James Hizon

FRS Data Science

Individual Report

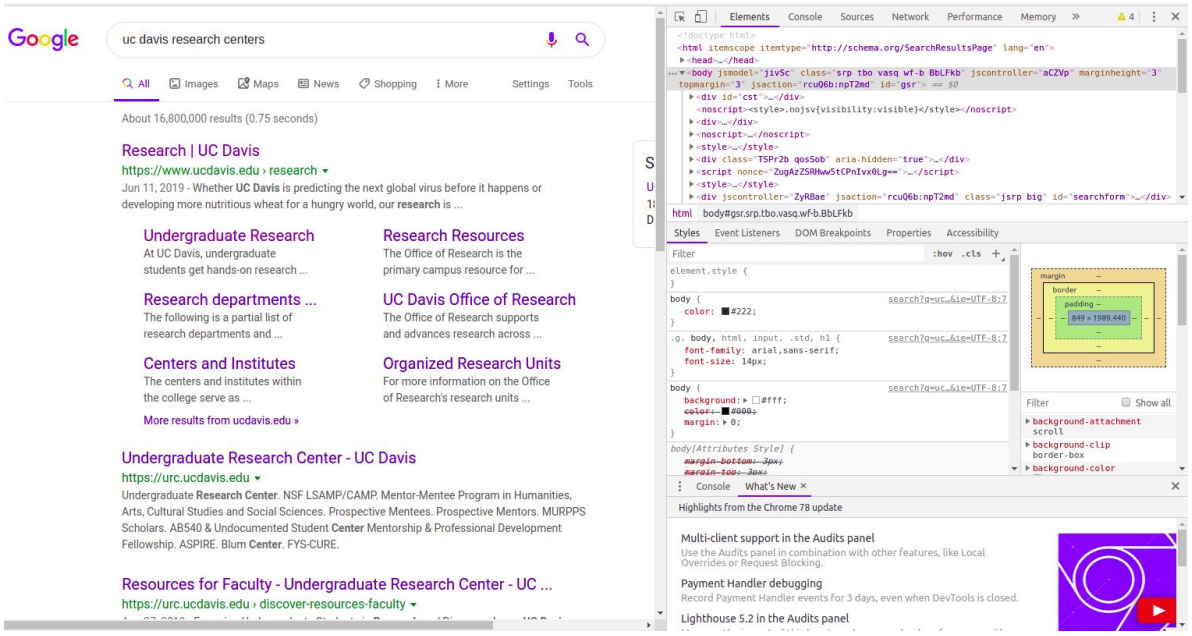
I. Introduction

In our research project, we are driven to improve the discoverability and accessibility of UC Davis's research center resources, and identify new opportunities for multidisciplinary research and engagement with STEM. As our main goal, we want to construct code inside RStudio to help in analyzing how University of California, Davis is represented online. Our desire is to give recommendations to Provost on what necessary changes need to take place within STEM Portal.

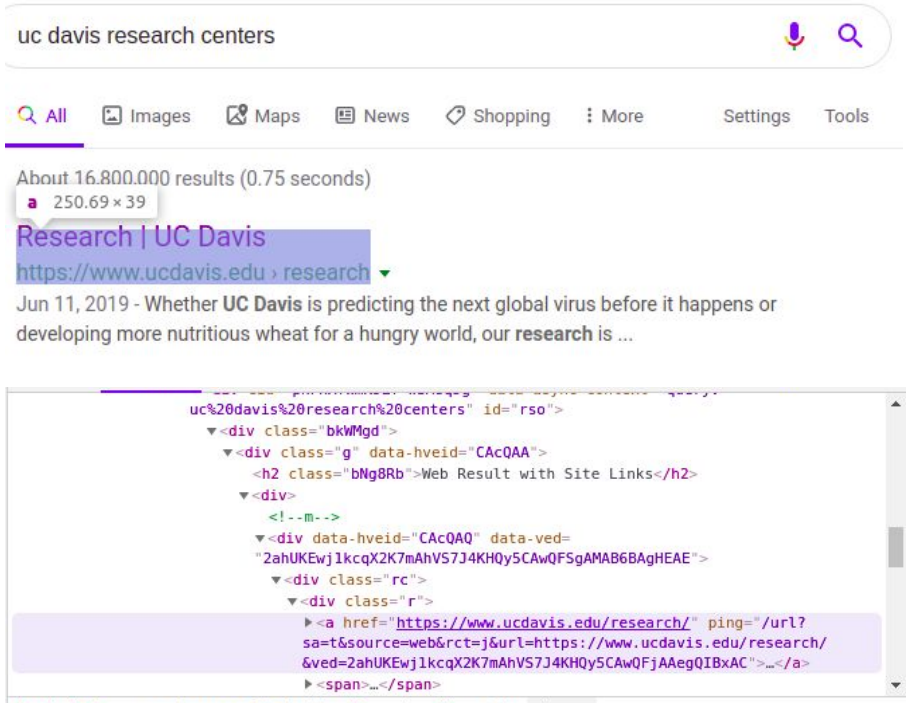
My specific research question involves understanding which topics should we expect most of our research to be related to. In other words, what are some of the most important research topics to take into consideration?

