Spring 2022

Survey Analysis: Senior and One-Year Out Surveys

Paige Phillips

Honors Senior Thesis
Release Agreement Form

The High Library supports the preservation and dissemination of all papers and projects completed as part of the requirements for the Elizabethtown College Honors Program (Honors Senior Thesis). Your signature on the following form confirms your authorship of this work and your permission for the High Library to make this work available. By agreeing to make it available, you are also agreeing to have this work included in the institutional repository, JayScholar. If you partnered with others in the creation of this work, your signature also confirms that you have obtained their permission to make this work available.

Should any concerns arise regarding making this work available, faculty advisors may contact the Director of the High Library to discuss the available options.

Release Agreement
I, as the author of this work, do hereby grant to Elizabethtown College and the High Library a non-exclusive worldwide license to reproduce and distribute my project, in whole or in part, in all forms of media, including but not limited to electronic media, now or hereafter known, subject to the following terms and conditions:

Copyright
No copyrights are transferred by this agreement, so I, as the author, retain all rights to the work, including but not limited to the right to use in future works (such as articles or books). With this submission, I represent that any third-party content included in the project has been used with permission from the copyright holder(s) or falls within fair use under United States copyright law (http://www.copyright.gov/title17/92chap1.html#107).

Access and Use
The work will be preserved and made available for educational purposes only. Signing this document does not endorse or authorize the commercial use of the content. I do not, however, hold Elizabethtown College or the High Library responsible for third party use of this content.

Term
This agreement will remain in effect unless permission is withdrawn by the author via written request to the High Library.

Signature: Paige Phillips **Date:** 5/2/22

Survey Analysis: Senior and One-Year Out Surveys

By

Paige Phillips

This thesis is submitted in fulfillment of the requirements for the Elizabethtown College Honors Program

Due Date: 5/2/22

Thesis Advisor (signature required)____Peilong Li_____

Second Reader _____Debra Sheesley_____

Third Reader _____
[only if applicable such as with interdisciplinary theses]

Survey Analysis: Senior and One-Year Out Surveys

Paige Phillips
phillipsp@etown.edu

ABSTRACT

Through surveys colleges are able to collect and analyze data pertaining to the institution's alumni, including valuable information about employment post-graduation. The number of alumni employed after graduation is an important statistic for colleges. In addition, the location, company, industry sector, and relevance to major are other aspects that can be analyzed by the college. Collecting alumni information not only provides pertinent employment status but can also maintain alumni connections. Therefore, this project works to organize and analyze alumni major feedback collected through alumni surveys, in order to provide pertinent information for departments and the college.

Keywords

Alumni, survey, senior, employment, graduation LinkedIn, data scraping, employment

1. INTRODUCTION

Alumni information collection is an integral part of a college's decision making, as well showing fulfillment of their mission statement.

Alumni networks are also important for company connections. Current students are able to reach out to alumni known within a particular industry. This boosts the employment of current students upon graduation, and even continuing students through internships. Building up this robust network of employers of alumni is the

Alumni information is not only crucial to the college itself, but it is also important for comparison amongst peer institutions and the national or state averages. This comparison, or benchmarking, will allow Elizabethtown to explore its competitive advantages, as well as where the college needs to improve in order to maintain the comparative standards. Identifying what sets Elizabethtown apart is significant in building a brand for attracting prospective students. Therefore, analyzing graduate statistics is substantial in the continuation of college, maintaining or growing enrollment, as well as strengthening academic programs.

Academic programs can be strengthened through analyzing graduate employment status, place, and industry relevance. Understanding the percentage of graduates that are able to gain employment is a drawing point for programs, and also a comparative benchmark within departments in order to identify high or low points within their programs. If a department identifies that their graduates are not being employed in industries related to their majors, then the program may need special inspection and reworking in order to generate successful graduates.

Through analysis of surveys taken by coming graduates and alumni, the college can gain insightful information regarding the students' college experience as well as important feedback regarding their major and the college in general.

