topic, anchor_strength=3, words=words, docs=row_label);
```

- Base method:

  def fit(self, X, anchors=None, anchor_strength=1, words=None, docs=None)

  a. X: scipy sparse CSR matrix (binary vectorize textbook content)
  b. anchors: list of anchor words
  c. anchor_strength: how much weight to assign to the anchor words relative to all the other words (anchor_strength=2 means to give twice the weight (MI score) to the anchor words compared to other words)
  d. words: list of strings that label the corresponding columns
  e. docs: list of Strings that label the corresponding rows

CorEx: Anchor Words

anchor_strength = 1

```
anchored_topic_model.get_topics(topic=0, n_words=15, weighted_rank=True)
```

```
[('including', 0.23221949686626517, 1.0),
 ('assumed', 0.2077027925989838, 1.0),
 ('relatively', 0.13816635599666688, 1.0),
 ('key', 0.12789516237224202, 1.0),
 ('sufficient', 0.12746610042248419, 1.0),
 ('term', 0.11695123639408368, 1.0),
 ('involve', 0.1165733383225404, 1.0),
 ('perfect', 0.11618822255776648, 1.0),
 ('denoted', 0.11326337260193749, 1.0),
 ('illustrates', 0.10939248932092667, 1.0),
 ('formed', 0.1003585185803298, 1.0),
 ('additive', 0.09475206645045689, 1.0),
 ('guaranteed', 0.09475206645045689, 1.0),
 ('four', 0.08681447323410248, 1.0),
 ('example', 0.08630833917353588, 1.0)]
```

anchor_strength = 2

```
anchored_topic_model.get_topics(topic=0, n_words=15, weighted_rank=True)
```

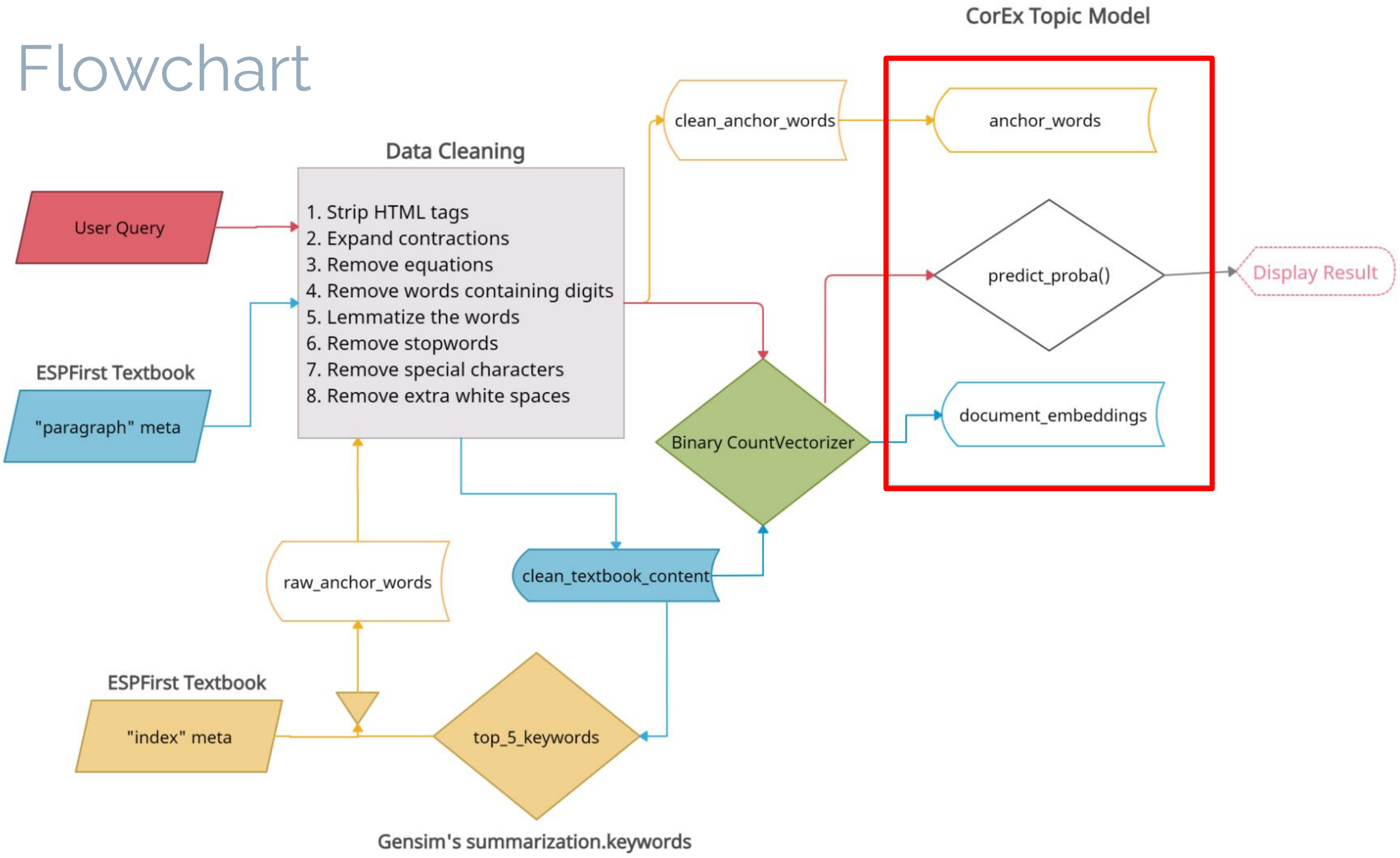
```
[('term', 0.2596919809890868, 1.0),
 ('example', 0.21194425672567554, 1.0),
 ('including', 0.19951052363899496, 1.0),
 ('assumed', 0.1786851585540792, 1.0),
 ('describe', 0.15837338062047038, 1.0),
 ('conclude', 0.13854707779476447, 1.0),
 ('illustrates', 0.13680379272141238, 1.0),
 ('involve', 0.13292876576475007, 1.0),
 ('relatively', 0.11918061415148956, 1.0),
 ('representation', 0.11639153018898576, 1.0),
 ('work', 0.11212290370246243, 1.0),
 ('formed', 0.10593894839086597, 1.0),
 ('sufficient', 0.10240590175947083, 1.0),
 ('infinite', 0.09835801967409115, 1.0),
 ('signal', 0.09808933990139937, 1.0)]
```

```
print(esp_seed_topic)
print(len(esp_seed_topic))
```

```
[['signal', 'system', 'mathematical', 'representation', 'example', 'term', 'audio', 'stored', 'store'], ['signal', 'ti
95
```

```
anchored_topic_model = ct.Corex(n_hidden=95, max_iter=500, seed=1)
# Anchor the main keywords
anchored_topic_model.fit(doc_word, anchors=esp_seed_topic, anchor_strength=3, words=words, docs=row_label);
```

CorEx: Flowchart



Recall: espfirst.json

- Used in CorEx Model
  - “index” meta is used as the anchor words
  - “paragraph” meta is used as the textbook content
  - “math” and “equation” meta are used as keys in dictionary to remove the raw equations from the raw textbook content

