**COVID-19 Project - Text Analysis in R**

|  |  |  |
| --- | --- | --- |
| **Source and Topic** | **Study Overview** | **Notes on Study and Usefulness** |
| [Research Note: Scraping Financial Data from the Web Using the R Language](https://drive.google.com/file/d/1twHOAPXX2UbcfzWcG5FjZSsa5AmIwvHS/view?usp=sharing)  **Web Scraping - Example** | This study provides an example of how to scrape financial statements from different websites, scrape data from the financial statements, and do text analysis to measure sentiment. | I am not sophisticated enough in my understanding of R code yet to know the extent to which this source is helpful, but it does provide screenshots of R code that can be used to replicate the web scraping.  The study also overviews R packages and resources that can be used to figure out web scraping. |
| [Web Scraping Using R](https://drive.google.com/file/d/1KG4joI9cMdtxjgJmhshJ1PD8ujMnLO5T/view?usp=sharing)  **Web Scraping - General Resource** | This is designed to be a tutorial that walks through processes associated with web scraping. This tutorial starts with a one webpage example and moves to multiple webpage examples. | This source provides a link to a website that includes videos, practice examples, and R coding for web scraping. There are several different modules and examples you can work through on the website.  Even though this isn’t a study, this source looks to provide helpful resources for building web scraping code using “rvest” in R. |
| [Topic Analysis of Road Safety Inspections Using Latent Dirichlet Allocation](https://drive.google.com/file/d/1SJMxV9icy-4qXLsL-1OLdIk64nfTyKpV/view?usp=sharing)  **Text Analysis Methods - Example** | This study is less about web scraping and more about text analysis once the text is gathered. This study lays out the methodology underlying Latent Dirichlet Allocation (LDA), which looks for word associations by topics within a text. | While web scraping is less emphasized in this study because the text is already gathered, this study focuses on the LDA method for analyzing text. This study walks through all of the assumptions and challenges of LDA, as well as the R packages and code needed to do it. |
| [A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)](https://drive.google.com/file/d/1ivgVlgpweC1crKrnIp-Mss_7F4_X5w03/view?usp=sharing)  **Text Analysis Methods - General Resource** | This article breaks down the different types of computer-aided text analysis that currently exists, as well as a suggested research methodology for a topic analysis paper. | I included this article because I found the typology of the different types of text analysis helpful. The article also outlines all of the steps that need to be undertaken in a computer-aided web analysis study, from web scraping to finished product.  The four types of text analysis presented in the article are: (1) Thematic Analysis, (2) Content Analysis/Dictionary, (3) Bag-of-Words (Count-Based), and (4) Natural Language Processing.  This article does not use R for the web scraping, but does provide tips for cleaning scraped text. |
| [Visualizing Student Opinion Through Text Analysis](https://drive.google.com/file/d/1ognJiFwnuDGNMCrtRWMxSPhQAqlGdyKP/view?usp=sharing)  **Text Analysis Methods - Example** | This study coded open-ended student survey responses from college course evaluations and visualized this data. | This study used Python’s Natural Language Toolkit to process the data.  This study used Latent Dirichlet Allocation (LDA) to split text documents (individual survey responses) into topics (e.g. lecture, tutorials) and then group words by topic. LDA also helped determine whether the topic was mentioned in a positive or negative light. |
| [Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web](https://drive.google.com/file/d/1HXHT8_wrMh6kwIU8yvGnusUwgPYYNQsz/view?usp=sharing)  **Text Analysis Methods - Example** | This study coded messages by investors on stock message boards as having a positive, negative, or neutral sentiment. | After building algorithms around stock lexicon, the model estimated sentiment in five different ways, with the final coding of a message reflecting what the majority of the five models coded in terms of sentiment.  This article does not use R for the web scraping. |