II. Methods

In the beginning was web scraping, and web scraping was with the process, and the web scraping was the process by which we obtained our data. In order to drive out our discoveries, we needed to first observe patterns within the Google web search engine for UC Davis research. We need to click on the top-right hand corner. After clicking "More tools", we want to click on "Developer tools," to obtain the following:



Then, we must manually find our patterns. After a series of clicking through CSS selectors, we find the code that gives us our primary links as follows:



The above HTML/CSS Selector helps use know what code to look for in our web scraping. The code controls our primary link.

We wrote out code inside RStudio, which enabled us to take urls with the ".edu" ending, collect our data after we added our nested for-loop with a "tryCatch" inside to catch our error, and we obtain text info.

```
ucdresearch <- read_html("https://www.google.com/search?q=uc+davis+research+centers&sxsrf=ACYBGNSNM
ucdresearch

# extract the nodes you need with css selectors

result <- html_nodes(ucdresearch, "a")
print(result)
length(result)

links <- grep("ucdavis.edu", result, value = TRUE)
```

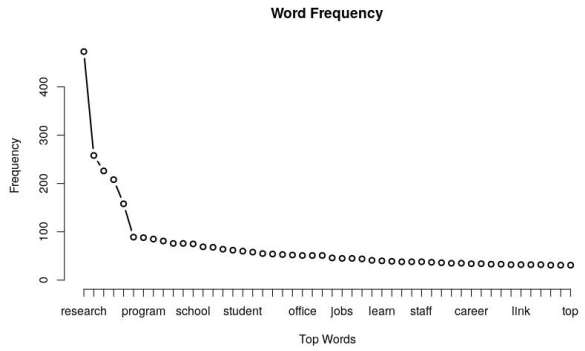
We used the function read_html() alongside html_nodes() to extract the nodes we need with our css selector. The grep() function was useful to grab all of our results that have "ucdavis.edu" in the url. Then, we created a nested for loop that ran our code and caught our errors. The following is the main code that we used to create a list of text files and then create a blob of text:

```
file_path <- "/home/james/FRS_Cure/Rstudio_Research/"
# Add txt_files_list to for loop.

content = ""
for (file in txt_files_list){
  full_filepath <- paste(file_path, file, collapse = "", sep ="")
  lines = readLines(full_filepath)
  #print(lines)
  combined = paste(lines, sep = "", collapse = " ")
  content[file] = combined
}
print(content)
```

After we create our blob of text, we were able to perform word frequency analysis with our own blob. In our natural language processing phase, we can change characters to lowercase, remove stopwords, remove other tokens on blacklist, remove punctuation, remove numeric elements, collapse multiple spaces, trim leading and trailing spaces, convert blob to a list of words, calculate total word count, set up a frequency table, sort the frequency table, convert to relative frequencies, find relative frequency of a given word, and plot the actual word frequencies.

