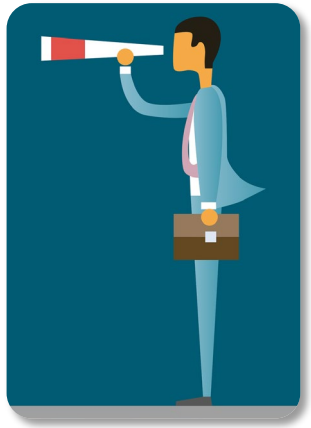
Unlocking Insights: Optimizing Your Text Data with Machine Learning

May 31, 2024

Xiaoxia Zhang, higher education doctoral student
Yih Tsao, higher education doctoral student
Brendan Dugan, Research Associate, NSSE, Indiana University

INDIANA UNIVERSITY

Overview



1. **Introduction: The Needs of ML Techniques in Text Data Analysis**
2. **Context of Case Studies: NSSE Topical Modules**
   - First-Year Experiences & Senior Transitions
   - Coping with Covid
3. **Data Analysis & Visualization of Case Studies**
   - Text Data Analysis Process and Techniques
4. **Conclusions & Implications**
5. **Demonstration and Steps to *Create Your Own***
   - R code and Packages
   - Workflow (steps)
6. **Discussion & Questions**

The Needs of ML Techniques in Text Data Analysis

Text data analysis (*Nawaz et al., 2022*)

- **Rich Detail**: Extensive information, including nuances, context, and specifics that may not be captured in numerical or structured data formats.
- **Actionable Insights**: To make informed decisions, develop strategies, and improve processes based on the findings extracted from textual sources.

Overall Challenges of analyzing text data (*Ittoo et al., 2016, Nawaz et al., 2022*)

- Scale and Scope Limitations
- Bias and Inconsistency
- Thematic analysis of large text data (*Braun & Clarke, 2023*)
  - Handling extensive textual data
  - Dealing with thematic diversity
  - Distinguishing between topic summaries and thematic codes

The Needs of ML Techniques in Text Data Analysis

Machine Learning techniques: Natural Language Processing (NLP) (*Roberts et al., 2014*)

1. ***Efficient Information Extraction*:** NLP efficiently identifies patterns, sentiments, and key themes from open-ended responses to uncover actionable insights and make informed decisions.

2. ***Time and Cost Reduction***: NLP significantly reduces the time and cost of manual analysis processes by automating the text analysis.

3. ***Advanced Analysis for Large Datasets***: One of the techniques for analyzing themes and patterns from large student response datasets, like NSSE, requires advanced NLP techniques.

Context of Case Studies

NSSE Topic Modules

First-Year Experiences & Senior Transitions

Coping With Covid (2020-21 only)

Case 1: Senior Transitions Topical Module

1. **Objective**:
   - This module includes separate sets of items for first-year students and seniors based on institution-reported class. The senior items explore post-graduation plans, links between the academic major and future plans, and confidence in skill development.
2. **Open-ended Question**:
   - ***Is there anything your institution could have done better to prepare you for your career or further education? Please describe.*** [text box]
3. **Case 1 Research Question:**
   - What are the differences in responses between **women of color in STEM** and other students regarding their institution's preparation for career or further education?

Case 1: Senior Transition Module – Descriptive Data

Word Counts

Module	min	p25	mdn	mean	p75	max
FYS	1	7	20	36.1	45	888

Responses

Module	Institutions	Comments	Not WoC STEM	WoC STEM
FYS	52	4896	281	4615

Case 2: Coping with COVID Topical Module (2021)

1. **Objective**:
   - To explores the impact of the COVID-19 pandemic on students' educational experiences, mental wellness, and everyday life experiences, including:
     - Perceptions of faculty and institutional responses, disruptions to educational plans, living situation details, stressors and negative emotional experiences, and changes in leisure activities and time demands are addressed.
2. **Open-ended Question**:
   - ***Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies***. [TEXT BOX]
3. **Case 2 Research Question:**
   - How do **international students** and **domestic students** differ in their feedback regarding their experiences as students on the response of their institutions during the COVID-19 pandemic?

Case 2: Coping with COVID Module – Descriptive Data

Word Counts

Module	min	p25	mdn	mean	p75	max
COV	1	15	34	51.9	66	872

Responses

Module	Institutions	Comments	International	Domestic
COV	51	3312	115	3197

Data Analysis & Visualization of Two Study Cases

NSSE data set and NLP in R

Review of Practices

1. Structural Topical Model overview
   - Mixed-membership, unsupervised machine learning method for text topic modelling (NLP)
   - Similar to LDA, but allows author metadata (“treatment”) to influence topical prevalence, topical content, or **both**
   - What is being discussed? How is it being discussed? Do groups discuss different topics, or discuss topics in different ways?
   - "The most important user input in parametric topic models is the number of topics [*k*]. There is no right answer to the appropriate number of topics." - `stm` package authors
   - Short, focused corpora: 3-10 topics
     Small corpora: 5-50 topics
     Medium sized corpora (10k to 100k documents): 60-100 topics

Review of Practices – NLP terms

1. **Tokens**: basic unit. Can be terms (words or features), n-grams, sentences, paragraphs
2. **Documents**: the text. Can be comments, reviews, tweets, books, chapters
3. **Beta**: probability of appearing in a topic assigned to each word
4. **FREX**: FRequent and EXclusive words in a topic
5. **Gamma**: Probability of a topic assigned to each document
6. **Metadata**: i.e. treament, covariate. Information about the author that might affect text