|  |  |  |
| --- | --- | --- |
| **Source and Topic** | **Study Overview** | **Notes on Study and Usefulness** |
| [New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings](https://drive.google.com/file/d/1PbT_S_NZQtpPfCkvWJywY91yI98yw62g/view?usp=sharing)  **Web Scraping - Example** | This study web scraped over 11 million ads from Craigslist and then did spatial and temporal visualizations to show rent prices in different markets relative to the fair market value. | This study walks through the process of using web scraping to conduct text analysis and eventually produce visualizations.  The study used Python for the web scraping and provides some depth in the explanation of the web scraping process. The study is less descriptive about how the text was cleaned and analyzed. |
| [What topics do presidential candidates discuss on Facebook?](https://www.brookings.edu/blog/techtank/2016/05/19/which-topics-do-presidential-candidates-discuss-on-facebook/)  **Web Scraping - Example**  **Text Analysis - Example** | This study is from Brookings.  It took Facebook posts for 2016 Presidential candidates and ran Latent Dirichlet Allocation (LDA) topic modeling on the text of these posts to determine which topics 2016 presidential candidates were posting about most. | While this source does not get into the details of web scraping and LDA, it does provide a helpful overview of LDA and the visualizations are intuitive.  It is seeming more and more like LDA is the preferred method of people doing topic analysis based on text.  These studies were done by professors at Arizona State University and Kevin Desouza, whose name I have seen on other students. These researchers may have more context or information on how to do web scraping and topic analysis. |
| [RCrawler: An R package for parallel web crawling and scraping](https://drive.google.com/file/d/1OOUkjUZGEF0Yupa6TxOMjy00iSP8iW9L/view?usp=sharing)  **Web Scraping - General Resources** | This article discusses and compares different R packages for web scraping. It is not a study - It is a software package overview article. | This article compares the ability of different R packages to web crawl, retrieve text, and parse text. The R packages covered in the article include scrapeR, Rvest, and Rcrawler.  scrapeR and Rvest are unable to web crawl, while RCrawler can web crawl. This might be helpful if there is not already a list of URLs.  This article is technical, but user-friendly. |
| [Automate Text Analysis for Consumer Research](https://drive.google.com/file/d/1OifI6fDv0Sqa2iDH8sU6JUfleIiOYfct/view?usp=sharing)  **Web Scraping and Text Analysis - General Resources** | This article walks through a step-by-step process for setting up a web scraping/text analysis research study. | This article is a little long, but it felt like a solid glossary and overview of the different stages in a web scraping and text analysis research study. |
| [Scraping Craigslist Posts](https://sia.planning.unc.edu/post/2020-02-04-scraping-craigslist-posts/)  **Web Scraping - Example** | This is a quick website tutorial, with screenshots and code, that show how to web scrape Cragslist posts. | This web scraping in this overview involved scraping Craigslist posts (like the example above), but provide a more user-friendly overview. This is an example of web scraping when the information is standardized in the same place in the html tree for each post. |
| [Tuning in to Psychological Change: Linguistic Markers of Psychological Traits](https://drive.google.com/file/d/1RqUaVAjV0Ll2eQqdB_rfW3v5-0KJmzvk/view?usp=sharing)  **Text Analysis - Example** | This study uses Linguistic Inquiry and Word Count (LIWC) to analyze how the sentiment, tone words, and word types of popular song lyrics have changed over the past few decades. | This study does not talk about the web scraping part because I think they already had the song lyrics pulled, but it does discuss LIWC, which seems to be another software/common method for analyzing text for common words, themes, and tone. |
| [The Use of Topic Modeling with Latent Dirichlet Analysis with Open-Ended Survey Items](https://drive.google.com/file/d/1twHOAPXX2UbcfzWcG5FjZSsa5AmIwvHS/view?usp=sharing)  **Text Analysis - Example** | This study applies LDA to open-ended survey responses. | This study does not do web scraping, but provides some helpful resources to conceptualize topic modeling. |
| **Source and Topic** | **Study Overview** | **Notes on Study and Usefulness** |
| [What topic modeling could reveal about the evolution of economics](https://drive.google.com/file/d/1xYtVXXiHd34w6XsfF2WnwkeR1nTW66xY/view?usp=sharing)  **Text Analysis - Example** | This study applies LDA topic modeling to Economics studies in JSTOR to track the evolution of topics discussed in Economics over time. | This study presents LDA data in a use-friendly manner. |

