Web Scraping for Agricultural Related Jobs on Indeed.com

An Analysis of the Agricultural Industry's Job Market

Douglas Abney

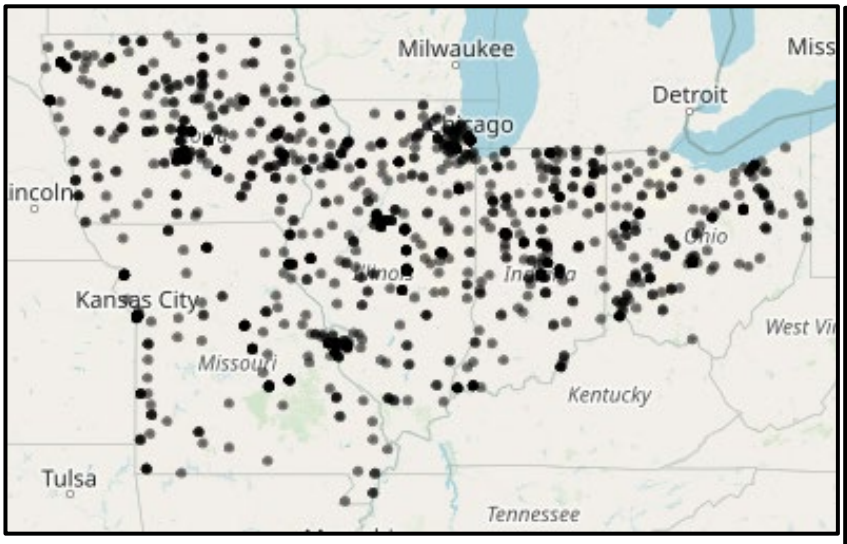

Figure 1: Map of all agricultural related jobs within the Corn Belt

ABSTRACT

Searching for jobs online can be challenging, especially when seeking opportunities that require unique skills. Are there other ways to search for jobs within agriculture without having to scroll through job listing sites endlessly? There is a wide belief that "big data" is the future of agriculture. This research project used data analysis to find data analysis jobs in agriculture. This paper outlines how web scraping and data analysis can be utilized to find specific agricultural related jobs. It also examines the potential for webs scraping in an agricultural sector job search and how data analysis can help identify new and helpful information.

Introduction

How long would it take someone to read through 1,622 job listings to find a desirable position? Job titles and descriptions have a tendency to be vague, and job seekers can become overwhelmed. They may miss their dream job after scrolling through hundreds of applications. These characteristics of the modern job search have become discouraging to many. I know how it feels to scroll through Indeed and LinkedIn endlessly and find nothing of interest. I have had trouble finding a job related to my passions.

When I dream of my perfect job, it involves the application of data science in the agricultural sector. If I search for "Data Science in Agriculture" in Indeed.com, however, I will find zero matches in Indiana, my home state, and less than fifty matches in in the entire United States. This begs the question: do these jobs exist? Is "big data" really the future of agriculture? Surely, these jobs have to be out there. It is possible that the algorithms used to filter postings on job listing sites are not perfect. After experiencing these struggles, I decided to look into why this is happening. Some questions came to mind. Why is job searching so hard? How can I find jobs I want? More importantly: Are there other ways to search for jobs with specific characteristic within the agricultural sector? This question challenged me to explore how data science could be used to find data science jobs in the agricultural sector. I decided to collect data from Indeed.com to better

understand the agricultural industry's job market and identify jobs that match my interests more closely.

There is a common saying that "modern agriculture requires modern solutions." It is widely believed that "big data" is the future of agricultural production. Within the agricultural sector, data can be collected and implemented in a variety of ways. Through data analysis, anecdotal information can be translated into actionable strategies from empirical patterns. Data can be used to provide information, to offer solutions, and to predict the future. Within data science, there is a practice called "web scraping." Web scraping allows data scientists to extract unstructured data that is already available on the internet and sort it into structured datasets that can be analyzed to create a better understanding of the world around us. For this project, I used web scraping to collect agricultural sector job listings from Indeed.com.

There are several factors that create difficulties for collecting job listings data at a national scale. The United States is diverse in both climate and geography. Differences in geography determine agricultural production within each state. Therefore, agriculture varies across states. The heterogeneity of job categories within the entire U.S. is vast. This factor, coupled with the size of data collected from each state, imposes difficulties for data management and analysis. Each state has hundreds to thousands of job postings within the agricultural sector. Collecting the descriptions of every available job in the agricultural sector could amount to immensely large "big data." For these reasons, I limit my analysis to job listings in Corn Belt states. The Corn Belt includes Indiana, Ohio, Illinois, Missouri, and Iowa. The Corn Belt serves as a preferred geographical location to explore because its agricultural sector closely matches the topics, I have studied at Purdue University. Analyzing the Corn Belt serves as an adequate first step for exploring how web scraping can be applied to the agricultural sector's job market.