```
Initital count of unique 'math' meta:  2133
Count of unique usage of 'math' meta:  1461
{'s(t)': '',
 '\\{\\dots, -2, -1, 0, 1, 2, \\dots\\}': '',
 't_s': '',
 '( )': '',
 '[ ]': '',
 'p(x,y)': '',
 'p(x_0,y_0)': '',
 '(x_0,y_0)': '',
 'v(x,y,t)': '',
 'p[m,n]': '',
 'x(t)': '',
```

	chap_sec	chapter	section	content	book_keywords	section_keywords	subsection_keywords	generated_keywords
0	1.0	1	0	introduction book signal system age multimedia...	signal system mathematical representation exam...			signal example term audio stored store
1	1.1	1	1	mathematical representation signal signal patt...	signal time_waveform speech continuous_time di...	mathematical representation signal		signal time represented function represent rep...
2	1.2	1	2	mathematical representation system already sug...	system definition one_dimensional continuous_t...	mathematical representation system		signal squarer_system squarer continuous_time ...
3	1.3	1	3	thinking system block_diagram useful represent...	cd audio system analog_to_digital a_to_d conve...	thinking system		conversion number sample example converter
4	1.4	1	4	next step cd audio_system good example discret...	multimedia information system tuning fork expe...	next step		concept signal understanding function chapter

Note: **book_keywords** values are from “index” meta, **subsection_keywords** values are not used, **generated_keywords** values are generated from gensim summarization.keywords using **content** values

Approach 2: Basics

- Declare CorEx model:

  ```
  anchored_topic_model = ct.Corex(n_hidden=95, max_iter=500, seed=1)
  ```

  - n_hidden: # of latent topics
    - It is 95 since there are 95 sections, and the esp textbook content is grouped by section
  - max_iter: # of iterations before ending (optional)
  - seed: a number that gives the same result if declared (optional)
  - **Note**: there are more, but are not used for our model

Approach 2: Data Input

- An information-theoretic approach
  - Takes in a binary word embeddings (presence or absence of a term instead of the raw counts)
    - e.g. *“a bird on on a tree” → [(“bird”, 1), (“on”, 1), (“tree”, 1)]*
    - Requires at least 2 characters for a match (pattern of “\b\w\w+\b”)
  - Aim to explain the relevance of words in documents through latent topics

```
vectorizerTemp = CountVectorizer(max_features=10000, binary=True)
X = ['a bird on on a tree', 'the tree on a mountain']
doc_word1 = vectorizerTemp.fit_transform(X)
doc_word1 = ss.csr_matrix(doc_word1)
```

	bird	mountain	on	the	tree
0	1	0	1	0	1
1	0	1	1	1	1

Binary word embeddings of:
'a bird on on a tree' and **'the tree on a mountain'**

```
anchored_topic_model.get_topics(topic=0, n_words=15, weighted_rank=True)
```

```
[('term', 1.0434725396558882, 1.0),
 ('example', 0.6738308284053873, 1.0),
 ('signal', 0.40485689214985165, 1.0),
 ('audio', 0.2398543846439643, 1.0),
 ('adding', 0.16712040590418778, 1.0),
 ('stored', 0.15065388372814484, 1.0),
 ('including', 0.10059147423503077, 1.0),
 ('assumed', 0.09019139519158798, 1.0),
 ('within', 0.08207001297736025, 1.0),
 ('representation', 0.0646294119884227, 1.0),
 ('formed', 0.061998150265960755, 1.0),
 ('naturally', 0.059970221445522914, 1.0),
 ('measure', 0.050217961242714944, 1.0),
 ('remarkable', 0.04543562902655192, -1.0),
 ('mathematical', 0.032329503091223066, 1.0)]
```

The top 15 keywords that represent the **topic 0**
(‘words’, mutual information, presence/absence)

Approach 2: CorEx Attributes

- Important attributes from CorEx Topic:
  - p_y_given_x
    - Probabilities of a topic given the words in a document (Y = topic, X = document's words).
  - total correlation (tc or tcs)
    - Relatively compare between two or more topics. Topics with higher TC will "explain" more about the collection of documents
    - Used in Total Correlation Graph (slide 17)
  - log_z
    - Pointwise estimate of total correlation for that topic (used for words)

```python
# Row label
rowLabel = df.chap_sec.tolist()
temp3 = anchored_topic_model.p_y_given_x
# The probabilities of documents in topic 5
for index, value in enumerate(temp3[5]):
  print("Section {} has a probability of {}% to be in topic 5.".format(rowL
```