[**Center on Reinventing Public Education Research Breakdown**](https://www.crpe.org/current-research/about-work)

**Overview of CRPE Research:**

* CRPE’s current analysis is focused on summer learning and planning for the 2020-21 school year.
* CRPE is continuously tracking the following variables with respect to school district response to COVID-19:
  + Technology access
  + Remote instruction
  + Student support
  + School operations
* The main goal of this research is to help schools, districts, and states develop best practices in terms of K-12 education’s response to the pandemic. This research is national in terms of its unit of analysis.
* There is also information in [this database](https://docs.google.com/spreadsheets/d/1zDTvGQ9x5Hmja87wO7LG5zmhX7-KIAnp6Z6j1imdTYk/edit#gid=0) on state responses to COVID-19 school closures. The state policy database tracks references to “special education”, “students with disabilities”, or “special needs” in state policies.

**Notes on Methodology and Process:**

* The Summer/Fall District Response [database](https://docs.google.com/spreadsheets/d/1WLwHgQvoWT1op6RALANmwOE7akFfDEJGbKLeXEQnEes/edit#gid=0) developed by CRPE includes **100 districts** serving nearly **10 million students.**  The Spring District Response [database](https://docs.google.com/spreadsheets/d/1n_5rHMCf2GAiOcgIa4pCZD8l9zh5-1NQRKmYEk3qPLg/edit#gid=778813724) includes **82 districts** and **18 charter school management organizations** serving nearly **9 million students.**
  + Information for the Spring database stopped being tracked on May 15, 2020, so information below relates to the currently active Summer/Fall database.
* The following criteria was used to select districts:
  + Variation in size and geography
  + Includes the 30 largest districts in the country
  + Includes members of the [Council of the Great City Schools](https://www.cgcs.org/domain/57) - This Council includes 76 of the nation’s largest urban public school systems
  + Includes smaller districts that participate in CRPE’s rural studies
  + Includes at least one district from otherwise unrepresented states
  + **The sample is not a representative sample**
* CRPE gathers information on school district response from district websites and media reports - The database is updated as new information on the districts is released
* Texas districts in the Summer/Fall database include the following:
  + Arlington Independent School District
  + Austin Independent School District
  + Cypress-Fairbanks Independent School District
  + Dallas Independent School District
  + El Paso independent School District
  + Fort Worth Independent School District
  + Houston Independent School District
  + Northside Independent School District
  + San Antonio Independent School District
  + Great Hearts Academies (Charter)
  + IDEA Public Schools (Charter)
  + KIPP Texas (Charter)
  + YES Prep Public Schools (Charter)
  + Uplift Charter Schools (Charter)
* Key Variables Tracked Include:

|  |  |  |  |
| --- | --- | --- | --- |
| **Elementary School** | **Middle School** | **High School** | **All Grades** |
| Summer Program - Yes or No  Learning Format - Virtual, Physical, Blended, Varies by Program/Grade, Independent Study  Review of Previous Year Content Planned?  Enrichment Opportunities? | Summer Program - Yes or No  Learning Format - Virtual, Physical, Blended, Varies by Program/Grade, Independent Study  Review of Previous Year Content Planned?  Enrichment Opportunities? | Summer Program - Yes or No  Learning Format - Virtual, Physical, Blended, Varies by Program/Grade, Independent Study  Review of Previous Year Content Planned?  Enrichment Opportunities?  Credit Recovery?  Credit Acceleration? | Summer Programming for Special Populations (e.g. SWD, ELL, Both, None)  Summer Program Tuition and Fees?  Teacher Training for Summer Programming?  Partner Programs for Summer Programming?  Summer Devices?  Meal Plan for Summer? |

**Note:** At this point, many of the variables being tracked relate to summer programming. It does not appear all of the variables that will be tracked during the Fall are in the database yet.