The data used in this research paper represents agricultural job listing posted on Indeed.com in late January of 2021. According to Indeed.com, the site is the number one visited job searching website in the world (Indeed.com, 2021). As a result, there are a number of resources for web scraping Indeed.com. All of the data used in this paper were collected from Indeed.com on January 25th, 2021.Therefore, the data provide a snapshot of the agricultural sector's job market on that specific date. This may impose some biases. Some potential biases include seasonal job market fluctuations, the impact of COVID-19 on unemployment, and the variability of job openings within companies. Due to the effects of COVID-19, the unemployment rate in January of 2021 was 6.3% according to the Bureau of Labor Statistics (bls.gov, 2021). Our findings could vary significantly between the time of year that data are collected.

Throughout this paper, I will explain the many steps within this research project. The methodology section outlines the use of web scraping including a discussion of how to determine which webpages are suitable for web scraping. The findings section discusses characteristics of the data collected. The conclusion and future research outlines and explain steps for continued research.

Methodology

Data Sourcing

Data for this research was sourced by web scraping Indeed.com. The web scraping scripts were programmed in R using a package called "rvest." The rvest package is designed for harvesting data from webpages (rvest.tidyverse.org, 2021). Web scraping is the process of collecting and extracting data from web pages by creating a script that manipulates a browser to search for information via the internet. Web scraping scripts gather data by interpreting the underlying HTML code that constructs webpages. HTML is the standard markup language for creating webpages (w3schools.com, 2021). In order to extract data from webpages, data scientists require a small amount of HTML knowledge. These elements include the tags in HTML that define the location of content, such as text and hyperlinks. These elements of interest are the data points within the web page that data scientists anticipate harvesting. The HTML of a webpage is identified by right clicking on the specific elements of interest. A popup will display on the user's screen (see Figure 2). Selecting the "Inspect Element" option allows the user to explore the underlying HTML of the web page. This will show all of the code used to construct a webpage. From here, data scientists can find all of the needed tags for extracting data.

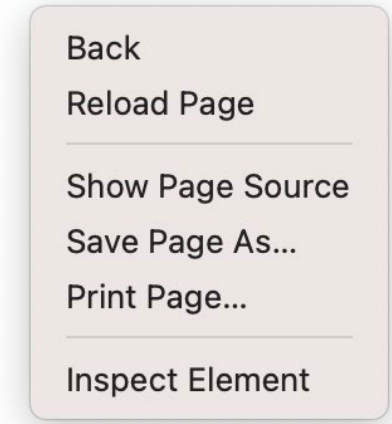

Figure 2: Right click to inspect element

Indeed.com is the most visited job searching website in the world (Indeed.com, 2021). Indeed's webpage layout allows for consistent web scraping. In fact, Indeed.com is a popular site for web scraping and there are several resources on how to extract data from Indeed.com. "A Detailed Guide to Web Scraping Indeed Jobs with R and rvest" by Pascal Schmidt, breaks down every step for extracting data from Indeed.com using R (thatdatatho.com, 2021).

Indeed.com is an ideal website for harvesting data because of its simple webpage structure. At the top of Indeed's page, there is a "What" input text box where one can type in the job category of interest. The web scraping script used for this paper, manually fills this input box with the term "agriculture." The script also fills the "Where" input text box with the states of interest. The code

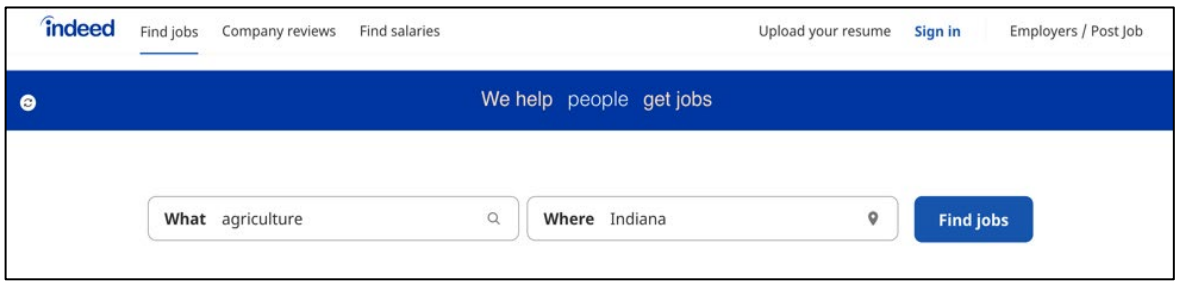

Figure 3: The top section of the main webpage for Indeed.com.

loops through each state of the Corn Belt. This process of iteration allows for collection of data in multiple geographical locations.