```
Section 1.0 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 1.1 has a probability of 99.9999% to be in topic 5.
Section 1.2 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 1.3 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 1.4 has a probability of 0.016784599641618627% to be in topic 5.
Section 2.0 has a probability of 99.9999% to be in topic 5.
Section 2.1 has a probability of 99.9999% to be in topic 5.
Section 2.2 has a probability of 99.9999% to be in topic 5.
Section 2.3 has a probability of 99.9999% to be in topic 5.
Section 2.4 has a probability of 0.2129895011192784% to be in topic 5.
Section 2.5 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 2.6 has a probability of 0.1946580210651914% to be in topic 5.
Section 2.7 has a probability of 99.9999% to be in topic 5.
Section 2.8 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 2.9 has a probability of 99.9999% to be in topic 5.
Section 3.0 has a probability of 99.9999% to be in topic 5.
Section 3.1 has a probability of 1.014980919642100% to be in topic 5.
Section 3.2 has a probability of 99.9999% to be in topic 5.
Section 3.3 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 3.4 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 3.5 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 3.6 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 3.7 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 3.8 has a probability of 0.00041941388527435521% to be in topic 5.
Section 3.9 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 3.10 has a probability of 99.9999% to be in topic 5.
Section 3.11 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 4.0 has a probability of 9.999999999999999e-05% to be in topic 5.
Section 4.1 has a probability of 9.999999999999999e-05% to be in topic 5.
```

How We Evaluate (I): Coherence Score

- Measured by CoherenceModel from Gensim
  - Coherence score: measures the relative distance between words within a topic
  - In our opinion, the score of 0.3 is bad, 0.4 is low, 0.55 is okay, 0.7 is great, 0.85+ is probably wrong
- Use c_v coherence measure (ranging from 0 to 1, the higher the better)
-

LDA (gensim)	Mallet LDA	LDA (tomotopy)	CorEx
0.4837	0.5056	0.5581	0.6400

Result 1: LDA

Coherence Score: 0.483666 (min = 0, max = 1)

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #1	window	signal	frequency	length	spectrogram	n_s	time	dft	analysis	short
Topic #2	system	signal	xn	output	input	sequence	yn	transform	example	use
Topic #3	time	period	signal	spectrum	xn	discrete	dft	coefficient	periodic	sample
Topic #4	pulse	cosine_wave	conversion	cubic	spline	shape	triangular	ynp	reconstruction	overlap
Topic #5	precision	optical	system	signal	audio_system	audio	recording	playback	time	part
Topic #6	signal	show	xn	frequency	sequence	time	ej	output	two	system
Topic #7	filter	ideal	frequency_response	passband	lpf	impulse_response	window	show	stopband	ripple
Topic #8	ej	signal	frequency	dft	dtft	xn	show	use	sum	plot
Topic #9	ej	zero	pole	filter	frequency_response	domain	unit_circle	system	show	system_function
Topic #10	signal	time	frequency	sample	discrete	sinusoid	show	system	continuous	f_s

Result 1.2: Mallet LDA

Coherence Score: 0.505598154

- Notice: Higher score than standard LDA's

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #1	ej	frequency_response	cos	complex	complex_exponential	function	phase	point	magnitude	cosine
Topic #2	system	output	input	impulse_response	filter	fir_filter	yn	difference_equation	xn	block
Topic #3	time	sample	discrete	continuous	f_s	signal	t_s	pulse	converter	show
Topic #4	dtft	filter	ej	ideal	frequency_response	passband	show	impulse_response	design	frequency
Topic #5	period	spectrum	coefficient	sum	fourier_series	a_k	integral	periodic	periodic_signal	ej
Topic #6	signal	frequency	time	plot	sinusoid	show	spectrum	hz	component	note
Topic #7	pole	system_function	transform	filter	order	unit_circle	domain	iir	polynomial	coefficient
Topic #8	xn	sequence	signal	property	time	transform	representation	result	sum	finite
Topic #9	dft	length	window	xk	sample	dtft	point	idft	interval	n_s
Topic #10	signal	system	lab	tuning_fork	chapter	cd	rate	produce	represent	number

Result 1.3: Tomotopy LDA

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #0	signal	system	represent	representation	chapter	example	diagram	value	mathematical	show
Topic #1	cd	synthesis	ak	complex_amplitude	lab	square_wave	finite	phasor_addition	rom	fourier_analysis
Topic #2	analysis	sound	music	vary	concept	write		synthesize	musical	result
Topic #3	signal	cosine	second	plot	time	phase_shift	multiply	envelope	high	beat
Topic #4	cos	sinusoid	call	form	example	rad	draw	simple	may	beat_note
Topic #5	rate	flash	spot	rotation	clockwise	strobe	disk	motor	rotate	rpm
Topic #6	period	function	sine	fourier_series	integral	coefficient	identity	property	cosine	periodic
Topic #7	sample	ts	cosine	interpolation	sampling_period	plot	curve	smooth	reconstruct	question
Topic #8	time	sample	discrete	continuous	signal	alias	xn	plot	output	sec
Topic #9	note	time	key	example	sec	make	duration	scale	axis	change
Topic #10	tuning_fork	equation	tine	produce	differential	physical	experiment	solution	model	tone
Topic #11	complex	complex_exponential	number	phasor	vector	real	rotate	representation	signal	real_part
Topic #12	show	original	sampling	sampling_rate	sampling_theorem	analog	aliase	line	would	input
Topic #13	frequency	spectrum	hz	one	component	line	sinusoid	contain	domain	high
Topic #14	result	per	also	possible	use	must	fact	case	thus	matlab
Topic #15	two	signal	use	sum	show	plot	waveform	term	define	general
Topic #16	function	signal	number	point	would	variable	many	although	call	often
Topic #17	give	time	make	equation	add	positive	find	exercise	involve	illustrate
Topic #18	pulse	converter	ideal	conversion	reconstruction	digital	continuous	time	yn	cosine_wave
Topic #19	signal	sinusoid	sinusoidal	see	formula	frequency	follow	amplitude	phase	since

Result 1.3: Tomotopy LDA

Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
tuning_fork	converter	function	lab	sample	complex	frequency	note	plot	time
produce	show	cosine	matlab	ts	phasor	sinusoid	frequency	sum	discrete
tine	output	mathematical	synthesis	pulse	signal	line	waveform	cosine	continuous
sound	input	variable	cd	reconstruction	vector	component	key	two	sample
tone	original	plot	involve	ideal	imaginary_part	sec	music	call	alias
real	clockwise	general	equation	analysis	complex_attitude	get	time	amplitude	differential
vector	strobe	zero	figure	would	periodic_signal	sec	variable	music	model
rotate	disk	term	possible	notation	synthesis	one	number	much	sound
representation	motor	result	may	vary	square_wave	original	example	show	experiment
represent	rotate	easy	consider	axis	fundamental_frequency	contain	show	sum	solution

Result 2: CorEx

Coherence Score: 0.6400 (min = 0, max = 1)

		Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
0	Topic #0	term	example	signal	audio	adding	stored	including	assumed	within	representation
1	Topic #1	represented	time	represents	period	mathematical	signal	representation	function	represent	two_dimensional
2	Topic #2	block	diagram	output	operation	operator	c_to_d	converter	signal	discrete_time	thinking
3	Topic #3	block	diagram	converter	cd	conversion	example	audio	sample	a_to_d	next
4	Topic #4	information	concept	chapter	understanding	signal	introduce	equally	multimedia	tuning	electrical
5	Topic #5	sinusoid	amplitude	frequency	sinusoidal	cosine	phase	function	signal	period	radian
6	Topic #6	sinusoid	sinusoidal	plot	signal	sound	middle	tuning_fork	plotted	wide	parameter
7	Topic #7	cosine	plotting	function	sinusoid	sine	angle	identity	trigonometric	examine	key
8	Topic #8	sinusoid	amplitude	frequency	sinusoidal	phase	time	signal	exponential	periodic	period
9	Topic #9	plotted	plot	plotting	period	function	sampled	curve	point	gray	ie
10	Topic #10	exponential	real	formula	magnitude	complex	complex_exponential_signal	euler	complex_exponential_signals	amplitude	inverse

How We Evaluate (II): Human Judgment

1. Provide example input to the model
    a. e.g. “What are finite impulse response filters?”
2. Predict the most probable topics
3. Compare the documents, namely textbook data, that were related to the same topic(s) when the model was trained
4. Repeat the process with a few example inputs including logistical questions and conceptual questions

How We Evaluate (II): Human Judgment

	Example 1	Example 2	Example 3
Raw Question	“What are FIR filters?”	“What are finite-impulse-response filters?”	“Are calculators allowed for the exams?”
Processed Input	“fir filter”	“finite_impulse_response filter”	“calculator allowed exam”
LDA	Chapter 5, 6, 2	Chapter 7	None
Mallet LDA	Chapter 6, 9	Chapter 8 & 2(equal weight)	Chapter 5 & 9(equal weight), 4 & 7 & 8(equal weight)
CorEx	Chapter 5, 6, 4	Chapter 5 & 6(equal weight), 4	None

How We Evaluate Human Judgement(Tomotopy)

What are FIR Filters?

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #3	signal	represent	representation	time	example	value	equation	take	define	general

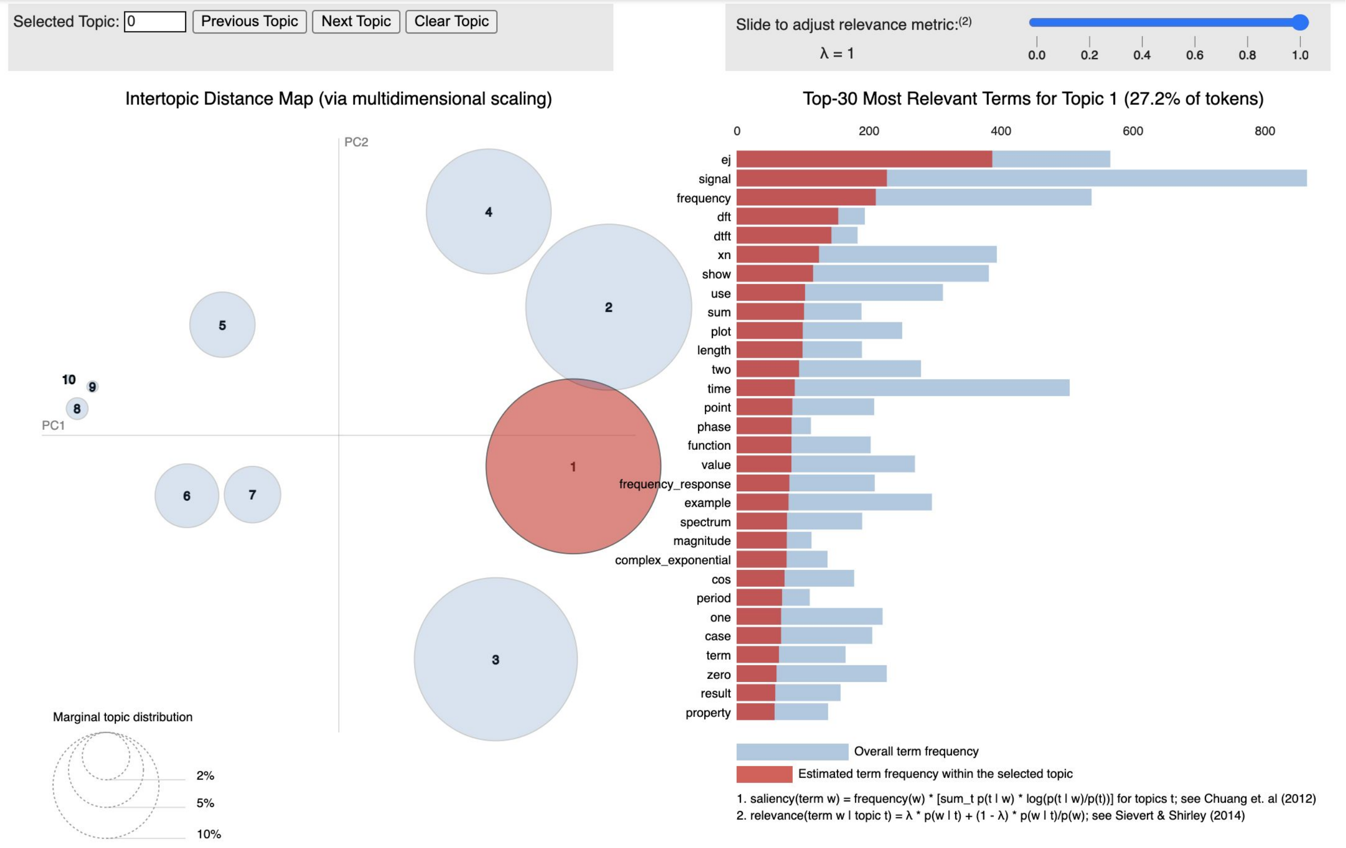
What are finite-impulse-response filters?

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #1	time	change	figure	analysis	section	would	constant	whose	many	notation

Are calculators allowed on the exam?

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #7	system	show	mathematical	diagram	form	may	although	square	operation	important

Result 1: LDA



Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1 2 3 4 5 6 7 8 9 10
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (27.2% of tokens)
0 200 400 600 800
ej
signal
frequency
dft
dtft
xn
show
use
sum
plot
length
two
time
point
phase
function
value
frequency_response
example
spectrum
magnitude
complex_exponential
cos
period
one
case
term
zero
result
property
Overall term frequency
Estimated term frequency within the selected topic
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)

Result 1: LDA (“What are FIR filters?”)

Processed Input: “fir filter”