Given our blob of text from primary link web scraping, we can obtain the following plot of a word frequency table:

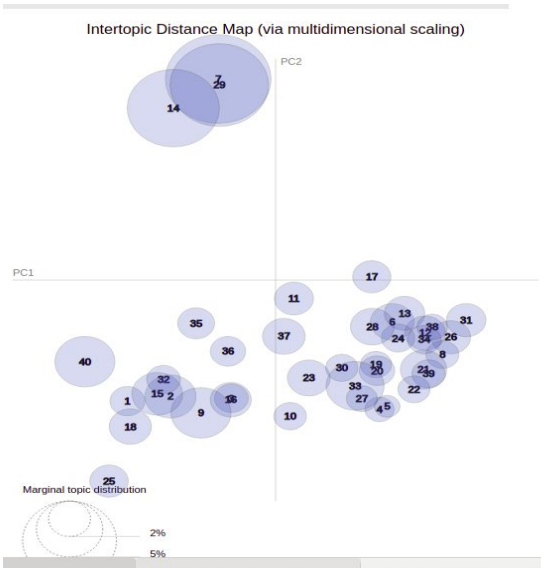


In our project, we ran into a lot of limitations, such as time constraints, understanding of programming and web scraping, and learning how to troubleshoot errors.

III. Results and Interpretations

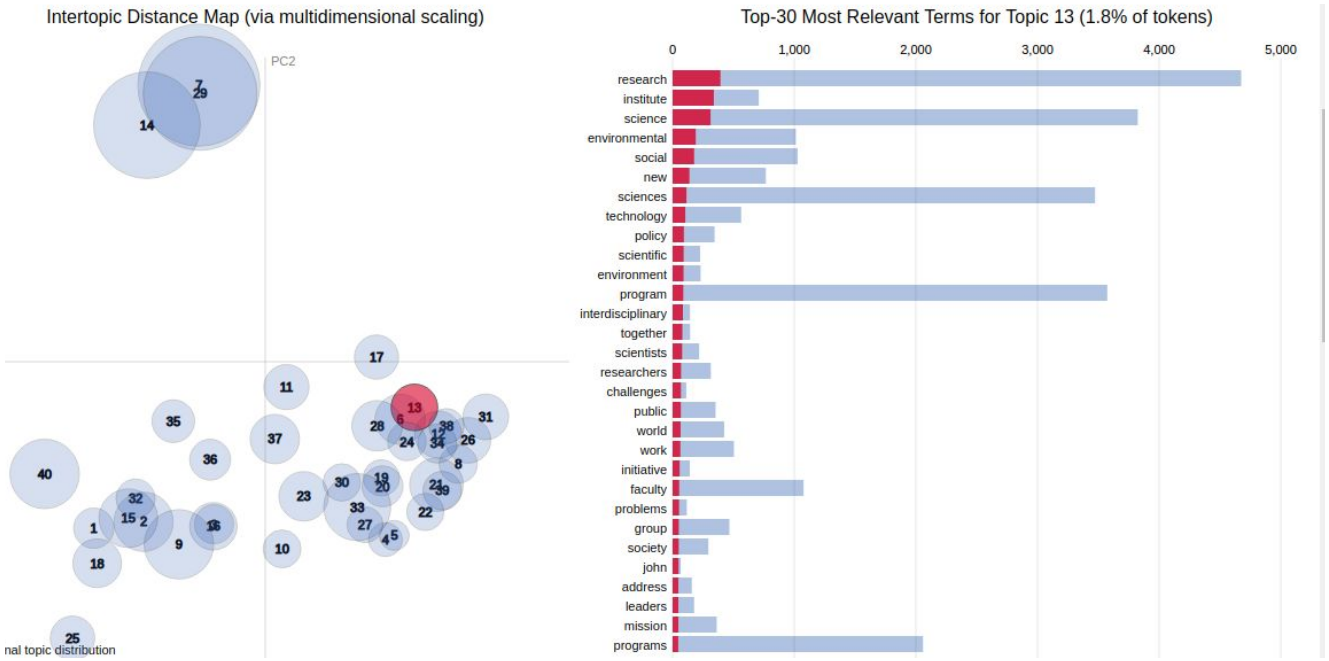
My specific research question involves understanding which topics should we expect most of our research to be related to. In other words, what are some of the most important research topics to take into consideration?

When answering our research question, we use an “Intertopic Distance Map”, in combination with the most salient and relevant terms. The following is a representation of our topics, where each circle is a representation of a given topic.