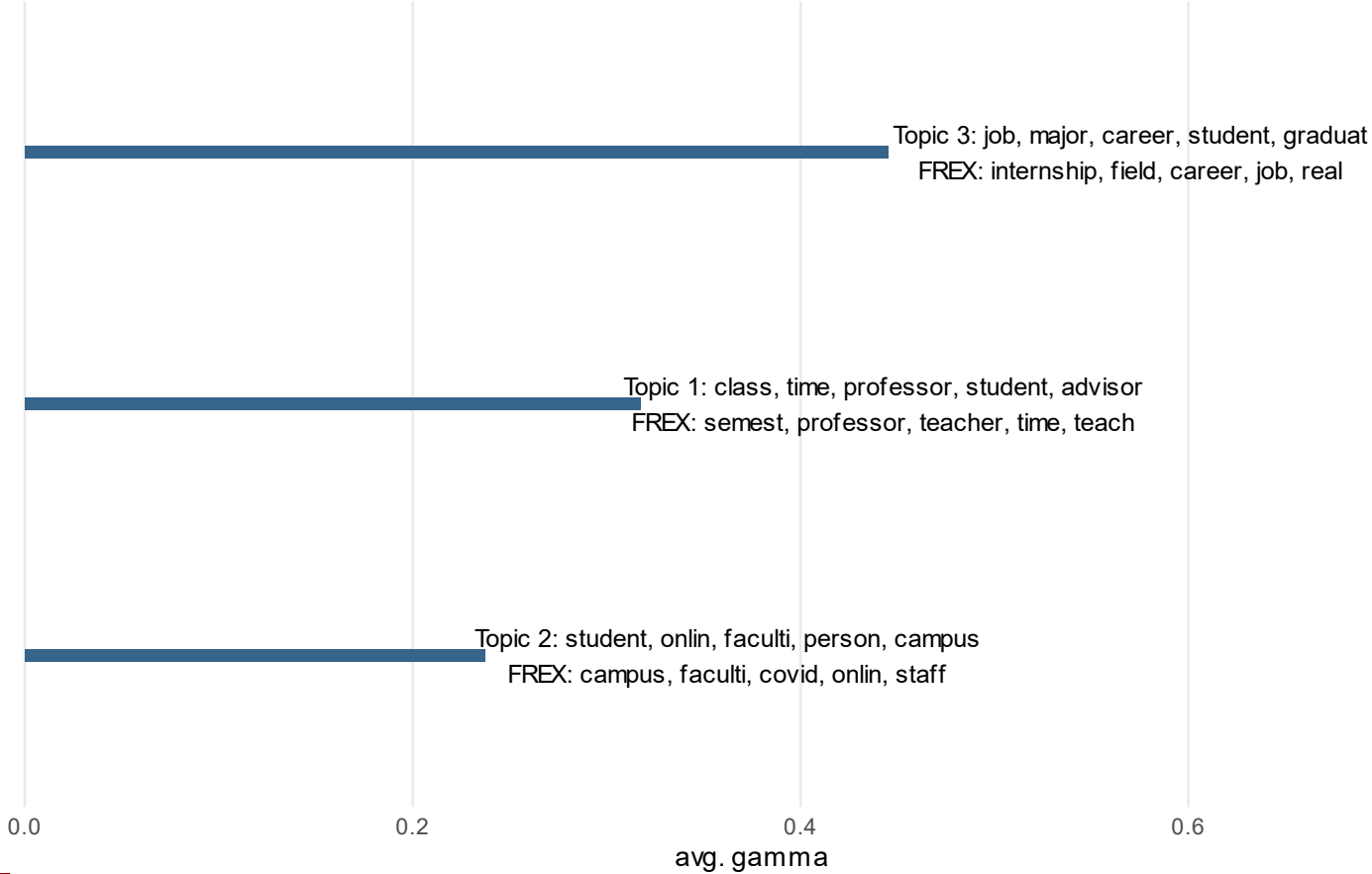
Review of Practices – NLP basics

1. "The parking at IU is terrible, but the professors are great! ☺"
2. **Tokenize**: "The“ "parking“ "at“ "IU" "is“ "terrible“ "but“ "the“ "professors“ "are“ "great“
3. **Remove stopwords**: "~~The~~“ "parking“ "~~at~~“ "~~IU~~" "~~is~~“ "terrible“ "~~but~~“ "~~the~~“ "professors“ "~~are~~“ "great“
4. **Stem**: “park~~*ing*~~” “terribl~~e~~” “professor~~s~~” “great”
5. **Restructure to document-feature matrix**:
6. **Model**!

```
Document-feature matrix of: 1 document,
       features
docs    park terribl professor great
  text1    1       1         1     1
```

Top 5 terms in 3 topics in NSSE Senior Transitions Topical Module

Is there anything your institution could have done better to prepare you for your career or further education?



Exemplar documents from topics in NSSE Senior Transitions Topical Module

Is there anything your institution could have done better to prepare you for your career or further education?

Topic 3 : job, major, career, student, graduat

(Doc ID 3427) more know ledge on careers in my field of supply chain. Also, access to more more support in gathering information on available internships, rotational programs, and other programs offered by companies. Handshake is great, but some form of tailored opportunities for specific majors w ould be great. My college, the College of Business, did a lo...

Topic 1 : class, time, professor, student, advisor

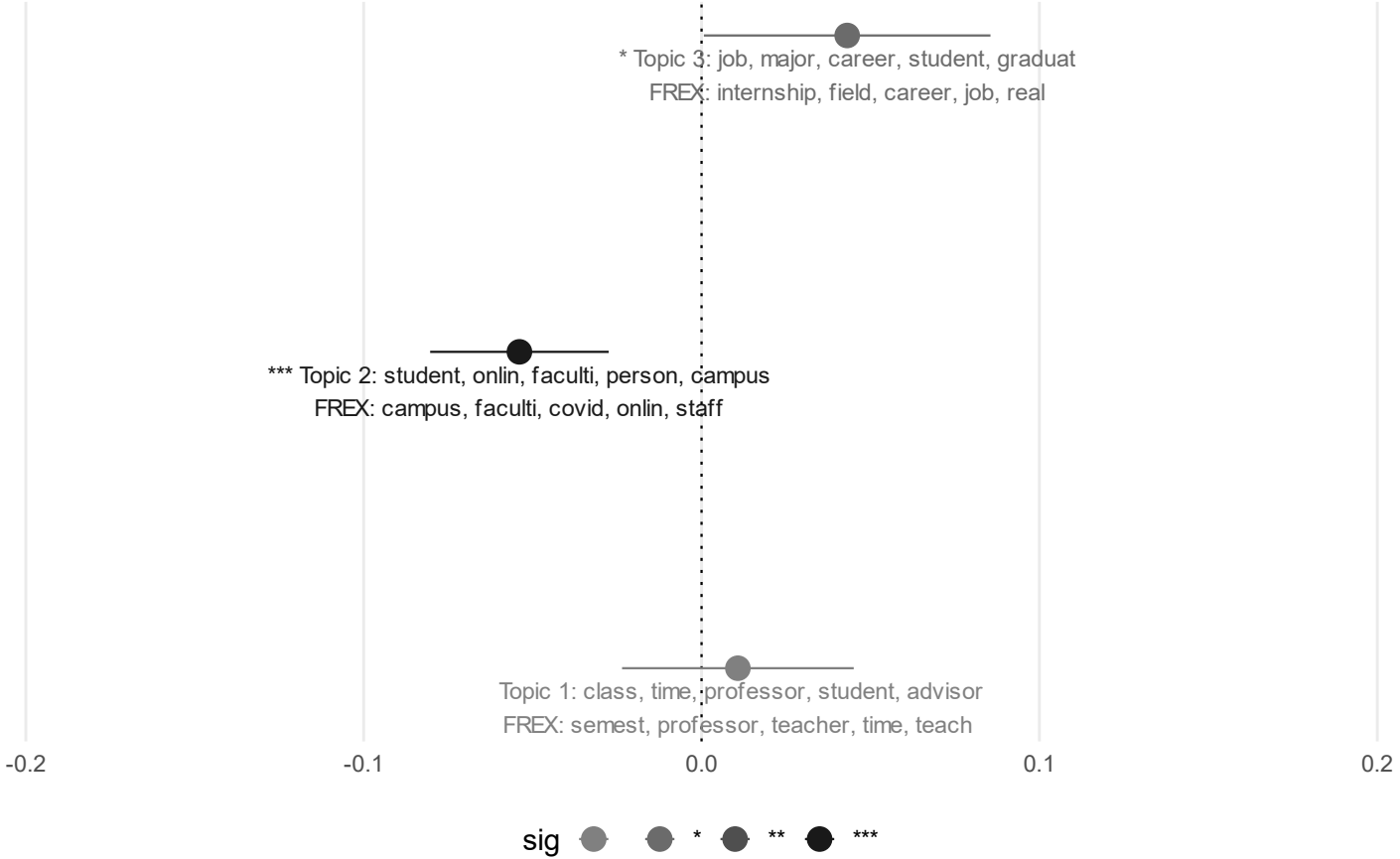
(Doc ID 1746) I honestly w ould of liked to see professors taking the time to brush up on their skill sets of w hatever it w as they w ere planning/having to teach. It w as very common for these professors to blow through an enormous amount of material but a lot of the information is outdated and not really used in the real w orld anymore. I understand if they ...