Understanding how students felt about their college experience is integral to the progress of the college. Adjustments to curriculum, departments, and college programs can be made based upon repeated feedback and call for changes. If there are apparent issues that are raised within the surveys, those can be addressed and investigated by the college. Making changes and listening to student will not only create a better college experience for current students, there will most likely me a positive implication on the side of marketing and admissions. Prospective students are able to see and benefit from college improvements. This is with all colleges, not just Etown. Colleges need to be able to adapt to student needs and feedback to create a positive learning and living environment. Analysis of these survey responses will aid the college in the area of recruiting and admissions.

In addition to college changes overall, it is important that departments consider feedback from alumni relating to the real-world work that they are completing. The goal of major courses and programs are to prepare students properly for the field they want to enter. Therefore, alumni feedback is very important in keeping the major programs inline with current and important preparations and skills needed upon graduation.

The departments and schools at Elizabethtown should also be interested in the employment information contained within the alumni surveys. There are fields pertaining to the timing of job offers that include selection such as at least 6 months before graduation, within 6 months of graduation, and several options about timing after graduation. In addition, there are questions related to the positions currently held and the relation of their current job to their major. Employment upon graduation is obviously a very important topic within departments as well as a key aspect of the college recruiting. Therefore, it is imperative for departments to understand if their graduates are getting jobs, what positions they are holding, and if those jobs are related to their major. Not only will this aid current students, but major programs can be adjusted to better suit the related industry if needed.

Overall, the alumni network is an important source of support and information, for both the college administrators and current students. Alumni are often helping in finding current students internship and jobs. Therefore, it is important to maintain the alumni network, and understand the important information that can be gained from it. foundation for a successful continuation of student employment.

Currently, the vast majority of Elizabethtown's alumni data collection is through surveys. These surveys are sent to alumni through emails that are held on record as their last known active personal email address. Therefore, there is a possibility that these emails have been deactivated or are completely ignored. In addition, these surveys are voluntary; therefore, this introduces the voluntary response bias in which only those who would like to respond are counted in the information collection. Students who are potentially unemployed, working in an industry outside of their major, or who deem their job less than their potential, may be less likely to answer the survey. Overall, there is an issue of potentially not being able to alumni, as well as a willingness to respond. Both

of these issues contribute to a lack of alumni information being gathered.

In addition to the answering of these surveys causing some potential issues, there is the fact that the surveys are very time consuming and manual. The surveys are emailed out at the graduation milestones, such as 1, 5, and 10 years post-graduation. Not only is the college hoping that alumni are answering the surveys to begin with, but then there is the fact that they would have to answer 10 years out of college, in a survey that is sent to their last know email address.

Overall, there is some streamlining of the alumni data collection process that can be aided by data mining techniques. Scraping LinkedIn profiles of alumni can provide information pertaining to employment, credentials, and industry that can be tracked throughout the years since their graduation from Elizabethtown. This would be a continuation of the project of alumni analysis.

2. DESIGN

The basis pipeline of the data processing in outlined in Figure 1. Each component of the design process will be broken down in the following sections.

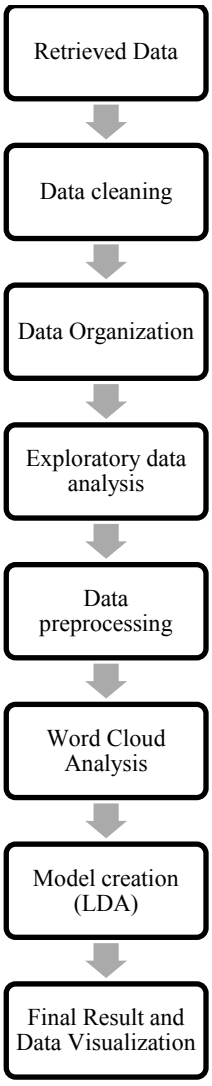

Figure 1. Data pipeline for project.

2.1 Data Layout

The senior survey is sent to graduating seniors within the last month of their college career. The senior survey consists of approximately 100 questions or question components. The senior survey contains questions pertaining to:

- Satisfaction rating of major and major advising
- Strengths/skill that were developed through college
- Impact of internships/student activities
- Feedback for majors and college
- A college-wide area of focus

In recent years, within the past five years, the college-wide area of focus has been the CORE program and what adjustments can be made.