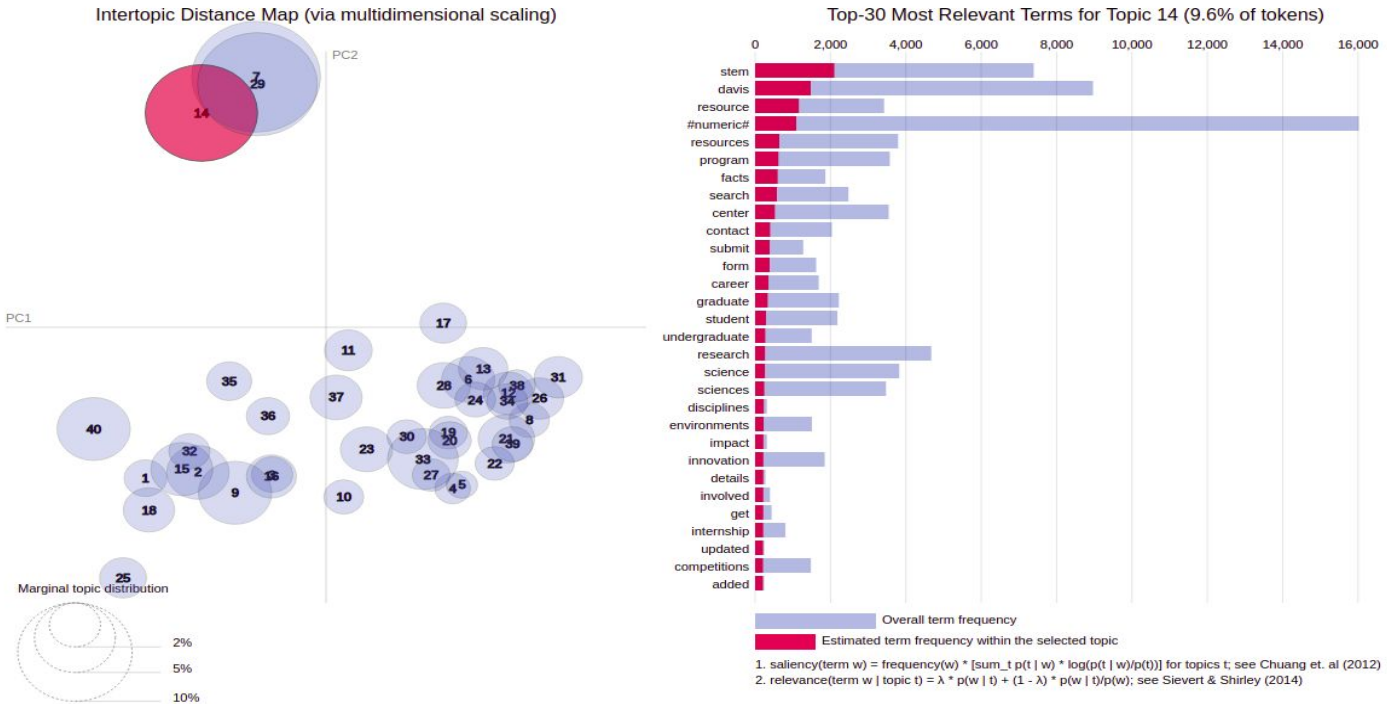


As we click on the word “research”, we are reduced to fewer terms. Note that the word “research” appears in the top-3 words most frequent words, which communicates well that our terms are indeed related to research. We clicked on two topics which had a small and a large circle. Topic 14 has a large circle to communicate how well-represented STEM Portal is to share how people can get involved in research. Topic 13, on the other hand, has a small cluster, to communicate how it is shortly represented. Given this information, we can communicate how a small portion of our data is related to environmental science as follows.

Topic 13:



Topic 14:



IV. Recommendations

Given the data we webscraped, there are several recommendations I can offer for the STEM Portal. One recommendation is to continue to communicate more opportunities in regards to research on webpages. Yet, this can also be based on how many opportunities are available (subjective). One strength can be found in how we are able to investigate the content of our research data by simply investigating the appearance of words. The limitation we would be found, however, in how these words are being put together. We have to do our best to figure out what our topic is, given list of words on the top-30 most relevant words.

The next recommendation, given the future outlook of events, is to further expand our research in the area of the environmental sciences alongside other disciplines. STEM Portal should continue to further research that aims to work together to create policies that care about environmental and social concerns, in addition to caring for a generation that will increase advancement in technology.

What I learned most about this research experience, is not only the essential skills of working in RStudio and increase my confidence from my own programming skills, but the preparation of working in the real-world with big companies. This undergraduate research experience was a great way to prepare me for other projects related to software development and data science in my career future.