Topic 2 : student, onlin, faculti, person, campus

(Doc ID 795) Yes, [INSTITUTION] could have made greater steps for equality in the educational setting and learning environment. Rather than putting clear guidelines for how teachers are supposed to w ork w ith students w ho have disabilities during a pandemic and supporting said students in an online learning environment, [INSTITUTION] has come out w ith w ishy w ashy and mixed me...

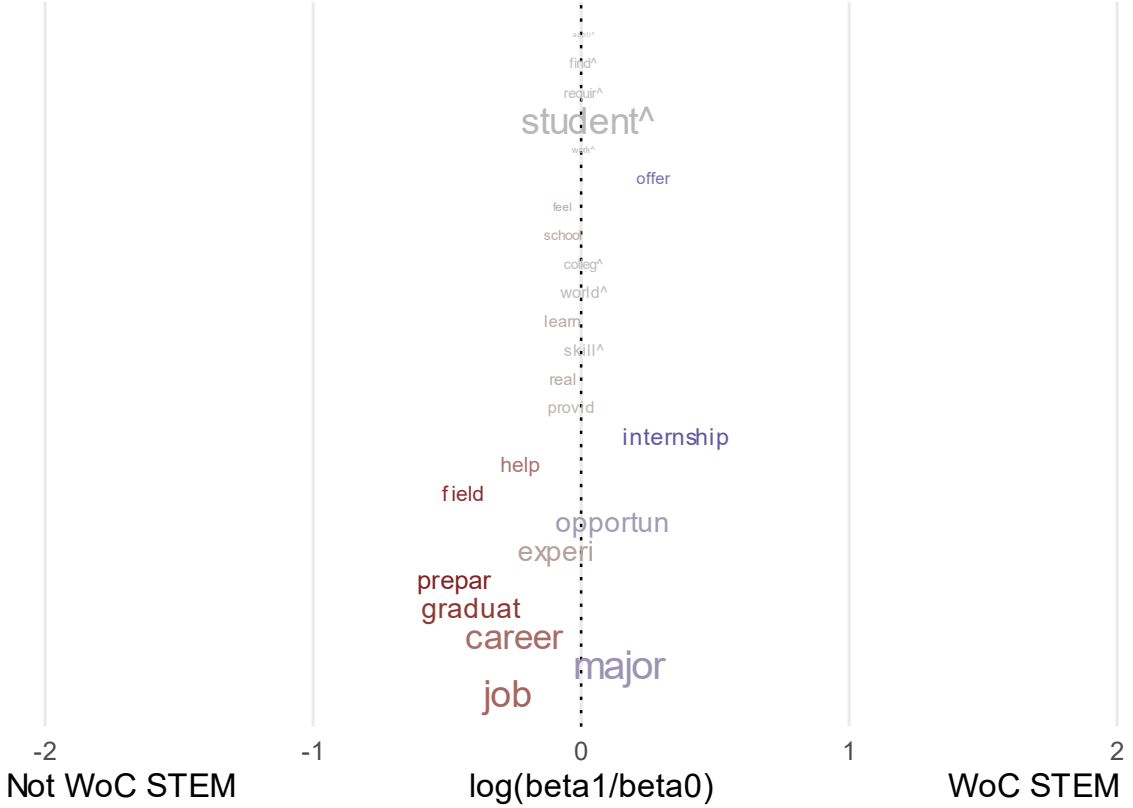
Estimated WoC STEM effects for 3 topics in NSSE Senior Transitions Topical Module

Is there anything your institution could have done better to prepare you for your career or further education?



Top 20 terms in topic 3 in NSSE Senior Transitions Topical Module

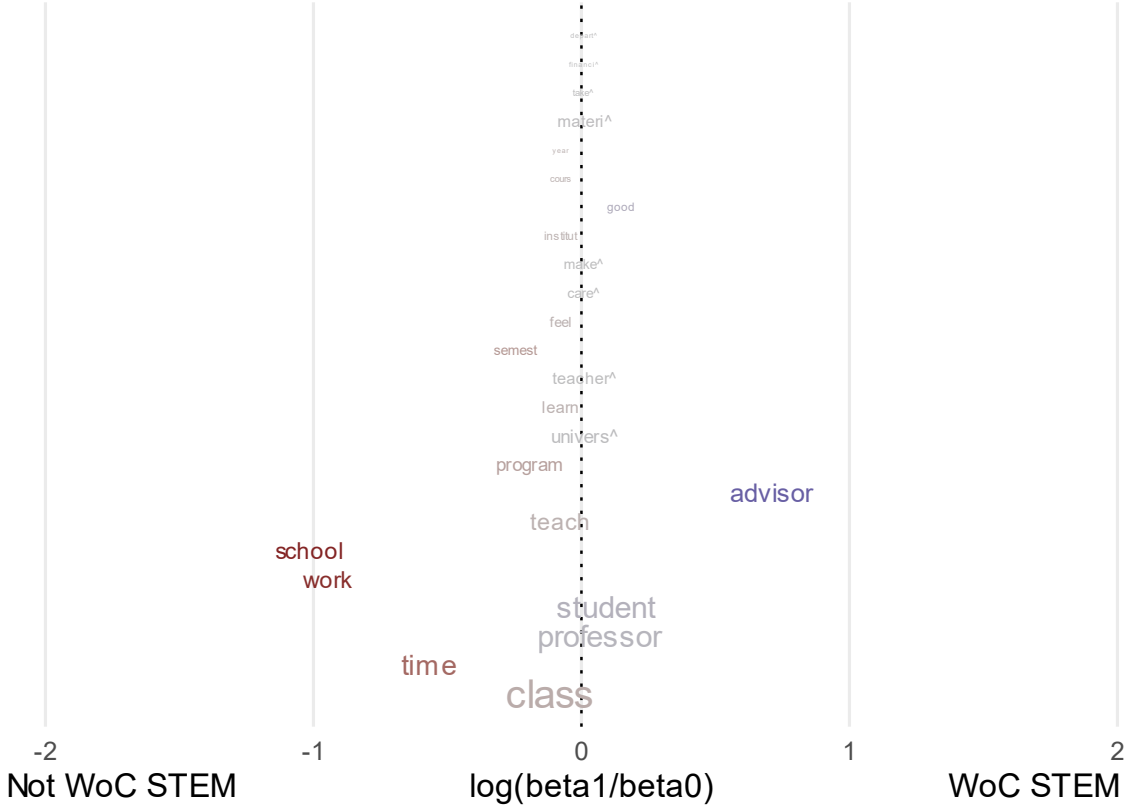
Is there anything your institution could have done better to prepare you for your career or further education?