Data Collection

Each job listing page hosts 15 unique job postings. Job postings details are positioned within a box, as shown in Figure 4. These boxes contain the job title, company name, location, a brief description, and, occasionally, salary. These boxes are imbedded with a URL. The first web scraping script extracts the job title, company name, location, brief description, and the URL imbedded within each box. This extracted data is saved within a data frame in RStudio, as shown in Figure 5. The data in RStudio is organized and can be analyzed. The data is now ready for the second step of web scraping.

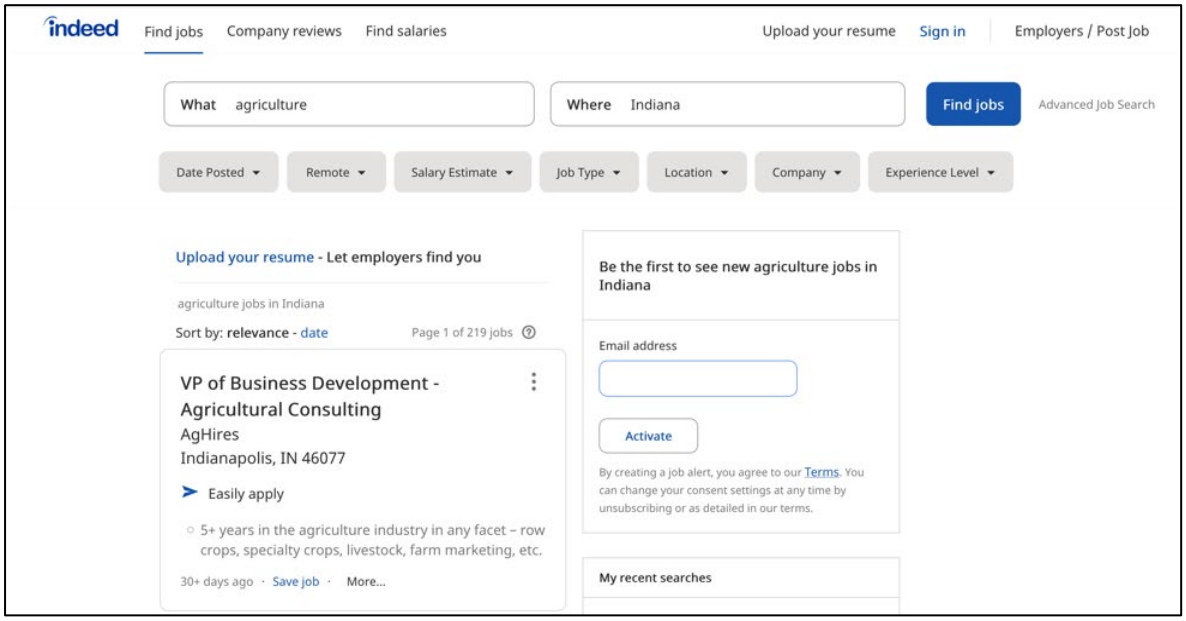

Figure 4: Job posting page

Company	Job_title	Location	Description	URL
Purdue University	Research Associate – Agricultural and Environmental ...	West Lafayette, IN	The Research Associate will manage a field research p...	https://www.indeed.com/rc/clk?jk=2a200fd5e53046...
Purdue University	Professor Assistant	West Lafayette, IN	The Agronomy Department of the College of Agricult...	https://www.indeed.com/rc/clk?jk=d68a6a20b18bd...
Purdue University	Laboratory Accessioning Clerk	West Lafayette, IN	The successful candidate will serve as an accessionin...	https://www.indeed.com/rc/clk?jk=287517f13a05ec...
Purdue University	Feed Administrator	West Lafayette, IN	Bachelor degree in animal science, veterinary science,...	https://www.indeed.com/rc/clk?jk=923c6e42a7cba8...
Purdue University	Dir, Lab Animal Prog/Central Animal Care	West Lafayette, IN	The Associate Vice President for Animal Resources an...	https://www.indeed.com/rc/clk?jk=b3891a6410098...
Purdue University	Farm Manager (ACRE)	West Lafayette, IN	Willingness to host touring agricultural groups and in...	https://www.indeed.com/rc/clk?jk=3052653308aae9...
Purdue University	Lecturer, Senior	West Lafayette, IN	This position will develop and administer laboratory-...	https://www.indeed.com/rc/clk?jk=f2646cd062b375...
Purdue University	Post Doctoral Research Associate	West Lafayette, IN	The successful applicant will hold a Ph.D. in agricultu...	https://www.indeed.com/rc/clk?jk=67ef0746d76f5e...
Purdue University	Assistant Professor of Ag Economics	West Lafayette, IN	The Department of Agricultural Economics is seeking ...	https://www.indeed.com/rc/clk?jk=c3fcf05fa2c5730...
Purdue University	Assistant Professor Botany & Plant Pathology, Turfgra...	West Lafayette, IN	The Department of Botany and Plant Pathology at Pur...	https://www.indeed.com/rc/clk?jk=835079027f309b...
Purdue University	Extension Educator – Ag and Natural Resources – Gib...	Princeton, IN + locations	Dedicated to helping improve the quality of life for cu...	https://www.indeed.com/rc/clk?jk=8fcf249b866514...
Purdue University	Assistant Professor of Animal Sciences, Muscle Biologist	West Lafayette, IN	The Department of Animal Sciences at Purdue Univers...	https://www.indeed.com/rc/clk?jk=f5cf257491f1b1b...
Purdue University	Extension Educator – Urban Agriculture – Vanderburg...	Evansville, IN (Downtown area)	Dedicated to helping improve the quality of life for cu...	https://www.indeed.com/rc/clk?jk=920d7e45190dce...