The one-year out survey is sent to Elizabethtown College alumni one year after graduation. The method of contact are the emails provided in the senior survey. The one-year out survey also contains approximately 100 questions or question components. The one-year out survey contains questions pertaining to:

- Satisfaction rating of major and major advising
- Timing of job offer in relation to graduation
- Relation of job to major
- Strengths/skill that were developed through college
- Feedback for majors and college

The senior survey and one-year out survey have overall similarities in the questions that are asked. However, the one-year out survey focuses more on the jobs held by the alumni, as well as the relation of the job to their college major and skills developed.

2.2 Data Preparation

The years that were focused on in this analysis were graduating years of 2017, 2018, and 2019. This is because the senior and one-year out surveys were both available and contained similar formatting.

The variables that were focused on were:

- Major and major department
- Major satisfaction rating
- Employment status
- Employment position
- Employment relation to major
- Would you pick Etown again if you started your college search over today?

The data preparation and organization were a large part of this project. Although each senior survey contained similar questions, there were some key fields that differed. Each year of the one-year out and the senior survey information was contained within separate excel files.

There were years of the surveys that did not contain student id numbers. This is the key identifier of students and acted as the method of joining the senior survey and one-year out survey information to collect those who answered both surveys. Therefore, to correct this information gap, the id numbers needed to be brought in. In addition, there were years of the survey that only contained the major department while other years only contained the individual major itself. Therefore, for years that contained the specific major, the major department was brought in.

In addition to the important field gaps, there were smaller adjustments that needed to be made. Although the answer choices

for all of the rating field contained Very Dissatisfied, Dissatisfied, Neither Satisfied nor Dissatisfied, Satisfied, and Very Satisfied the capitalization differed and therefore was adjusted for the aggregation of the data to be successful.

2.3 Exploratory Data Analysis

As previously stated, the graduating years of 2017, 2018, and 2019 were analyzed; therefore, it is important to understand the number of graduates that responded to each of the surveys.

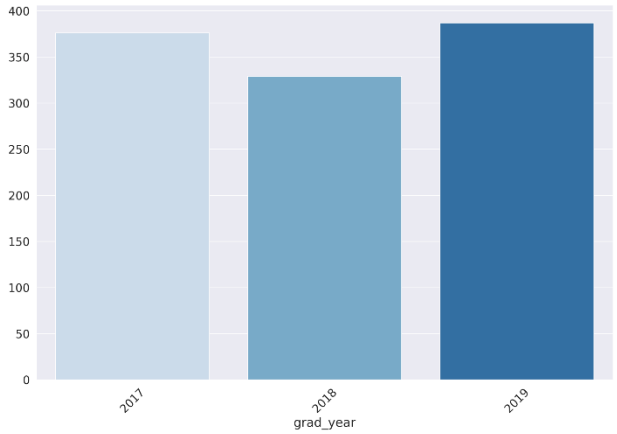

Figure 2. Count of Student Responses to Senior Survey.

As seen in Figure 2, there were approximately 375 responses to the senior survey in 2017, approximately 340 responses in 2018, and nearly 400 responses in 2019.

Each graduating year had a high percentage response rate of well over 90% in 2017 and 2019, and hovering right around 90% in 2018.

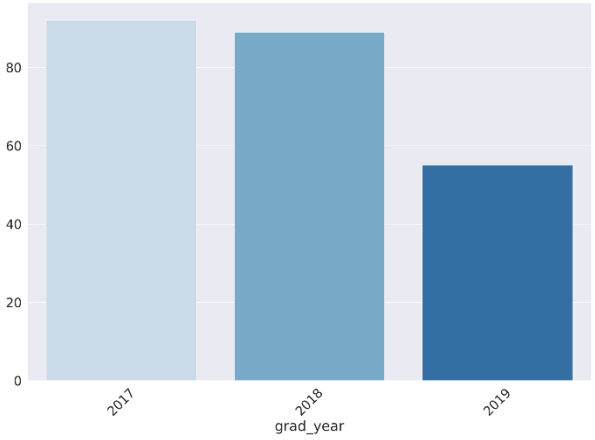

Figure 3. Count of Student Responses to One-Year Out Survey.