```
Predicted Topic:  1
```

	content	chap_sec	raw_content	content_length	lda_topic
10	complex_exponential phasor show cosine signal ...	2.5	Complex Exponentials and Phasors \r\n ...	1053	1
34	chapter fir_filter point focus attention signa...	5.0	<h3><span class="chapter text-center">CHAPTER ...	128	1
35	discrete time system discrete time system comp...	5.1	Discrete-Time Systems \r\n \r\n A disc...	187	1
36	running_average_filter simple useful transform...	5.2	The Running-Average Filter \r\n \r\n ...	492	1
37	general fir_filter note special case general d...	5.3	The General FIR Filter \r\n Note that ...	1087	1
38	implementation fir_filter recall general defin...	5.4	Implementation of FIR Filters \r\n Rec...	816	1
39	linear time invariant lti_system section discu...	5.5	Linear Time-Invariant (LTI) Systems \r...	473	1
40	convolution lti_system consider lti discrete t...	5.6	Convolution and LTI Systems \r\n \r\n ...	517	1
41	cascade lti_system cascade_connection two syst...	5.7	Cascaded LTI Systems \r\n \r\n In a c...	316	1
42	example fir_filtering conclude chapter example...	5.8	Example of FIR Filtering \r\n \r\n We...	322	1
44	chapter frequency_response fir_filter chapter ...	6.0	<h3><span class="chapter text-center">CHAPTER ...	113	1
45	sinusoidal_response fir_system linear time inv...	6.1	Sinusoidal Response of FIR Systems \r\...	355	1

Result 1: LDA (“What are finite-impulse-response filters?”)

Processed Input: “finite_inpulse_response filter”

```
Predicted Topic:   6
```

	content	chap_sec	raw_content	content_length	lda_topic
57	ideal filter practical application lti discret...	7.4	Ideal Filters \n\n\nIn any practical a...	677	6
58	practical fir_filter ideal filter useful conce...	7.5	Practical FIR Filters \n Ideal filters...	1250	6

Result 1: LDA: ("Explain continuous-to-discrete conversion.")

Processed Input: "explain continuous_to_discrete conversion"

```
Predicted Topic:  9
```

	content	chap_sec	raw_content	content_length	lda_topic
0	chapter introduction book signal system age mu...	1.0	<h3><span class="chapter text-center">CHAPTER ...	217	9
1	mathematical representation signal signal patt...	1.1	<h3>Mathematical Representation of Signals</h3...	479	9
2	mathematical representation system already sug...	1.2	<h3>Mathematical Representation of Systems</h3...	301	9
3	think system block diagram useful represent co...	1.3	Thinking About Systems \r\n Block diag...	229	9
4	next step cd audio_system good example discret...	1.4	The Next Step \r\n The CD audio system...	102	9
5	chapter sinusoid begin discussion introduce ge...	2.0	<h3><span class="chapter text-center">CHAPTER ...	122	9
6	tuning_fork experiment one reason cosine_wave ...	2.1	Tuning Fork Experiment \r\n One of the...	301	9
7	review sine cosine function sinusoidal signal ...	2.2	Review of Sine and Cosine Functions \r...	308	9
8	sinusoidal signal general mathematical formula...	2.3	Sinusoidal Signals \r\n The most gener...	683	9
9	sampling plot sinusoid plot sinusoid chapter c...	2.4	Sampling and Plotting Sinusoids \r\n A...	406	9
12	physics tuning_fork section describe simple ex...	2.7	Physics of the Tuning Fork \r\n In\r\...	562	9
13	time signal formula purpose chapter introduce ...	2.8	Time Signals: More Than Formulas \r\n ...	263	9

Result 1: LDA (“Explain C-to-D conversion.”)

Processed Input: “explain c_to_d conversion”

```
Predicted Topic:  9
```

	content	chap_sec	raw_content	content_length	lda_topic
0	chapter introduction book signal system age mu...	1.0	<h3><span class="chapter text-center">CHAPTER ...	217	9
1	mathematical representation signal signal patt...	1.1	<h3>Mathematical Representation of Signals</h3...	479	9
2	mathematical representation system already sug...	1.2	<h3>Mathematical Representation of Systems</h3...	301	9
3	think system block diagram useful represent co...	1.3	Thinking About Systems \r\n Block diag...	229	9
4	next step cd audio_system good example discret...	1.4	The Next Step \r\n The CD audio system...	102	9
5	chapter sinusoid begin discussion introduce ge...	2.0	<h3><span class="chapter text-center">CHAPTER ...	122	9
6	tuning_fork experiment one reason cosine_wave ...	2.1	Tuning Fork Experiment \r\n One of the...	301	9
7	review sine cosine function sinusoidal signal ...	2.2	Review of Sine and Cosine Functions \r...	308	9
8	sinusoidal signal general mathematical formula...	2.3	Sinusoidal Signals \r\n The most gener...	683	9
9	sampling plot sinusoid plot sinusoid chapter c...	2.4	Sampling and Plotting Sinusoids \r\n A...	406	9
12	physics tuning_fork section describe simple ex...	2.7	Physics of the Tuning Fork \r\n In\r\...	562	9

Result 1: LDA (“What is phase difference?”)

Processed Input: “phase difference”