Figure 5: Job information data frame.

The next step requires the collection of description text. Collecting the descriptions of each job listing requires looping through each iteration in the URL column. A second web scraping script browses through each URL. This script extracts the job description text that outlines the requirements of employment, responsibilities, and benefits pertaining to each specific job. Figure 6

illustrates the layout of the description page for a specific job listing. The data, that is extracted and stored into a data frame, is the text in the "Full Job Description" section of the page, as shown in Figure 6. As job descriptions are collected, they are uploaded and stored in a new column named "URL Description." Each full job description is saved in the indexed row and column that is specific to the job title, company, and location. This organization allows for each row to represent a specific job listing.

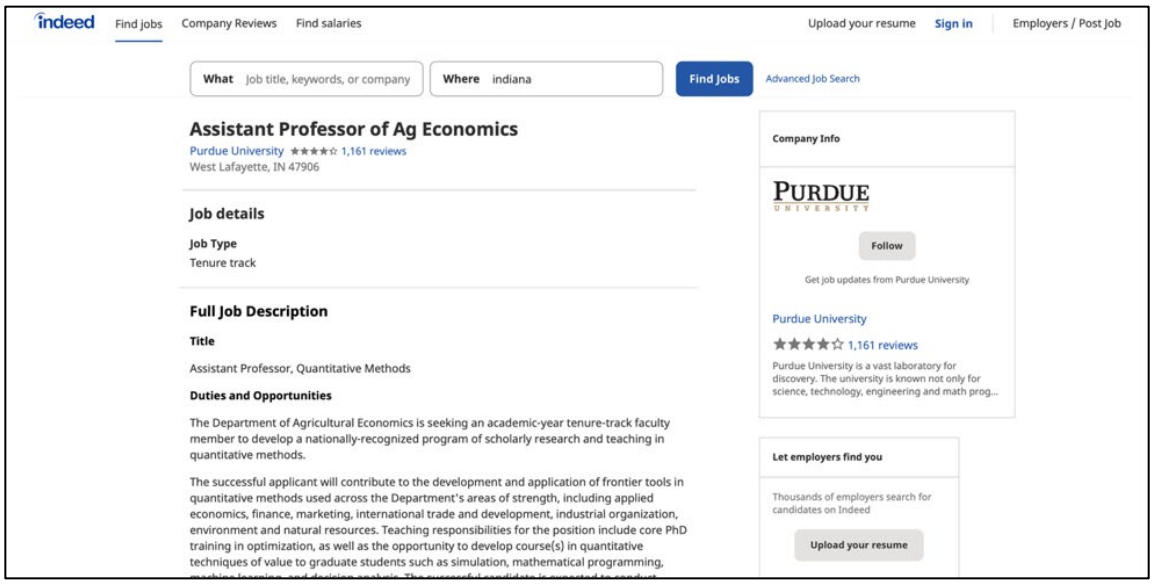

Figure 6: Job description page

Data Cleaning

Lastly, the data needs to be cleaned before analyzing. When initially collecting data, I noticed that some job listings were duplicating. This duplication was most likely created when the original web scraping script ran through the pages of the website. It is likely that some of the jobs were appearing on multiple pages as the script collected data. The duplicated job listings are eliminated to limit data errors and mismeasurement. This step required a simple function on the entire data frame. The "distinct function" cycles through the entire data frame by row and retains the unique rows while eliminating any duplicated rows. This function allows for the data to only represent distinct job listings. At the conclusion, the data can be analyzed.

Findings

After sourcing, collection, and cleaning, the data was ready for examination. The first and easiest questions to address was: What is the distribution of agricultural related jobs among Corn Belt states?