Note: vertical alignment is random. Terms unique to one group denoted w ith ^.

Top 20 terms in topic 1 in NSSE Senior Transitions Topical Module

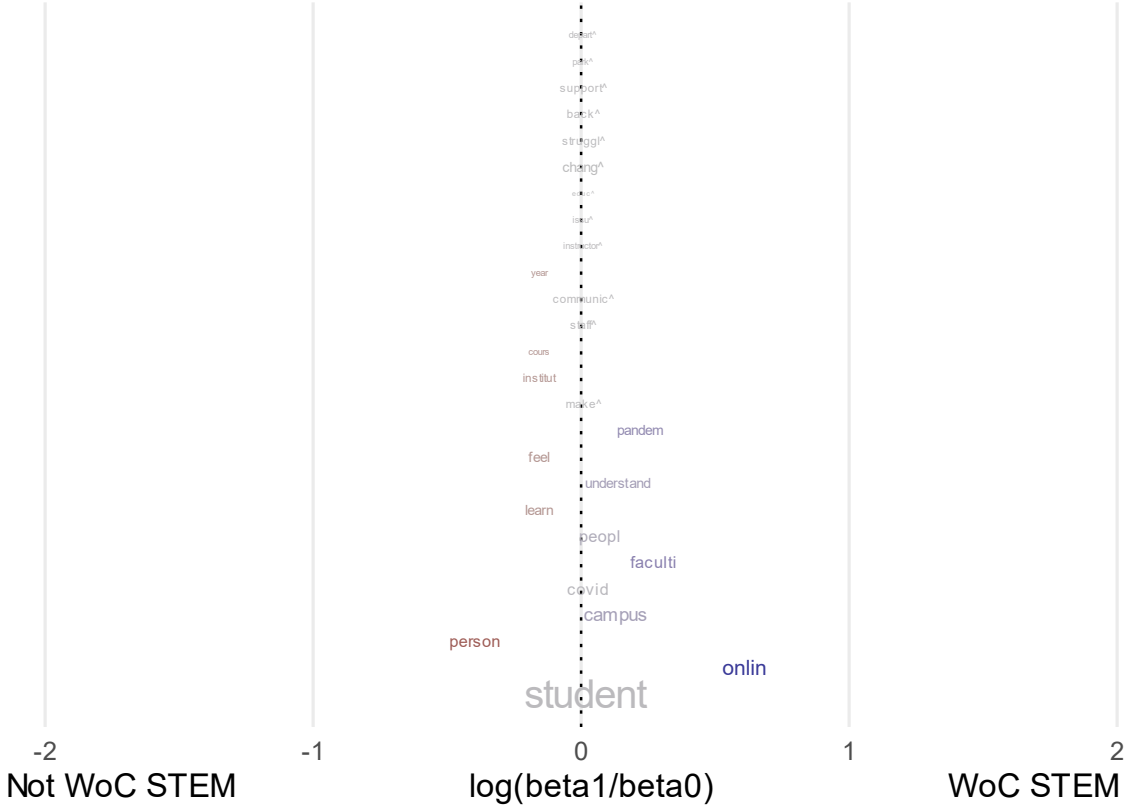
Is there anything your institution could have done better to prepare you for your career or further education?



Note: vertical alignment is random. Terms unique to one group denoted w ith ^.

Top 20 terms in topic 2 in NSSE Senior Transitions Topical Module

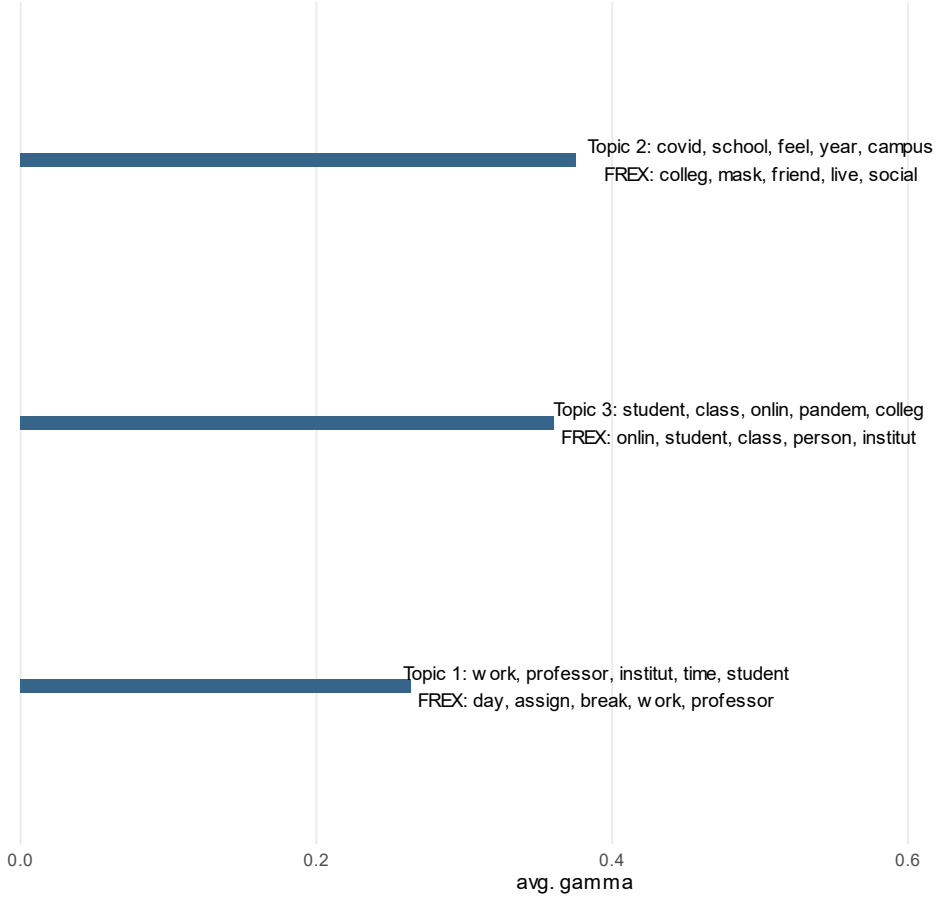
Is there anything your institution could have done better to prepare you for your career or further education?