As seen in Figure 3, there were approximately 90 responses to the one-year out survey for the graduating class of 2017, approximately 85 for the graduating class of 2018, and approximately 55 for the graduating class of 2019.

Even though the graduating class of 2019 had the largest response to the senior survey, there was a significant drop in the number of graduates that responded to the one-year out survey.

Table 1. Top 5 Alumni count by State who responded to One-Year Out Survey

State	Number of Alumni
Pennsylvania	103
Maryland	12
New Jersey	10
Virginia	6
Delaware	5

From the alumni graduating between 2017-2019 who responded to the one-year out survey and provided an employment state, the vast majority reside in Pennsylvania. There are 103 alumni who work in Pennsylvania, with the next highest being Maryland with 12. The high concentration of respondent in Pennsylvania can be explored from two different angles: that Elizabethtown alumni predominantly remain within Pennsylvania upon graduation, or that the Elizabethtown alumni that remain in Pennsylvania feel a stronger connection to the college and obligation to respond. As a majority of students that attend Elizabethtown come from Pennsylvania, it makes sense that a majority would remain in the state; however, the sharp difference of any other state can be explored in the regionality of alumni.

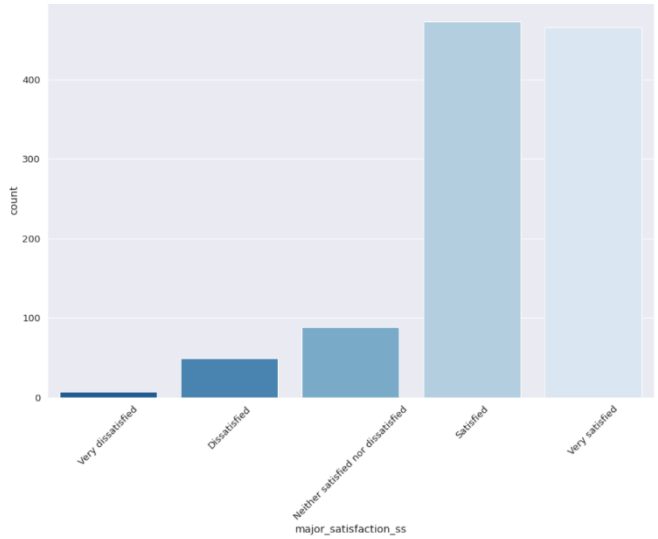

Figure 4. Count of Major Satisfaction on Senior Survey.

One other important aspect of the data to inspect before moving into sentiment analysis is the major satisfaction of the graduates who responded to the surveys. The count of graduates by major satisfaction for the senior survey can be seen in Figure 4.

The vast majority of students that responded were either satisfied or very Satisfied. The “Satisfied” grouping slightly outpaced the “Very Satisfied” grouping. There were just over 1200 students who responded to the senior, of which only approximately 150

responded that they were neither satisfied nor dissatisfied, dissatisfied, or very dissatisfied with their major.

Overall, the percentage of seniors who were pleased with their major neared 90%, which is an impressive number.

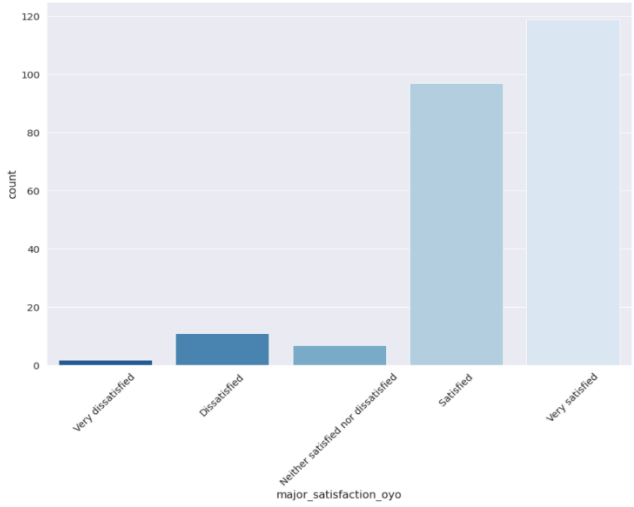

Figure 5. Count of Major Satisfaction on One-Year Out Survey.