	Indiana		Illinois		Ohio		Missouri		Iowa		Total	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Job listings	269	17%	427	26%	181	11%	307	19%	438	27%	1622	100%

Table 1: Agricultural Related Jobs among states

Table 1 details the count and percentage of jobs listings by state. Figure 6 illustrates the differences in job opportunities among Corn Belt states. Agricultural related job listings are not equally divided between each state. Iowa sits at the highest with 438 job listings accounting for 27% of observations. Meanwhile, Ohio is last with 181 job listings that represent 11% of listings. This frequency table easily illustrates how the data can be used to explain characteristics of job listings within the Corn Belt.

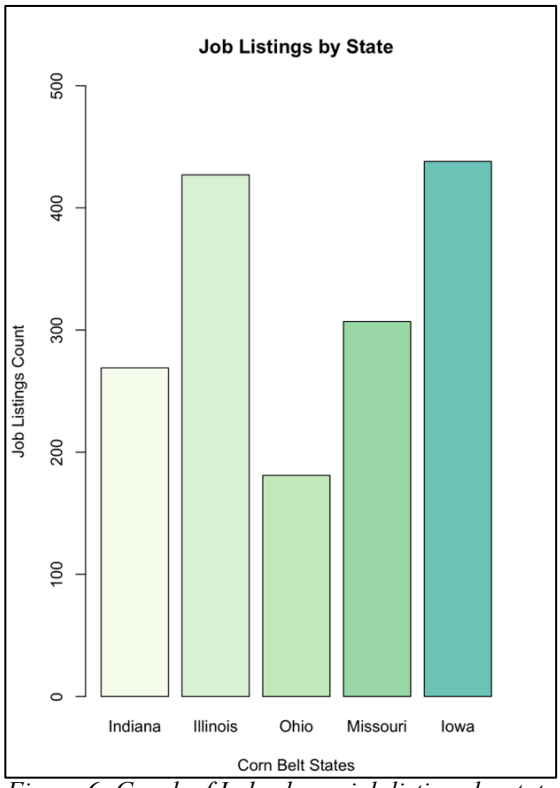

Figure 6: Graph of Indeed.com job listings by state

Attribute Identification

From all of the job descriptions, attributes needed to be collected and summarized. The attributes consist of desired skills, career characteristics, and working environments. Table 2 lists each concept with a list of related terms. For example, I am interested in job listings that mention experience with data analysis skills. These job listings were identified by searching for jobs with the words: "data," "data analysis," "big data," "data analyst," "data wrangling," and "data collection." Any job listings that used words from the data skills list in the description was categorized as requiring data analysis skills.

Counting Attribute Mentions

I wrote a script in R that takes a list of related terms for the attribute and counts every time any word from these lists appears in the description of job listings. The code loops through each row of the description column and counts the words matched from the lists. If the amount of term mentions is greater than or equal to 1, then the individual job descriptions are counted as containing the term. Using this code, the listings were categorized into job listing attributes. Table 2 shows an arranged list of all attributes and similar terms that justify categorizing job listings.

Table 3 shows the count and share of job listing with each set of attributes as defined in Table 2, by state. For example, the Indiana column shows that 30% of Indiana agricultural related job listings mention data skills. From this data, one can logically say that the highest demand for data skills within the Corn Belt is in Illinois. This data could identify where potential jobs are likely to be sourced and where demand is located.

Job Attributes	Term List
Benefits	Benefit, Retirement, Health, Dental, Vision, 401k, Gym, Pension, Union, Company Vehicle, Company Truck
Communication	Communication, Communicate, Talk, Call, Email, Explain, Present, Speak
Compensation	Compensation, Wage, Wages, Earning, Earnings, Pay, Pays, Paying, Income, Salary, Commission
Critical Thinking	Critical Thinking, Problem Solving, Analysis, Deliberate, Brainstorm, Reasoning
Data Skills	Data, Data Analysis, Big Data, Data Analyst, Data Wrangling, Data Collection
Education	Degree, Degrees, College, Graduate, Undergrad, Doctorate, PhD, Bachelor, Bachelors, Education, GPA
Experience	Work experience, Experience, Training, Prior job, Familiarity, Practice, Knowledge, Background, Exposure to, Understanding of
Leadership	Leadership, Lead, Leading, Leader, Manager, Management, Mentor, Supervision, Supervisor, Supervise
Programming Skills	Programming, Program, Coding, Code, Python, R, SQL, Java, C, Computing, Algorithm
Public Speaking	Public Speaking, Speaking, Speech, Speeches, Speak, Present, Presentation
Quantitative Skills	Quantitative, Analysis, Analyze, Analyst, Insight, Predict, Prediction, Predictive, Math, Mathematics, Statistics, Statistical, Optimization, Optimize, Regression
Remote Working	Remote, Remote Working, Home Office, Work from Home, Work at Home
Team	Collaboration, Team, Teams, Teamwork
Technical Skills	Technology, Microsoft, Microsoft Excel, Excel, spreadsheets, Microsoft word, PowerPoint
Travel	Travel, Drive, Driving, Fly, Flying