Note: vertical alignment is random. Terms unique to one group denoted w ith ^.

Top 5 terms in 3 topics in NSSE Coping with Covid Topical Module

Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.



Exemplar documents from topics in NSSE Coping with Covid Topical Module

Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.

Topic 2 : covid, school, feel, year, campus

(Doc ID 2485) I fond the policies with dorm housing stupid. You can't visit with others unless in the lounge which is stupid if you and your friend wanna just chill or watch a movie or play video games or something. It's also stupid because I can't have my actual family members or friends even visit when they have the chance because of Covid policies and ...

Topic 3 : student, class, onlin, pandem, colleg

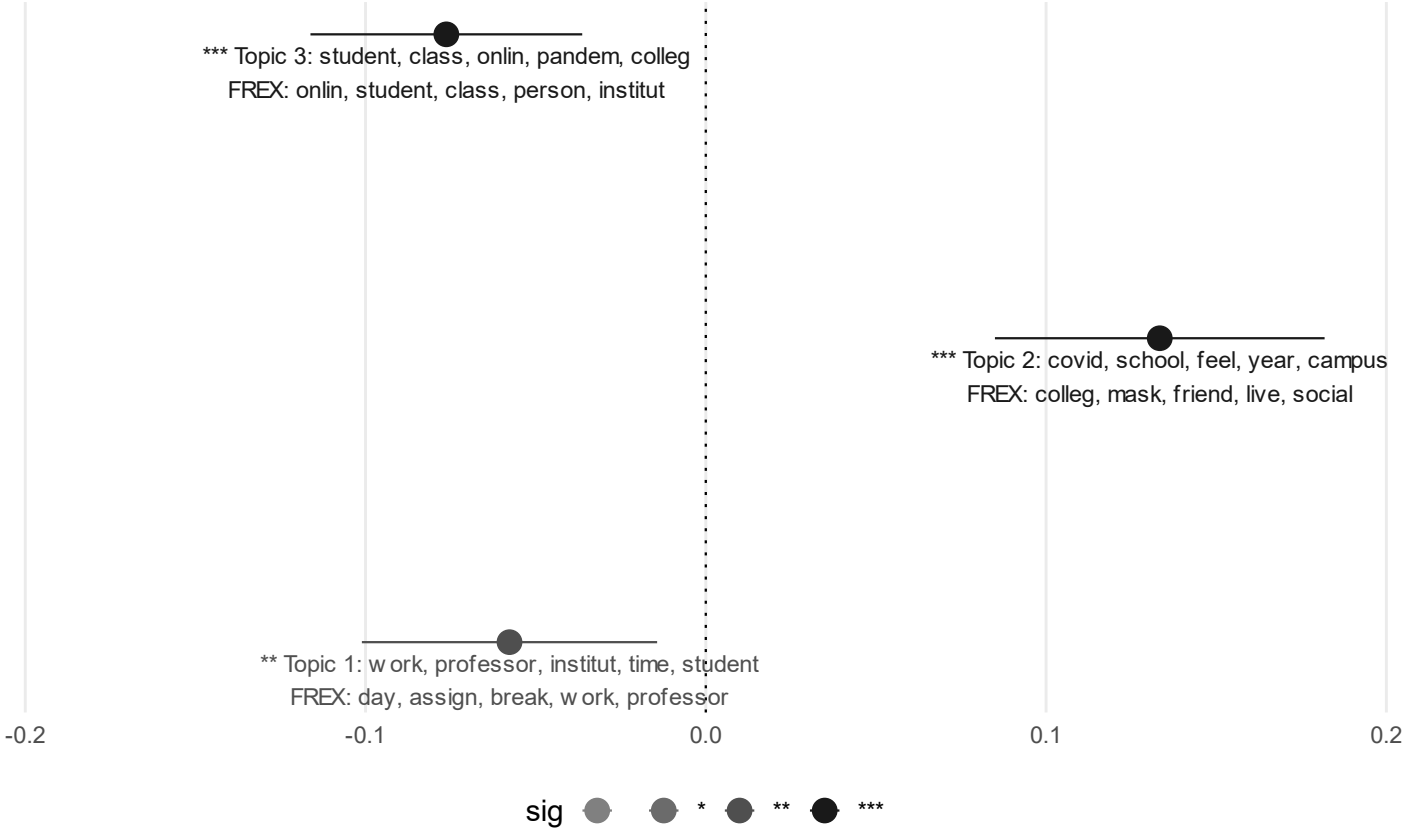
(Doc ID 2351) I still can not get over the fact I'm paying the full price tuition for IN PERSON learning when all of my classes have been fully online. This is especially frustrating because in my major you DO NOT get the same quality of learning online as you do in person. That might work for a liberal arts degree but not for a bachelors of science. I pa...

Topic 1 : work, professor, institut, time, student

(Doc ID 1144) With how this year's academic schedule was set up, students only had a week long break during the week of Thanksgiving, which was a week AFTER many courses had ended. Additionally, for this spring semester, there is no actual break. We get 2 "wellness days" on Feb 23 and March 24. While teachers could not assign homework for those 2 days, th...