```
Predicted Topic:   7
```

	content	chap_sec	raw_content	content_length	lda_topic
11	phasor_addition many situation necessary add t...	2.6	Phasor Addition \r\n There are many\r...	584	7
15	chapter spectrum representation chapter introd...	3.0	<h3><span class="chapter text-center">CHAPTER ...	92	7
16	spectrum sum sinusoid one reason sinusoid impo...	3.1	The Spectrum of a Sum of Sinusoids \r\...	556	7
18	periodic waveform periodic_signal satisfie con...	3.3	Periodic Waveforms \r\n A periodic s...	571	7
19	fourier_series example sec show synthesize per...	3.4	Fourier Series \r\n \r\n The example...	433	7
20	spectrum fourier_series discuss spectrum secti...	3.5	Spectrum of the Fourier Series \r\n\r\...	204	7
21	fourier_analysis periodic_signal synthesize pe...	3.6	Fourier Analysis of Periodic Signals ...	763	7
22	parsevals theorem one reason fourier_series us...	3.7	Parseval's Theorem \r\n One reason...	153	7
23	nature fourier_series approximation one conseq...	3.8	<h3>The Nature of the Fourier Series Approxima...	243	7
46	superposition frequency_response principle_of_...	6.2	<h3>Superposition and the Frequency Response</...	392	7
48	property frequency_response frequency_response...	6.4	Properties of the Frequency Response \...	322	7
49	graphical representation frequency_response tw...	6.5	<h3>Graphical Representation of the Frequency ...	607	7

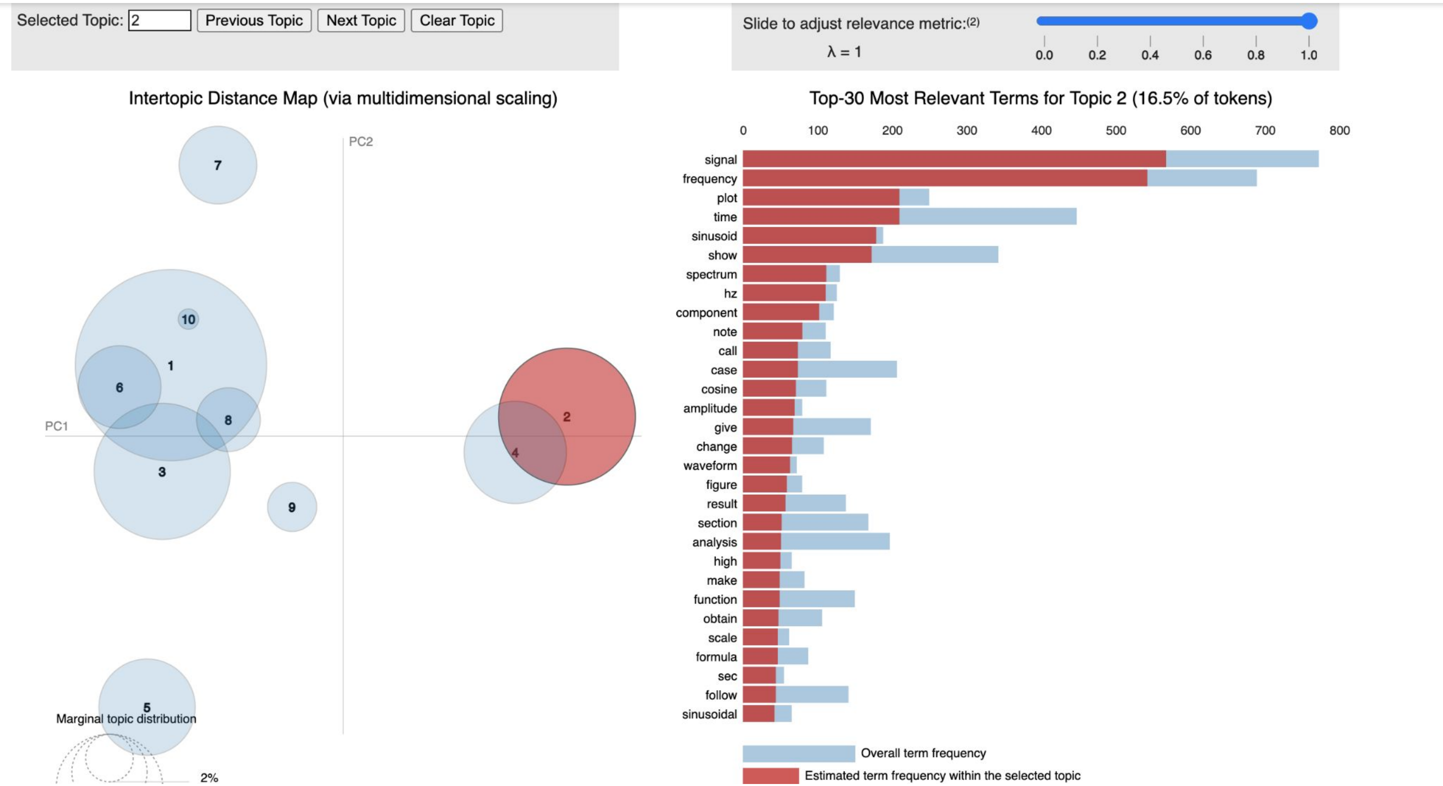
Result 1: LDA (“Are calculators allowed for the exams?”)

Processed Input: “calculator allowed exam”

```
Predicted Topic:  4
```

content	chap_sec	raw_content	content_length	lda_topic

Result 1.2: Mallet LDA (Visualization)



Result 1.2: Mallet LDA (“What are FIR filters?”)

```
Predicted Topic:  6
```

	content	chap_sec	raw_content	content_length	ldamallet_topic
23	nature fourier_series approximation one conseq...	3.8	<h3>The Nature of the Fourier Series Approxima...	243	6
38	implementation fir_filter recall general defin...	5.4	Implementation of FIR Filters \r\n Rec...	816	6
45	sinusoidal_response fir_system linear time inv...	6.1	Sinusoidal Response of FIR Systems \r\...	355	6
48	property frequency_response frequency_response...	6.4	Properties of the Frequency Response \...	322	6
50	cascade lti_system section show two lti_system...	6.6	Cascaded LTI Systems \r\n \r\n In Sec...	249	6
51	run average filter simple linear time invarian...	6.7	Running-Average Filtering \r\n \r\n A...	1034	6
53	summary link chapter introduce concept frequen...	6.9	Summary and Links \r\n This chapter in...	177	6
57	ideal filter practical application lti discret...	7.4	Ideal Filters \n\n\nIn any practical a...	677	6
76	convolution transform section observe unit_del...	9.5	Convolution and the \(z\)-Transform \r...	742	6
78	useful filter understand tie domain exploit kn...	9.7	Useful Filters \r\n \r\n Now that we u...	710	6
79	practical bandpass_filter design although much...	9.8	Practical Bandpass Filter Design \r\n ...	411	6
80	property linear_phase filter filter discuss se...	9.9	Properties of Linear-Phase Filters \r\...	297	6
86	pole zero interesting fact transform system_fu...	10.4	Poles and Zeros \r\n An interesting fa...	370	6

Result 1.2: Mallet LDA
(“What are finite-impulse-response filters?”)

```
Predicted Topic:  2
```

	content	chap_sec	raw_content	content_length	ldamallet_topic
11	phasor_addition many situation necessary add t...	2.6	Phasor Addition \r\n There are many\r...	584	2
62	discrete_fourier_transform dft dtft discrete t...	8.2	Discrete Fourier Transform (DFT) \n\n\...	1180	2

Result 1.2: Mallet LDA:
("Explain continuous-to-discrete conversion.")

Predicted Topic: 0

	content	chap_sec	raw_content	content_length	ldamallet_topic
7	review sine cosine function sinusoidal signal ...	2.2	Review of Sine and Cosine Functions \r...	308	0
9	sampling plot sinusoid plot sinusoid chapter c...	2.4	Sampling and Plotting Sinusoids \r\n A...	406	0
20	spectrum fourier_series discuss spectrum secti...	3.5	Spectrum of the Fourier Series \r\n\r\...	204	0
26	summary link chapter introduce concept spectru...	3.11	Summary and Links \r\n This chapter in...	218	0
31	discrete continuous conversion purpose ideal d...	4.4	Discrete-to-Continuous Conversion \r\...	869	0
61	chapter discrete_fourier_transform chapter bui...	8.1	<h3><span class="chapter text-center">CHAPTER ...	161	0
66	spectrum analysis periodic_signal section show...	8.7	Spectrum Analysis of Periodic Signals ...	393	0

Result 1.2: Mallet LDA (“Explain C-to-D conversion.”)