Table 2: List of Job Listing Attributes

Job Attributes	Indiana		Illinois		Ohio		Missouri		Iowa		Total	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Benefits	221	82%	321	75%	126	70%	215	70%	319	73%	1202	74%
Communication	217	81%	344	81%	151	84%	255	83%	335	77%	1301	80%
Compensation	173	65%	192	45%	113	63%	170	56%	220	50%	868	54%
Critical Thinking	62	23%	128	30%	38	21%	83	27%	96	22%	403	25%
Data Skills	81	30%	140	33%	61	34%	110	36%	106	24%	494	30%
Education	186	69%	335	79%	126	70%	233	76%	303	69%	1181	73%
Experience	261	97%	410	96%	174	97%	299	98%	422	97%	1566	97%
Leadership	236	88%	370	87%	144	80%	267	87%	364	83%	1381	85%
Programming Skills	18	7%	28	7%	11	6%	30	10%	19	4%	102	6%
Public Speaking	88	33%	160	38%	66	37%	125	41%	129	30%	564	35%
Quantitative Skills	114	43%	231	54%	70	39%	144	47%	167	38%	722	45%
Remote Working	39	15%	55	13%	28	16%	42	14%	56	13%	216	13%
Team	179	67%	310	73%	126	70%	204	67%	312	71%	1130	70%
Technical Skills	164	61%	281	66%	110	61%	186	61%	265	61%	1005	62%
Travel	152	57%	243	57%	110	61%	165	54%	255	58%	923	57%

Table 3: Distribution of Job Listings Attributes

Job Attribute Overlapping

In Table 3, the columns do not sum to 100% because many of these job attributes are overlapping. The data consists of many jobs that require several skills and offer different working environments. By creating categorical values that overlap, there is an opportunity to search for jobs that require a specific combination of skills. The Ven diagram in Figure 7 helps illustrate one notable example of overlapping skills. Figure 7 demonstrates the overlap between jobs that require data skills and quantitative skills. For example, one job that may require only data skills is field scouting. Often, field scouts walk fields and collect data, but do not make decisions from the data. A job that may require just quantitative skills would be that of a general engineer that uses math or problem solving to come up with solutions or create models. A job that could encompass both of these skill sets is a pricing analyst that observes data and uses predictive models to create decisions. R and the data collected in for this research paper can effectively be used to identify jobs with specific attribute overlap.

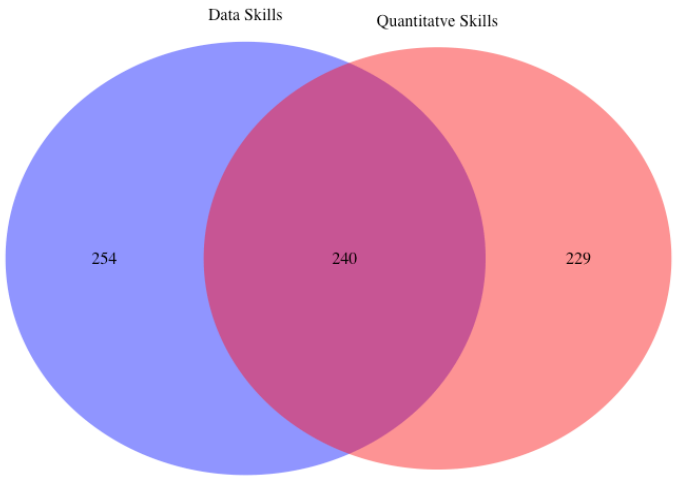

Figure 7: Agricultural jobs listings that require data skills and quantitative skills

Future Steps for Attribute Identification

By quantifying these job attributes, it is possible to track changes overtime. Data that is collected over a series of weeks, months, or years could distinguish trends in job attributes. For example, if data was collected from 2019 it would likely have fewer remote working jobs than 2020 and 2021, due to the effects of COVID-19 on the job market. Continued web scraping could determine if jobs in the future are more likely to be remote or based out of an office. From this data one could even track the growth of particular skill requirements, such as data management, quantitative reasoning, and computer programming. More data could aid in tracking the growth of "Big Data" in the agricultural sector.

Job Categorization

The job postings were also categorized by job type in a similar manner. A script in R categorized the jobs by related terms. The job categories and the list of terms are shown in Table 4. The code counts every time any word from these lists appears in the description of job listings. The code loops through each row of the description column and counts the words matched from the lists. If the amount of term mentions is greater than or equal to 1, then the individual job descriptions are classified as containing the term. Therefore, job listings that possess terms related to categories are then classified as belonging to that specific category.