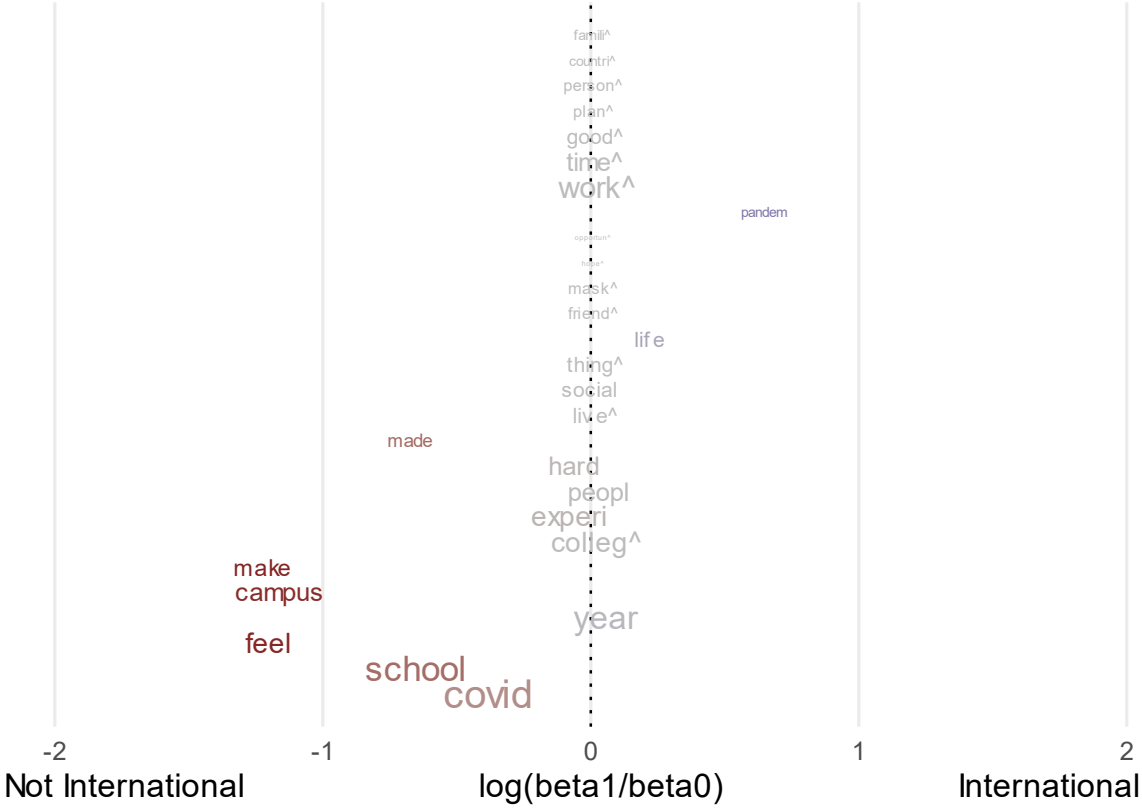
Estimated international student effects for 3 topics in NSSE Coping with Covid Topical Module

Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.



Top 20 terms in topic 2 in NSSE Coping with Covid Topical Module

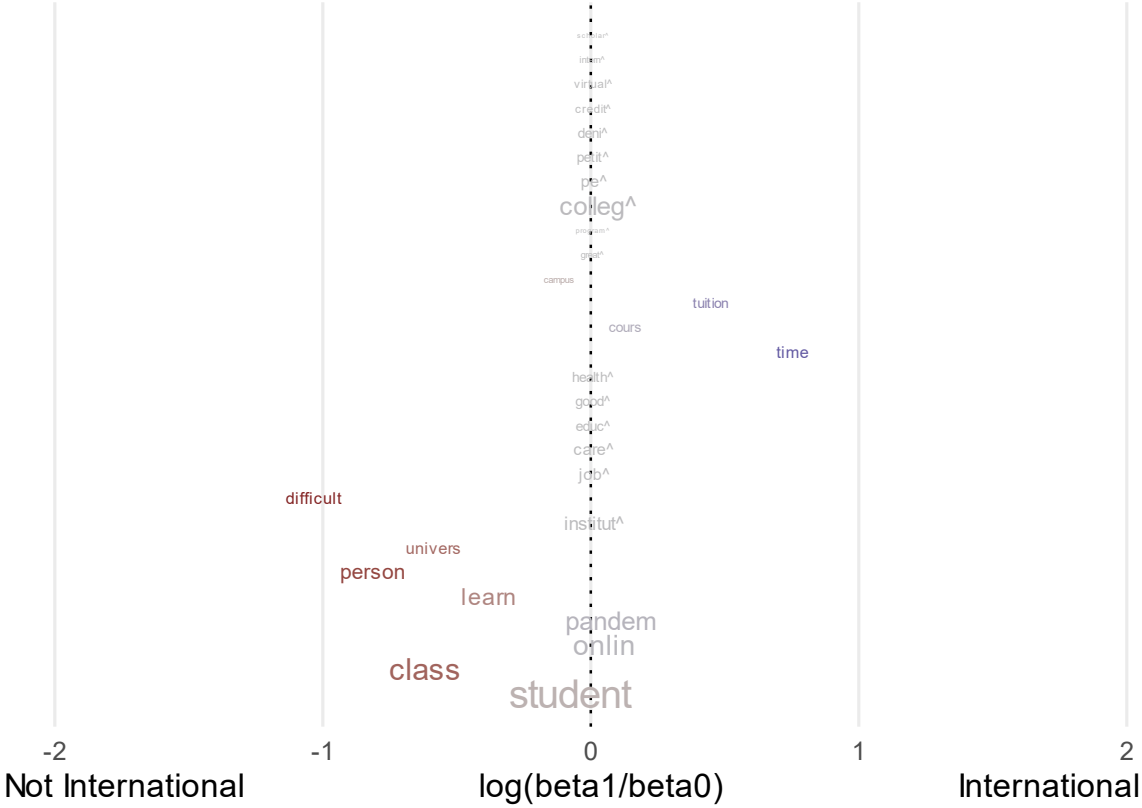
Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.



Note: vertical alignment is random. Terms unique to one group denoted w ith ^.