* Categories tracked in terms during remote learning curriculum in the Spring included (Tableau Map):
  + Formal curriculum, instruction, progress monitoring
  + Formal curriculum and progress monitoring, no instruction
  + Formal curriculum and instruction, no progress monitoring
  + Formal curriculum
  + Access to general education resources only

**Notes on Takeaways:**

* [The Wait for Fall Reopening Plans, and the First Details to Emerge](https://www.crpe.org/thelens/wait-fall-reopening-plans-and-first-details-emerge) - June 29, 2020
  + This article is based on the sample of 100 districts and 18 CMOs in the published databases
  + Only 3 districts in the sampe have released either a partial plan or full plan for the Fall
  + 61 districts have a plan underway for the Fall
  + 36 districts have not provided any information on a planning process for the Fall
* [Too many schools leave learning to chance during the pandemic](https://www.crpe.org/sites/default/files/final_national_sample_brief_2020.pdf) - June 10, 2020
  + This study includes analysis of a new nationally representative sample of 477 school systems, including the 81 in the Spring database.
    - Collected and coded publicly available information
    - Project is in collaboration with RAND Corporation and stems from the ongoing American School District Panel Project designed to build a nationally representative panel of American school districts
    - Appendix B includes coding definitions
    - Pulled information from public available sources, including school district website and social media and publicly available teachers’ online learning platform pages
  + Only 1 in 3 districts expect teachers to provide instruction, track student engagement, or monitor academic progress for all students.
  + Rural districts are less likely to provide instruction than urban and suburban counterparts.
  + Affluent school districts are more than twice as likely to provide real-time lessons than economically disadvantaged districts.
  + 85% of districts made sure students received some form of a grade and subject-specific curriculum, either in packets, on Google Classroom, or on some other tech platform.
  + 27% of school districts required schools to track their students’ attendance.
* [Federal Special Education Guidance is Clear; Now States Must Step Up](https://www.crpe.org/thelens/federal-special-education-guidance-clear-now-states-must-step) - May 5, 2020
  + There is wide variation in the SPED policy being set by states.
  + Special education guidance was gathered from all 50 states and Washington, D.C.
  + There are three levers states can pull to influence SPED - (1) Compliance, (2) Resources, (3) Learning Plan. Only 6 states require all three.
  + Providing resources is a strength of Texas.
  + Only 27 states require districts to submit remote learning plans, and only 18 of these states require districts to include special education in those plans.
* There are [databases](https://www.crpe.org/current-research/impact) for Massachusetts, North Carolina, Delaware, and Connecticut that house school district distance learning plans.
* CRPE encourages policy researchers on K-12 COVID-19 policy to reach out for [collaboration](https://www.evidence-project.org/).

**COVID-19 Study by REACH at the National Center for Research on Education Access and Choice at Tulane**

**Sources Describing the Study:**

* [The 74 Million - Ambitious Research Project - To Review How Every School in America Responded to COVID-19 - Aims to Deliver Its First Findings in Early July](https://www.the74million.org/article/ambitious-research-project-to-review-how-every-school-in-america-responded-to-covid-19-aims-to-deliver-its-first-findings-in-early-july/)
* [REACH - REACH to Study How Coronavirus Has Impacted Schools Nationwide](https://www.reachcentered.org/news/reach-to-study-how-coronavirus-has-impacted-schools-nationwide)

**Updates:**

* The grant and study was announced in April
* Initial findings are set to be published as soon as early July to help facilitate discussions around fall planning

**Methodology and Scope of the Study:**

* The study aims to track how every K-12 public, private, and charter school in the U.S. is responding to the coronavirus pandemic using a $100,000 grant from the U.S. Department of Education. The part of the study will use human analysis and AI/machine learning, but the exact methodology has not yet been released.
* The study will start with a random sample of **3,500** **schools.** This data will be manually collected and coded by Tulane students.
* The team of researchers is working with the Tulane Computer Science Department to then extend the study to the approximately **150,000 schools** nationwide.