Job Categorization Overlapping

Most jobs within the agricultural sector are a mesh of various categories. For example, a sales manager for a seed company could be classified into multiple job categories. These categorized areas may include economics, sales, agronomy, and farming. Allowing overlap of job categories for a specific job listing makes data analysis much easier. This characteristic also limits the number of individual job categories. As job listings become unique, accurate categorization becomes intractable. Text mining for specific terms helps avoid the taxing and time-consuming process of reading through each job description and manually categorizing job listings.

Job Categories	Term List
Ag Economics	Ag Economics, Agriculture Economics, Economic, Ag Business, Agricultural Business, Agri-business, Econometrics, Market, Market Research, Market Analysis
Ag Education	Ag Education, Agricultural Education, Ag Ed, Extension educator, Professor, FFA, Future Farmers of America, 4-H, Teacher, Ag Teacher, Agricultural Teacher
Ag Engineering	Ag Engineering, Agricultural Engineering, Ag Systems Management, Agricultural Systems Management, ASM, Mechanic, Diesel Mechanic
Ag Sales	Ag Sales, Agricultural Sales, Sales, Sell, Customer, Account, Seller, Buyer, Selling, Buying, Commission
Agronomy	Agronomy, Agronomist, Corn, Soybeans, Soil, Seed, Crop, Crop Adviser, Row Crop
Animal Science	Animal Science, Animal, Veterinarian, Sheep, Cattle, Dairy, Swine, Pig, Pork, Beef, Livestock, Breed, Poultry, Equestrian, Horse
Farming	Farming, Farmhand, Planting, Plant, Harvest, Harvester, Farm, forage, Combine, Farmer
Food Science	Food Science, Food Production, Meat Packing, Butcher, Food Distribution, Food Handling, Food Processes, Food Product, Meat, Produce
Natural Resources	Natural Resources, Conservation, Wildlife, Environment, Environmental, Ecosystem, Raw Materials, Nature, Energy, Pollution, Renewable

Table 4: List of Job Categories

Job Categories	Indiana		Illinois		Ohio		Missouri		Iowa		Total	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
Ag Economics	104	39%	179	42%	80	44%	100	33%	161	37%	621	38%
Ag Education	11	4%	18	4%	9	5%	6	2%	14	3%	54	3%
Ag Engineering	109	41%	115	27%	46	26%	92	30%	137	31%	497	31%
Ag Sales	217	81%	337	79%	130	72%	221	72%	342	78%	1247	77%
Agronomy	93	35%	188	44%	37	21%	86	28%	161	37%	562	35%
Animal Science	72	27%	137	32%	30	17%	71	23%	86	20%	392	24%
Farming	154	57%	237	56%	82	46%	151	49%	243	56%	863	53%
Food Science	50	19%	78	18%	41	23%	68	22%	84	19%	317	20%
Natural Resources	195	34%	182	43%	48	27%	91	30%	150	34%	558	34%

Table 5: List of Job Categories

Table 5 illustrates the results from categorizing job listings by mentions in the job description data. Table 5 describes the number of jobs listings that meet a category's requirements within each state. The analysis indicates that there are 104 jobs in Indiana that are defined as agricultural economic jobs. Agricultural economics jobs make up 39% of agricultural related job listings in Indiana. The data also shows that a majority of agricultural job listings within the Corn Belt are agricultural

sales jobs at 77%. Two assumptions that can be made from the data. Either there is a low demand for jobs in agricultural education or Indeed.com is not a good resource for identifying these jobs. These assumptions were made because only 3% of agricultural job listings in the Corn Belt are categorized as agricultural education.

Future Steps for Job Categorization

In future projects, I would suggest the term list for each job category be expanded to encompass more job possibilities. I also suggest creating a new job category that targets academic/higher education positions. The term "ag education" is very broad, but that also makes the data easy to categorize. There are definitely tradeoffs that exist when choosing terms to match the categories. This is beneficial to recognize for future research because term lists could be tailored to search for more specific job categories.

Mapping by Job Category

Since all of the data was categorized, it is possible to map job categories by location. Figure 8 shows the geographical locations of agricultural jobs categorized as having data skills.