```
Predicted Topic:  7
```

	content	chap_sec	raw_content	content_length	ldamallet_topic
18	periodic waveform periodic_signal satisfie con...	3.3	Periodic Waveforms \r\n A periodic s...	571	7
19	fourier_series example sec show synthesize per...	3.4	Fourier Series \r\n \r\n The example...	433	7
43	summary link chapter introduce concept fir_fil...	5.9	Summary and Links \r\n This chapter in...	75	7
56	property dtft motivate study dtft primarily co...	7.3	Properties of the DTFT \n\nWe have mot...	997	7
67	window since dft finite sum use analyze finite...	8.8	Windows \n\nSince the DFT is a finite ...	820	7
68	spectrogram see dft compute exact frequency do...	8.9	The Spectrogram \n\n We have seen that...	2306	7
70	summary link chapter introduce discrete_fourie...	8.11	Summary and Links In this chapter we ...	61	7
81	summary link transform method introduce chapte...	9.10	Summary and Links \r\n The \(z\)-trans...	176	7

Result 1.2: Mallet LDA (“What is phase difference?”)

```
Predicted Topic:  5
```

	content	chap_sec	raw_content	content_length	ldamallet_topic
1	mathematical representation signal signal patt...	1.1	<h3>Mathematical Representation of Signals</h3...	479	5
3	think system block diagram useful represent co...	1.3	Thinking About Systems \r\n Block diag...	229	5
12	physics tuning_fork section describe simple ex...	2.7	Physics of the Tuning Fork \r\n In\r\...	562	5
21	fourier_analysis periodic_signal synthesize pe...	3.6	Fourier Analysis of Periodic Signals ...	763	5
24	timefrequency_spectrum see wide range interest...	3.9	Time–Frequency Spectrum \r\n \r\...	709	5
59	table fourier_transform property pair chapter ...	7.6	<h3>Table of Fourier Transform Properties and ...	140	5
64	table discrete_fourier_transform property pair...	8.5	<h3>Table of Discrete Fourier Transform Proper...	158	5
65	spectrum analysis discrete periodic_signal cha...	8.6	<h3>Spectrum Analysis of Discrete Periodic Sig...	1839	5
85	system_function iir filter see chapter fir cas...	10.3	System Function of an IIR Filter \r\n ...	916	5
89	inverse transform application see three domain...	10.7	<h3>The Inverse \(z\)-Transform and Some Appli...	538	5

Result 1.2: Mallet LDA
(“Are calculators allowed for the exams?”)

Predicted Topic: 1

	content	chap_sec	raw_content	content_length	ldamallet_topic
33	summary link chapter introduce concept samplin...	4.6	Summary and Links \r\n This chapter in...	112	1
37	general fir_filter note special case general d...	5.3	The General FIR Filter \r\n Note that ...	1072	1
39	linear time invariant lti_system section discu...	5.5	Linear Time-Invariant (LTI) Systems \r...	473	1
42	example fir_filtering conclude chapter example...	5.8	Example of FIR Filtering \r\n \r\n We...	322	1
60	summary link chapter introduce discrete time f...	7.7	Summary and Links \n In this chapter w...	58	1
63	inherent periodicity_of xn dft section study p...	8.4	<h3>Inherent Periodicity of \(x[n]\) in the DF...	1320	1
72	definition transform finite length signal xn r...	9.1	Definition of the \(z\)-Transform \r\n...	266	1
73	transform linear system transform indispensab...	9.2	<h3>The \(z\)-Transform and Linear Systems</h3...	306	1
75	transform operator delay property state secti...	9.4	The \(z\)-Transform as an Operator \r\...	279	1

Result 1.3: Tomotopy LDA

Example Input: *“What are FIR Filters?”*

Predicted Topics(in order of most probable to least probable):

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #3	signal	represent	representation	time	example	value	equation	take	define	general
Topic #7	system	show	mathematical	diagram	form	may	although	square	operation	important
Topic #1	time	change	figure	analysis	section	would	constant	whose	many	notation
Topic #2	sum	sinusoidal	term	period	identity	periodic	integer	add	angle	obtain
Topic #4	chapter	point	different	use	concept	matlab	result	show	amplitude	cycle
Topic #5	time	sample	discrete	continuous	alias	output	sampling	xn	converter	input
Topic #11	frequency	spectrum	sinusoid	hz	component	line	show	signal	example	phase
Topic #6	plot	cos	signal	use	sec	case	also	original	simple	obtain
Topic #0	note	key	music	sound	vary	duration	much	spectrogram	play	msec

```
Topic With Highest Document Probability:  3
```

Result 1.3: Tomotopy LDA

Example Input: *"What are finite-impulse-response filters?"*

Predicted Topics(in order of most probable to least probable):

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #1	time	change	figure	analysis	section	would	constant	whose	many	notation
Topic #2	sum	sinusoidal	term	period	identity	periodic	integer	add	angle	obtain
Topic #6	plot	cos	signal	use	sec	case	also	original	simple	obtain
Topic #4	chapter	point	different	use	concept	matlab	result	show	amplitude	cycle
Topic #5	time	sample	discrete	continuous	alias	output	sampling	xn	converter	input
Topic #3	signal	represent	representation	time	example	value	equation	take	define	general
Topic #11	frequency	spectrum	sinusoid	hz	component	line	show	signal	example	phase
Topic #0	note	key	music	sound	vary	duration	much	spectrogram	play	msec

```
Topic With Highest Document Probability:  1
```

Result 1.3: Tomotopy LDA

Example Input: *“Are calculators allowed for the exams?”*

Predicted Topics(in order of most probable to least probable):

	Keyword #1	Keyword #2	Keyword #3	Keyword #4	Keyword #5	Keyword #6	Keyword #7	Keyword #8	Keyword #9	Keyword #10
Topic #7	system	show	mathematical	diagram	form	may	although	square	operation	important
Topic #18	function	cosine	call	variable	period	note	two	denote	sinusoidal	parameter
Topic #15	phasor	sine	show	cosine	use	two	real_part	define	imaginary_part	multiply
Topic #3	signal	represent	representation	time	example	value	equation	take	define	general
Topic #6	plot	cos	signal	use	sec	case	also	original	simple	obtain
Topic #1	time	change	figure	analysis	section	would	constant	whose	many	notation
Topic #2	sum	sinusoidal	term	period	identity	periodic	integer	add	angle	obtain
Topic #5	time	sample	discrete	continuous	alias	output	sampling	xn	converter	input
Topic #10	formula	since	sinusoid	must	phase_shift	positive	second	equal	find	ak
Topic #4	chapter	point	different	use	concept	matlab	result	show	amplitude	cycle
Topic #0	note	key	music	sound	vary	duration	much	spectrogram	play	msec

Topic With Highest Document Probability: 7

Result 2: CorEx