The count of graduates by major satisfaction for the one-year out survey can be seen in Figure 5. As seen with the senior survey, the "Very Satisfied" and "Satisfied" grouping made up a majority of responses.

In the one-year out survey responses there are still all 5 categories being represented, and the "Dissatisfied" grouping is in similar ratio to the "Very Satisfied" as it was in the senior survey responses. Therefore, we are receiving different viewpoints in the responses, which will be important for the sentiment analysis that will be completed.

2.4 Model Creation and Analysis

The field of interest for the sentiment analysis is the major feedback contained within the one-year survey. Alumni were asked to provide any comments that related to their major field and/or department. This feedback field was completely open to any length or words that the student provided.

Initial analysis and visualization of the major feedback field occurred through word clouds. First, the feedback field from those who responded to this question were merged and transformed into a document term matrix. After aggregation, this matrix contained each individual word as the column headers, with the rows being the grouping of interest, and the values being the word count.

There were two groupings of interest that were used throughout this analysis. The first is by graduation year. This is to explore the difference in sentiments between students of different graduating years. The second is by the response to "Would you pick Etown again if you started your college search over today?" The options for this field were "Definitely Not", "Probably Not", "Not Sure", "Probably", and "Definitely".

Once the document term matrix was created, it was converted into a dictionary in which the individual word became the key and the number of occurrences of the word was the value. A common English stop words list was imported and these words were eliminated from the feedback dictionary. These stop words are commonly used words such as "a", "the", "is", and "are" which provide little useful information.

For grouping by graduation year, if a word was within the top occurring words of all three years (2017, 2018, and 2019) it was added to the list of stop words that were then removed from the feedback analysis. The commonly shared words were removed from the analysis as the difference in the sentiments by graduation years is the focus of the analysis. Therefore, removing the commonalities allows the variation between years to shine through.

For grouping by if they would pick Etown again, if a word was within the top occurring words of at least 3 of the 5 responses it was added to the list of stop words that were then removed from the feedback analysis. The commonly shared words were removed from the analysis as the difference in the sentiments by response to this question is the focus of the analysis. Therefore, removing the commonalities allows the variation between response groupings to be shown.

Once the feedback is parsed, organized, and filtered, the information is organized in a word cloud. The word cloud provides a visual representation of the words used by alumni in their major feedback on the one-year survey. The size of the words in the visualization represents the quantity that the words are used. Word clouds were generated for both of the data organizations: by graduation year, and by response to if they would pick Etown again.

For sentiment by graduation year, the TextBlob package was imported to evaluate the polarity and subjectivity of alumni feedback. Polarity scores are between [-1,1] in which -1 is a negative sentiment and +1 is a positive sentiment. Subjectivity scores lies between [0,1] where 0 refers to judgement and 1 refers to opinion. The subjectivity and polarity of each individual words are computed, and then the average of the entire feedback matrix is computed by taking the average of the word values.

For sentiment by response to picking Etown again, the feedback matrices were fed into a Latent Dirichlet Allocation (LDA) model. The LDA model is able to group similar words based on overarching topics that are identified. The response groupings were then fed into the model results in order to assign each response with a corresponding topic. Each response grouping had a unique topic assignment when run.

3. EXPERIMENT

Results of the model creation and analysis outlined above will be explained in the following sections.

3.1 Graduation Year

The major feedback from the one-year survey was organized by graduation year and was structured and filtered as outlined above. The word cloud generated from this data can be seen in Figure 6.

The graduating class of 2017 focused on words such as "major", "working", "job", and "professional." Overall, the Class of 2017 was highly focused on their professional career after graduation, and how their major pertained to their future work.

The graduating class of 2018 had a large focus on the word "job", as well as "experience", "graduate", "information", and "major." Overall, the Class of 2018 was highly focused on their job after graduation, as well as the course information and experience for their profession.

Figure 6. Word cloud of major sentiments by graduation year.

The graduating class of 2019 was particularly focused on "emphasis", "course", "focus", "concentration", and "include." Overall, the Class of 2019 was concentrated on coursework, including the emphasis, focus, and inclusion of information in their course.