Top 20 terms in topic 3 in NSSE Coping with Covid Topical Module

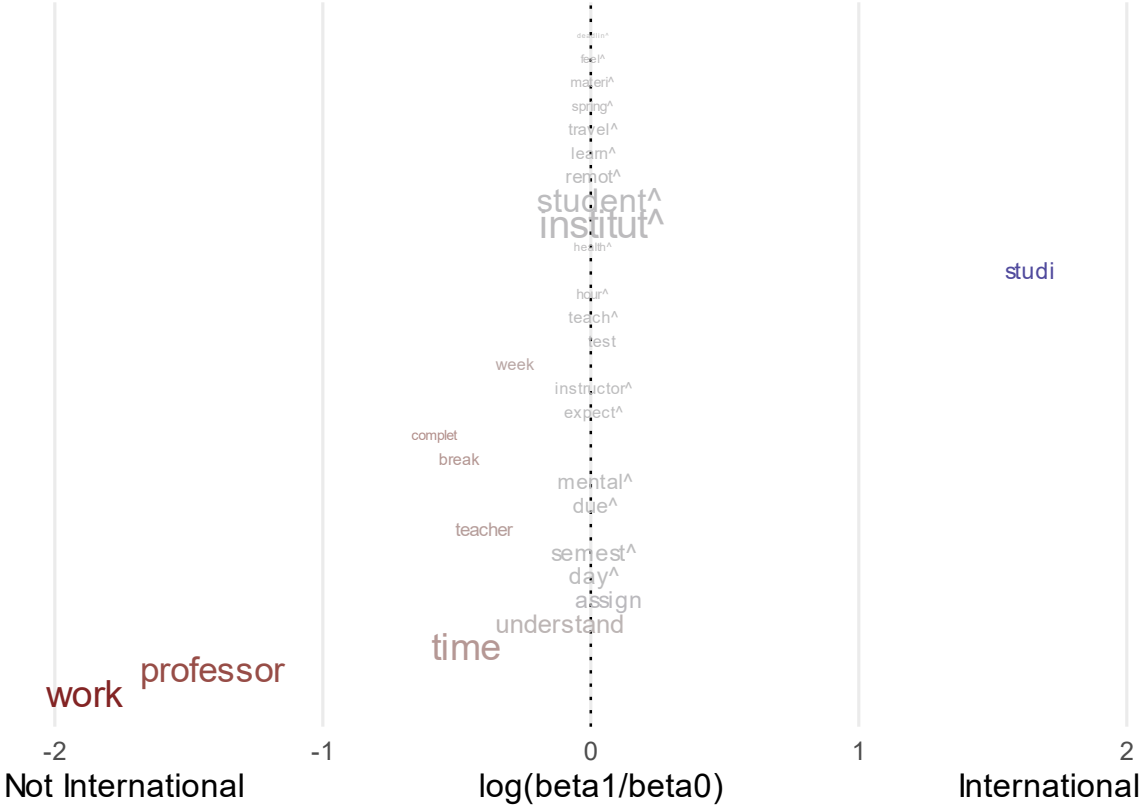
Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.



Note: vertical alignment is random. Terms unique to one group denoted w ith ^.

Top 20 terms in topic 1 in NSSE Coping with Covid Topical Module

Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.



Note: vertical alignment is random. Terms unique to one group denoted w ith ^.

Conclusions

Case 1: Senior Transition - Post-Graduation Plan

3 Key Topics from NLP Text Data analyses

- **Topic 3** (+): Match Curriculum with Industry Standards and Expectations**
  - Industry Alignment and Career-Specific Preparation and Opportunities,
  - Curricular Adaptation for Skills Development,
  - Enhancing Competitiveness and Marketability.
- **Topic 1: Inclusive and Modern Education for Diverse Student Populations**
  - Holistic Support for Non-Traditional Students (e.g. advising)
  - Possible disparities among Different Type of Students.
- **Topic 2*** (-): Real-World Readiness through Practical, Interactive Teaching and Learning**
  - Aligning course content with real-world demands: practical skill development in courses, industry-relevant coursework,
  - Expecting more interactive and dynamic learning environments.

Exemplar documents from topics in NSSE Senior Transitions Topical Module

Is there anything your institution could have done better to prepare you for your career or further education?

Topic 3 : job, major, career, student, graduat

(Doc ID 3427) more know ledge on careers in my field of supply chain. Also, access to more more support in gathering information on available internships, rotational programs, and other programs offered by companies. Handshake is great, but some form of tailored opportunities for specific majors w ould be great. My college, the College of Business, did a lo...

Topic 1 : class, time, professor, student, advisor

(Doc ID 1746) I honestly w ould of liked to see professors taking the time to brush up on their skill sets of w hatever it w as they w ere planning/having to teach. It w as very common for these professors to blow through an enormous amount of material but a lot of the information is outdated and not really used in the real w orld anymore. I understand if they ...

Topic 2 : student, onlin, faculti, person, campus

(Doc ID 795) Yes, [INSTITUTION] could have made greater steps for equality in the educational setting and learning environment. Rather than putting clear guidelines for how teachers are supposed to w ork w ith students w ho have disabilities during a pandemic and supporting said students in an online learning environment, [INSTITUTION] has come out w ith w ishy w ashy and mixed me...

Case 1: Senior Transition - Post-Graduation Plan

- **Major Terms for *WoC STEM* vs. *Other* Students**

	Topic 3*	Topic 1	Topic 2***
WoC STEM	*Appli, find, requir, student*	*Depart, financi, advisor, take, materi*	*Depart, park, support, struggl, chang*
Other Students	*Work, colleg, world, skill*	*Student, professor, class, teach, program,*	*Educ, instructor, staff, communic*
Both	*Feel, school, internship, provi,*	*Make, care, teacher, univer, school, work*	*Student, onlin, campus, cours, understan, feel*

Case 2: Coping with COVID – Institutional Response Feedback

3 significant topics identified from NLP Text Data analyses

- **Topic 2*** (+) : Social challenges imposed by policies**
  - Lack of understanding on the differences between interacting in public spaces vs. in private spaces
  - Frustration with lack of individualized social outlets
  - Strict policies even for family members
- **Topic 3*** (-) : Frustration with the modality of online classes**
  - Feeling unfair the education they paid for
  - Feeling unsatisfied with the quality of education received online
  - Pointing out disciplinary differences on the need for in-person classes with science majors
- **Topic 1** (-) : Learning fatique due to break polices**
  - One week-long break between semesters is not enough
  - Getting two "Wellness Days" with no assignments instead of a full spring break is not enough
  - Faculty intentions for course learning conflicting with institutional intentions on breaks

Exemplar documents from topics in NSSE Coping with Covid Topical Module

Please describe anything else you would like to share about your experiences as a student during the COVID-19 pandemic or about your institution's related actions and policies.

Topic 2 : covid, school, feel, year, campus

(Doc ID 2485) I fond the policies with dorm housing stupid. You can't visit with others unless in the lounge which is stupid if you and your friend wanna just chill or watch a movie or play video games or something. It's also stupid because I can't have my actual family members or friends even visit when they have the chance because of Covid policies and ...

Topic 3 : student, class, onlin, pandem, colleg

(Doc ID 2351) I still can not get over the fact I'm paying the full price tuition for IN PERSON learning when all of my classes have been fully online. This is especially frustrating because in my major you DO NOT get the same quality of learning online as you do in person. That might work for a liberal arts degree but not for a bachelors of science. I pa...

Topic 1 : work, professor, institut, time, student

(Doc ID 1144) With how this year's academic schedule was set up, students only had a week long break during the week of Thanksgiving, which was a week AFTER many courses had ended. Additionally, for this spring semester, there is no actual break. We get 2 "wellness days" on Feb 23 and March 24. While teachers could not assign homework for those 2 days, th...