```
Original input:  What are FIR filters?
Processed input:  fir filter
Latent topics and probabilities:
Topic 79 has a probability of 99.9999%.
Topic 37 has a probability of 99.9999%.
Topic 73 has a probability of 99.9999%.
Topic 34 has a probability of 99.9999%.
Topic 81 has a probability of 99.9999%.
```

	chap_sec	content	book_keywords	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Word #8	Word #9	Word #10
28	4.1	sampling si...	sampling di...	signal	frequency	spectrum	sinusoid	output	fourier	filter	period	sample	sinusoidal
34	5.0	fir filter ...	chapter fir...	filter	output	fir	system	sequence	input	response	lti	linear	link
35	5.1	discrete_ti...	discrete_ti...	filter	output	fir	system	input	sequence	signal	sampling	sample	response
36	5.2	running_ave...	running ave...	filter	output	fir	system	signal	sequence	input	response	property	plot
37	5.3	general fir...	weighted ru...	signal	filter	fir	output	frequency	system	sinusoid	sinusoidal	sequence	input
38	5.4	implementat...	multiplier ...	filter	output	fir	system	signal	sequence	input	response	lti	sample
39	5.5	linear time...	linear time...	filter	output	fir	system	sequence	input	response	lti	signal	delay
40	5.6	convolution...	convolution...	output	filter	fir	system	sequence	input	response	lti	property	signal
41	5.7	cascaded lt...	cascaded lt...	filter	fir	system	output	input	response	sequence	lti	convolution	linear
42	5.8	example fir...	fir filteri...	system	filter	fir	output	input	response	signal	sequence	lti	sampled
43	5.9	summary lin...	fir cd_rom ...	fir	filter	lab	sinusoid	sinusoidal	sample	chapter	convolution	system	filtering
44	6.0	frequency_r...	sinusoid fr...	fir	filter	system	input	output	signal	frequency_r...	sinusoid	response	lti
45	6.1	sinusoidal_...	sinusoid re...	filter	fir	output	system	input	response	frequency_r...	signal	frequency	sequence
46	6.2	superpositi...	superpositi...	filter	frequency	fir	signal	sinusoid	output	system	sinusoidal	input	response
47	6.3	steady_stat...	steady stat...	filter	fir	output	system	input	response	sequence	frequency_r...	lti	linear
48	6.4	property fr...	periodicity...	filter	fir	input	output	system	frequency_r...	periodic	response	plotting	function
49	6.5	graphical r...	delay syste...	frequency	signal	filter	fir	sinusoid	system	output	period	periodic	function
50	6.6	cascaded lt...	cascaded lt...	fir	filter	system	input	output	frequency_r...	response	lti	convolution	impulse_res...

Result 2: CorEx

```
Original input:  What are finite-impulse-response filters?
Processed input:  finite_impulse_response filter
Latent topics and probabilities:
Topic 79 has a probability of 99.9999%.
Topic 34 has a probability of 99.9999%.
Topic 81 has a probability of 99.9999%.
Topic 73 has a probability of 99.9999%.
```

	chap_sec	content	book_keywords	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Word #8	Word #9	Word #10
28	4.1	sampling si...	sampling di...	signal	frequency	spectrum	sinusoid	output	fourier	filter	period	sample	sinusoidal
34	5.0	fir filter ...	chapter fir...	filter	output	fir	system	sequence	input	response	lti	linear	link
35	5.1	discrete_ti...	discrete_ti...	filter	output	fir	system	input	sequence	signal	sampling	sample	response
36	5.2	running_ave...	running ave...	filter	output	fir	system	signal	sequence	input	response	property	plot
37	5.3	general fir...	weighted ru...	signal	filter	fir	output	frequency	system	sinusoid	sinusoidal	sequence	input
38	5.4	implementat...	multiplier ...	filter	output	fir	system	signal	sequence	input	response	lti	sample
39	5.5	linear time...	linear time...	filter	output	fir	system	sequence	input	response	lti	signal	delay
40	5.6	convolution...	convolution...	output	filter	fir	system	sequence	input	response	lti	property	signal
41	5.7	cascaded lt...	cascaded lt...	filter	fir	system	output	input	response	sequence	lti	convolution	linear
42	5.8	example fir...	fir filteri...	system	filter	fir	output	input	response	signal	sequence	lti	sampled
43	5.9	summary lin...	fir cd_rom ...	fir	filter	lab	sinusoid	sinusoidal	sample	chapter	convolution	system	filtering
44	6.0	frequency_r...	sinusoid fr...	fir	filter	system	input	output	signal	frequency_r...	sinusoid	response	lti
45	6.1	sinusoidal_...	sinusoid re...	filter	fir	output	system	input	response	frequency_r...	signal	frequency	sequence
46	6.2	superpositi...	superpositi...	filter	frequency	fir	signal	sinusoid	output	system	sinusoidal	input	response
47	6.3	steady_stat...	steady stat...	filter	fir	output	system	input	response	sequence	frequency_r...	lti	linear
48	6.4	property fr...	periodicity...	filter	fir	input	output	system	frequency_r...	periodic	response	plotting	function
49	6.5	graphical r...	delay syste...	frequency	signal	filter	fir	sinusoid	system	output	period	periodic	function
50	6.6	cascaded lt...	cascaded lt...	fir	filter	system	input	output	frequency_r...	response	lti	convolution	impulse_res...
51	6.7	running_ave...	running_ave...	frequency	filter	signal	fir	output	system	period	input	frequency_r...	function
52	6.8	filtering s...	filtered sa...	signal	frequency	filter	fir	output	sinusoid	system	sample	sampling	sampled
53	6.9	summary lin...	filtering l...	sinusoid	filter	sinusoidal	fir	system	frequency	signal	input	lti	frequency_r...

Result 2: CorEx

```
Original input:  Explain continuous-to-discrete conversion.
Processed input:  explain continuous_to_discrete conversion
Latent topics and probabilities:
Topic 28 has a probability of 99.9999%.
Topic 3 has a probability of 10.087221364130325%.
```

	chap_sec	content	book_keywords	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Word #8	Word #9	Word #10
1	1.1	mathematica...	signal time...	signal	sample	sampled	sampling	frequency	period	function	plot	plotting	time
2	1.2	mathematica...	system defi...	output	signal	sequence	input	system	response	block	diagram	operation	operator
3	1.3	thinking sy...	cd audio sy...	output	input	response	signal	block	diagram	converter	conversion	sample	sampled
6	2.1	tuning_fork...	tuning fork...	signal	frequency	sinusoid	sinusoidal	spectrum	period	function	sampled	plot	plotting
10	2.5	complex exp...	complex exp...	signal	frequency	sinusoid	spectrum	sinusoidal	function	period	plot	amplitude	plotting
11	2.6	phasor_addi...	phasor addi...	signal	frequency	sinusoid	sinusoidal	function	amplitude	period	plot	plotting	spectrum
24	3.9	timefrequen...	time ndash ...	signal	frequency	spectrum	sinusoid	fourier	sinusoidal	period	periodic	function	amplitude
27	4.0	sampling al...	reconstruct...	sampling	sample	frequency	spectrum	sampled	signal	reconstruction	conversion	reconstructed	fourier
28	4.1	sampling si...	sampling di...	signal	frequency	spectrum	sinusoid	output	fourier	filter	period	sample	sinusoidal
29	4.2	spectrum vi...	sampling sp...	frequency	signal	spectrum	sinusoid	sinusoidal	sampling	sample	sampled	output	function
30	4.3	strobe demo...	strobe rota...	frequency	signal	spectrum	sinusoid	sampling	sample	sampled	sinusoidal	plot	amplitude
31	4.4	discrete_to...	discrete_to...	frequency	signal	sinusoid	sampling	sample	sampled	spectrum	sinusoidal	function	output
32	4.5	sampling_th...	shannon sam...	signal	frequency	spectrum	sampling	sample	sampled	sinusoid	periodic	fourier	sinusoidal
51	6.7	running_ave...	running_ave...	frequency	filter	signal	fir	output	system	period	input	frequency_r...	function
52	6.8	filtering s...	filtered sa...	signal	frequency	filter	fir	output	sinusoid	system	sample	sampling	sampled
61	8.1	discrete fo...	sampled sig...	spectrum	frequency	signal	periodic	fourier	dft	transform	dtft	discrete_time	period
68	8.9	spectrogram...	time_depend...	signal	frequency	spectrum	sinusoid	fourier	period	sinusoidal	periodic	sampled	sampling

Result 2: CorEx