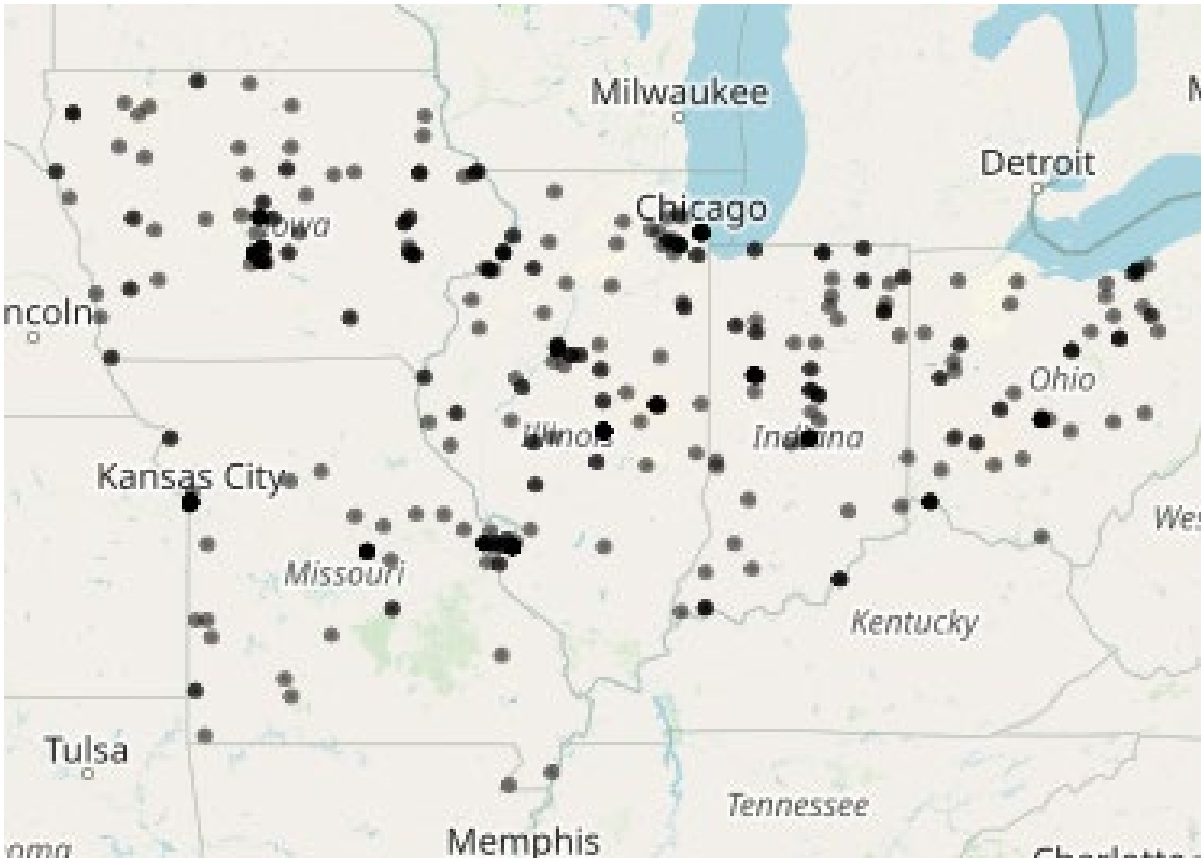

Figure 8: Map of Data Related Jobs

Conclusion and Future Research

The primary purpose of this research was to create a process for identifying specific jobs within a large pool of online listings. This project offered an opportunity to developed tools, within computer programming, that would collect data from existing sources and analyze job listings faster than a manual process. This project offered an amazing learning experience to apply data analysis to create a solution. I found that searching for jobs online can be challenging, especially when seeking opportunities that require unique skills. Through this research, I created a process to pinpoint agricultural jobs without the taxing work of scrolling and reading online job listings. This research project expressed a solution for searching for unique jobs with specific skill requirements.

Web scraping the agricultural industry's job market has opened doors for potential research. Future research may include tracking skill requirements change over time, job categories change over time, and change in job listings over time. The future research opportunities are vast and only limited by the programming skills of researchers. Future research like this could benefit the College of Agriculture at Purdue University. Findings in these research areas could help define future areas of study and curriculum. If "modern agriculture requires modern solutions," I believe data analysis and computer programming offer the tools to find these modern solutions.

References

Indeed. "About Indeed." *Indeed*, 2021, www.indeed.com/about.

Refsnes Data. *HTML Tutorial*, 2021, www.w3schools.com/html/default.asp.

Schmidt, Pascal. "A Detailed Guide to Web Scraping Indeed Jobs With R and Rvest." *A Detailed Guide to Web Scraping Indeed Jobs With R and Rvest* , 1 Nov. 2018, thatdatatho.com/web-scraping-indeed-jobs-r-rvest/.

United States Department of Labor. "Charts Related to the Latest 'The Employment Situation' News Release | More Chart Packages." *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, Jan. 2021, www.bls.gov/charts/employment-situation/civilian-unemployment-rate.htm#.

Whickham, Hadley. "Easily Harvest (Scrape) Web Pages." *Rvest Part of the Tidyverse*, 2014, rvest.tidyverse.org/.