**Variables and Research Questions Underlying the Study:**

|  |  |
| --- | --- |
| **Research Questions** | **Variables** |
| * To what extent do COVID-19 responses differ across school types (e.g. district, charter, private)? * To what extent is there a correlation between student demographics and how schools react to the crisis? * How are students learning when school buildings are closed? * How do COVID-19 responses in schools differ based on spending levels? | * How quickly schools and districts respond to school closures * Personalization and engagement in instruction * Non-academic engagement (e.g. morning meetings and office hours) * Academic expectations (e.g. grading policies and time spent on subjects) * Continuation of services (e.g. free lunch and counseling) * Equity of access for English leaders, students with disabilities, and students who need internet connectivity or devices * Do schools offer remote learning online? * Are schools offering laptops and WiFi hotspots to families who lack these resources? |

**Critiques of the Study to Consider:**

Concerns articulated by Morgan Polikoff from USC in *The 74 Million* article:

* There are concerns about the quality of information gleaned from school websites because school websites are often hard to navigate and lacking in information.
* There are concerns about how messaging on school websites will match the reality of what actually happens in schools, so survey results might be a consideration to check implementation.
* It might be difficult to take all of this information and convert it into a user-friendly format.

[**American Enterprise Institute COVID-19 Education Response Longitudinal Survey (C-ERLS)**](https://www.aei.org/covid-19-education-response-longitudinal-survey-c-erls/)

**Methodology and Scope of the Study:**

* School district websites are used to gather the data under the assumption that they are the central communication hub for transitions to remote learning.
* Collected district-level data and not school-level data.
* This is a nationally representative sample of **250 “regular school districts”** across 50 states and D.C. using the 2017-2018 Common Core of Data (CCD) collected by the National Center for Education Statistics (NCES) as the data source that provided the universe of districts.
* Sampling probability was proportional to size, meaning large districts were more likely to be selected than if a simple random sample were done.
* This study is **longitudinal** and data is collected every 7-10 days to track performance over time. Currently, the **6th** wave of the survey is reported. The first wave was done March 26-27. The most recent wave was May 27-29.
* The most recent reported margin of error is +/- 6.1%.

**Variables and Research Questions Underlying the Study:**

|  |  |
| --- | --- |
| **Research Questions** | **Variables** |
| * Do school districts’ efforts differ across school levels (e.g. elementary, middle, high)? * Do districts’ responses vary by district size? * What online platforms are districts using for asynchronous and synchronous instruction? * What technology and internet accommodations are being made? * How are accommodations for special education and ELL students being handled? | * Closures - Are districts closed and when do they plan to reopen? * Food Services - Do districts have a plan for offering meals on their website? Are meal pickups daily? Can meals be delivered? Are multi day meal pickups allowed? * Technology Assistance - How many districts mention technology support on their website (e.g. device support and internet support)? * Education Programs - Are districts offering the following programs:   + Virtual supplemental content   + Packets   + Asynchronous Instruction   + Synchronous Instruction   + Hybrid instructions, mostly online, or mostly packets/in-person * Expectations - Are districts implementing the following expectations:   + Expected participation   + Taking attendance remotely   + Grading student work for performance or completion |

**Key Findings:**

* In the Spring, 1 in 5 school districts offered rigorous remote instruction most closely resembling in-classroom learning, while two in five offered perfunctory instructional programs.
* In terms of graduation ceremonies - 43% were in-person and 28% were virtual.
* By the end of May, 97% of districts had some sort of education program or offering available.
* By the end of May:
  + 20% of districts offered rigorous instruction with engagement expectations and attendance
  + 40% of districts offered moderate instruction
  + 40% of districts offered perfunctory instruction relying on packets, lack of participation requirements, no attendance, or no grades
* 95% of districts offered some meal service by the end of May.