```
Original input:  Explain C-to-D conversion.
Processed input:  explain c_to_d conversion
Latent topics and probabilities:
Topic 28 has a probability of 99.9999%.
Topic 3 has a probability of 10.087221364130325%.
```

	chap_sec	content	book_keywords	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Word #8	Word #9	Word #10
1	1.1	mathematica...	signal time...	signal	sample	sampled	sampling	frequency	period	function	plot	plotting	time
2	1.2	mathematica...	system defi...	output	signal	sequence	input	system	response	block	diagram	operation	operator
3	1.3	thinking sy...	cd audio sy...	output	input	response	signal	block	diagram	converter	conversion	sample	sampled
6	2.1	tuning_fork...	tuning fork...	signal	frequency	sinusoid	sinusoidal	spectrum	period	function	sampled	plot	plotting
10	2.5	complex exp...	complex exp...	signal	frequency	sinusoid	spectrum	sinusoidal	function	period	plot	amplitude	plotting
11	2.6	phasor_addi...	phasor addi...	signal	frequency	sinusoid	sinusoidal	function	amplitude	period	plot	plotting	spectrum
24	3.9	timefrequen...	time ndash ...	signal	frequency	spectrum	sinusoid	fourier	sinusoidal	period	periodic	function	amplitude
27	4.0	sampling al...	reconstruct...	sampling	sample	frequency	spectrum	sampled	signal	reconstruction	conversion	reconstructed	fourier
28	4.1	sampling si...	sampling di...	signal	frequency	spectrum	sinusoid	output	fourier	filter	period	sample	sinusoidal
29	4.2	spectrum vi...	sampling sp...	frequency	signal	spectrum	sinusoid	sinusoidal	sampling	sample	sampled	output	function
30	4.3	strobe demo...	strobe rota...	frequency	signal	spectrum	sinusoid	sampling	sample	sampled	sinusoidal	plot	amplitude
31	4.4	discrete_to...	discrete_to...	frequency	signal	sinusoid	sampling	sample	sampled	spectrum	sinusoidal	function	output
32	4.5	sampling_th...	shannon sam...	signal	frequency	spectrum	sampling	sample	sampled	sinusoid	periodic	fourier	sinusoidal
51	6.7	running_ave...	running_ave...	frequency	filter	signal	fir	output	system	period	input	frequency_r...	function
52	6.8	filtering s...	filtered sa...	signal	frequency	filter	fir	output	sinusoid	system	sample	sampling	sampled
61	8.1	discrete fo...	sampled sig...	spectrum	frequency	signal	periodic	fourier	dft	transform	dtft	discrete_time	period
68	8.9	spectrogram...	time_depend...	signal	frequency	spectrum	sinusoid	fourier	period	sinusoidal	periodic	sampled	sampling

Result 2: CorEx

```
Original input:  What is phase difference?
Processed input:  phase difference
Latent topics and probabilities:
Topic 49 has a probability of 99.9999%.
```

	chap_sec	content	book_keywords	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Word #8	Word #9	Word #10
6	2.1	tuning_fork...	tuning fork...	signal	frequency	sinusoid	sinusoidal	spectrum	period	function	sampled	plot	plotting
7	2.2	review sine...	trigonometr...	plotting	function	plotted	plot	frequency	sinusoid	period	frequency_r...	periodicity	phase
8	2.3	sinusoidal_...	amplitude p...	frequency	signal	period	sinusoid	periodic	sinusoidal	spectrum	function	plot	amplitude
9	2.4	sampling pl...	sample spac...	frequency	signal	sinusoid	sampled	spectrum	sampling	period	sinusoidal	periodic	sample
10	2.5	complex exp...	complex exp...	signal	frequency	sinusoid	spectrum	sinusoidal	function	period	plot	amplitude	plotting
11	2.6	phasor_addi...	phasor addi...	signal	frequency	sinusoid	sinusoidal	function	amplitude	period	plot	plotting	spectrum
12	2.7	physic tuni...	physic tuni...	signal	sinusoid	frequency	sinusoidal	amplitude	function	sample	phase	tuning_fork	spectrum
14	2.9	summary lin...	lab matlab ...	sinusoid	sinusoidal	signal	frequency	amplitude	function	phase	plot	tuning_fork	plotted
15	3.0	spectrum re...	spectrum de...	signal	frequency	sinusoid	sinusoidal	amplitude	spectrum	fourier	phase	tuning_fork	complex
16	3.1	spectrum su...	negative fr...	signal	frequency	sinusoid	fourier	spectrum	sinusoidal	period	periodic	function	amplitude
17	3.2	beat_notes ...	beat note a...	signal	frequency	sinusoid	sinusoidal	spectrum	function	plot	amplitude	plotting	plotted
18	3.3	periodic wa...	period fund...	frequency	signal	spectrum	period	sinusoid	sinusoidal	periodic	function	plot	fourier
19	3.4	fourier ser...	fourier ana...	signal	frequency	sinusoid	fourier	spectrum	period	sinusoidal	periodic	function	amplitude
20	3.5	spectrum fo...	fourier ser...	frequency	spectrum	fourier	signal	periodic	series	complex	period	formula	sum
21	3.6	fourier ana...	square wave...	signal	frequency	spectrum	sinusoid	fourier	sinusoidal	period	periodic	function	plot
24	3.9	timefrequen...	time ndash ...	signal	frequency	spectrum	sinusoid	fourier	sinusoidal	period	periodic	function	amplitude
25	3.10	frequency_m...	frequency m...	frequency	signal	sinusoid	sinusoidal	spectrum	amplitude	function	plot	plotted	phase
28	4.1	sampling si...	sampling di...	signal	frequency	spectrum	sinusoid	output	fourier	filter	period	sample	sinusoidal
29	4.2	spectrum vi...	sampling sp...	frequency	signal	spectrum	sinusoid	sinusoidal	sampling	sample	sampled	output	function
30	4.3	strobe demo	strobe rota	frequency	signal	spectrum	sinusoid	sampling	sample	sampled	sinusoidal	plot	amplitude

Result 2: CorEx

```
Original input:  Are calculators allowed for the exams?
Processed input:  calculator allowed exam
Latent topics and probabilities:
```

chap_sec	content	book_keywords	Word #1	Word #2	Word #3	Word #4	Word #5	Word #6	Word #7	Word #8	Word #9	Word #10

Conclusion

- Difficult to form many topics without additional data or specified anchored words
- Still need to find the sweet spot between generality vs specificity
  - Number of Topics vs Number of Documents per Topic
- User query could be way shorter than the training documents
  - Not enough tokens in user query to assign it topics with enough accuracy

Merits and Plans for Future Semesters

- Apply LDA (and its variants) and CorEx to Piazza data
- Integrate the improved data preprocess to the chatbot
- Add the topic labeling to the main chatbot program and evaluate the performance
  - Does looking only at documents of the same topics speed up the program
- CorEx
  - Explore more on the Hierarchical Topic Modeling
  - Find the optimal number of topics

Works Cited

- Gallagher, R. J., Reing, K., Kale, D., and Ver Steeg, G. "Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Transactions of the Association for Computational Linguistics (TACL), 2017.