Case 2: Coping with COVID – Institutional Response Feedback

- **Major Terms for *International* vs. *Domestic* Students**

	Topic 2***	Topic 3***	Topic 1**
International	*work, time, good, plan, person, countri, famili*	*colleg, petit, deni, credit, virtual*	*institut, student, studi, remot, learn, travel, spring, materi*
Domestic	*feel, campus, make, colleg, live, thing, friend, mask*	*class, learn, person, univers, institut, difficult, job, care, educ, good, health*	*work, professor, time, day, semest, due, mental, expect, instructor, teach, hour*
Both	*year, peopl, hard, made, social, life, pandem*	*student, online, pandem, time, cours, tuition, campus*	*assign, teacher, break, complet, week, test*

Overall Conclusions and Implications

1. The NLP technique is efficient for generating topics qualified by frequent and exclusive key terms to offer insights such as:
    - Students appreciate curricula matched with industry standards and expectations
    - Social challenges imposed by institutional policies were a major source of complaints
2. The topics and key terms frequently used by different student populations can be visualized and reveal, for example:
    - International students may be less interested in the modality of instruction while domestic students may be more concerned with mental health and expectations such as due dates
3. Common interests or polarity of topics among different student groups can also be spotted
4. Still requires your input!

Demonstration and Steps to *Create Your Own*

Corpus-Tokenize-Stopwords-Stem-DFM-STM

```
library(quanteda)
 (x <- "The parking at IU is terrible, but the professors are great! :)")
  [1] "The parking at IU is terrible, but the professors are great! :)"

(x <- tokens(x, what = "word", remove_punct = TRUE))
  Tokens consisting of 1 document.
  [1] "The"  "parking"  "at"  "IU"  "is"  "terrible"  "but"  "the"  "professors"  "are"  "great"

(x <- tokens_remove(x, c(stopwords("english", source = "smart"), "IU") ) )
  Tokens consisting of 1 document.
  [1] "parking"    "terrible"   "professors" "great"

(x <- tokens_wordstem(x))
  Tokens consisting of 1 document.
  [1] "park"      "terribl"   "professor" "great"

dfm(x)
  Document-feature matrix of: 1 document, 4 features (0.00% sparse) and 0 docvars.
         features
  docs    park terribl professor great
  text1    1       1         1     1
```

Corpus-Tokenize-Stopwords-Stem-DFM-STM

```
library(dplyr)
library(quanteda)
 library(stm)

# quanteda functions to create and clean document-feature matrix
 my_dfm <- read.csv("my_data.csv") |> # base pipe %>%
  select(id, metadata, text) |>
  mutate(text = tolower(text)) |>
  corpus(docid_field = "id",
         text_field = "text",
          meta = list(metadata))
```

Corpus-Tokenize-Stopwords-Stem-DFM-STM

```
my_dfm <- tokens(
  my_dfm,
  what = "word",
  remove_punct = TRUE,
  remove_numbers = TRUE,
  remove_separators = TRUE,
  split_hyphens = TRUE,
  include_docvars = TRUE) |>
  tokens_remove(
    c(stopwords(language = "english", source = "smart"),
     "IU", "IUB", "other", "custom", "stopwords")) |>
  tokens_wordstem() |>
  dfm()
```

Corpus-Tokenize-Stopwords-Stem-**DFM**-STM

```
library(tibble)
convert(my_dfm, "data.frame") |>
  as_tibble() |>
  head()
 # A tibble: 6 × 4,472
  doc_id   general professor understand  work student affect
   <chr>    <dbl>     <dbl>      <dbl> <dbl>   <dbl>  <dbl>
1 70002…       1         1          1     1       1      1
2 70003…       0         0          1     0       0      0
3 70004…       0         0          0     0       0      0
4 70004…       0         0          0     0       0      0
# ℹ 4,462 more variables: pandem <dbl>, applaud <dbl>, …
```

Tokenize-Corpus-Stopwords-Stem-DFM-STM

```
library(stm)
library(broom)
library(tidytext)
my_stm <- stm(my_dfm, K = 3,
              prevalence = ~metadata,
              content = ~metadata, # doing both can take time!
              init.type = "spectral")
my_stm
## A topic model with 3 topics, 3018 documents and a 4471 word dictionary

glance(my_stm)
       k  docs terms  iter alpha
   <int> <int> <int> <int> <dbl>
 1     3  3018  4471    85  16.7
```

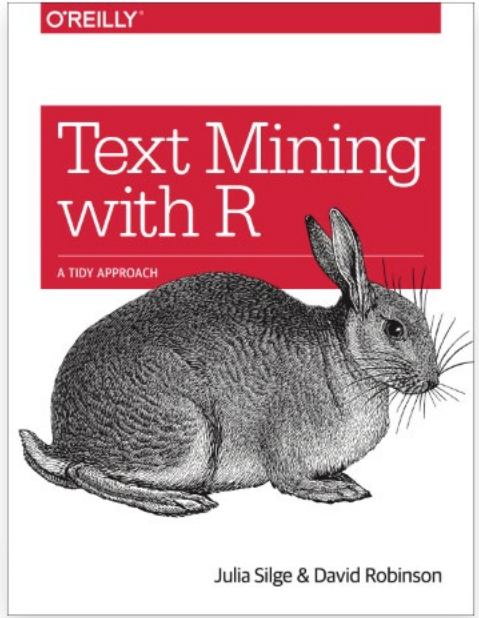
Tokenize-Corpus-Stopwords-Stem-DFM-**STM**

```
tidy(my_stm, matrix = "gamma") |>
  arrange(document) |>
  head()
# A tibble: 6 × 3
 document topic gamma
    <int> <int> <dbl>
1        1     1 0.392
 2       1     2 0.182
 3       1     3 0.426
 4       2     1 0.233
 5       2     2 0.538
 6       2     3 0.229
```

```
tidy(my_stm, matrix = "beta") |>
  arrange(desc(beta)) |>
  head()
 # A tibble: 6 × 4
   topic term        beta y.level
   <int> <chr>      <dbl> <chr>
1      3 student   0.0729 FALSE
2      3 student   0.0690 TRUE
3      1 work      0.0502 FALSE
4      3 class     0.0438 FALSE
5      1 professor 0.0404 FALSE
6      2 covid     0.0394 FALSE
```

Creating your own

1. Roberts et. al. 2019 “stm: R package for Structural Topic Models” (https://www.jstatsoft.org/article/view/v091i02)
2. Roberts et. al. 2014. "Structural Topic Models for Open-Ended Survey Responses" (https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12103)
3. Text Mining with R (https://www.tidytextmining.com/) (LDA)
4. R packages:
   - Cleaning and extracting model summaries: dplyr, broom, tidyr (tidyverse)
   - Tokenizing and text cleaning: quanteda, stringr
   - Modeling: stm, topicmodels (LDA), quanteda (QUANtitative analysis of TExtual DAta)
   - Plotting: ggplot2, stm
   - Other useful packages: tidytext, rmarkdown, tokenizers



Q & A