There is a similarity in the words and focus of feedback for the Class of 2017 and Class of 2018 in terms of emphasis on relation of major to jobs, as well as profession experience. There is a shift in the Class of 2019 towards more focus on the course material, as well as the emphasis and inclusion of material.

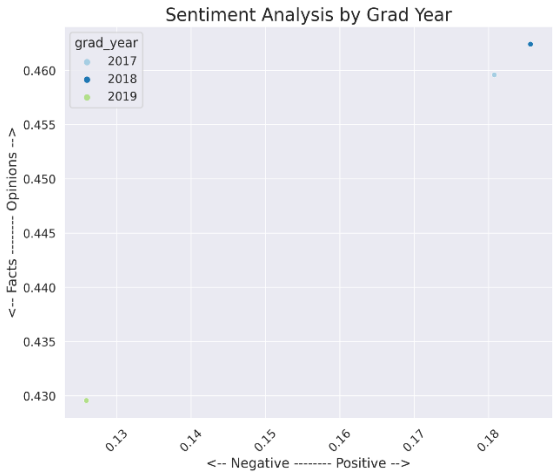

Figure 7. Polarity vs Subjectivity of major sentiments by graduation year.

The TextBlob package was applied to the feedback organized by graduation year. The subjectivity and polarity ratings of the major sentiments by graduation year can be seen in Figure 7. The Class of 2018 had the highest polarity score, meaning the feedback contained the most positive sentiments.

The polarity of 2017 alumni was similar in positivity. However, those who graduated in 2019 had a noticeable decrease in polarity score. This indicates that those who graduated in 2019 had the most negative comments about their major programs.

In addition to the polarity, it can seen that there is a difference in the subjectivity of 2017, 2018, and 2019 graduating years. Both 2017 and 2018 have comments based on opinions while those who graduated in 2019 have feedback more rooted in facts.

The Class of 2019 has the most negative major feedback out of the three graduating classes, and focused more on the course material and emphasis. The reason for the shift to more negative comments in 2019 may be an area that the college would be interested in investigating further.

3.2 Would You Pick Etown Again?

In the one-year out survey, the college posed the question "Would you pick Etown again if you started your college search over today?" The answer options are "Definitely Not", "Probably Not", "Not Sure", "Probably", and "Definitely." The outlay of responses for graduation classes between 2017-2019 who responded to the one-year out survey can be seen in Figure 8.

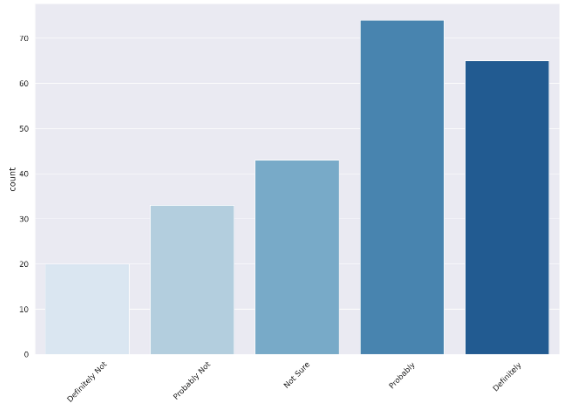

Figure 8. Count of "Would you pick Etown again?" from one-year out survey.

The majority of alumni selected "Definitely" or "Probably." However, approximately 40% of students selected either "Not Sure", "Probably Not", or "Definitely Not." This is a decent percent of students who were unsure or said they would not choose Elizabethtown again, and the reasoning behind these responses should be inspected.

The major feedback from the one-year survey was organized by response to this question and was arranged and filtered as outlined above. The word cloud generated from this data can be seen in Figure 9.

Those who responded "Definitely Not" focused on words such as "application", "curriculum", "changed", "department", and "hire." This feedback focused on changes to departments and curriculum, as well as the need for hiring new faculty.

Those who responded "Probably Not" concentrated on words such as "urban", "career", "education", "department", and "college." Some students did not like the setting of Etown, as well as focused on preparation of their departments and education for their future career.

Those who responded "Not Sure" focused on "marketing", "graduate", "concentration", "real world", and "required." Students in this category centered on the major and graduation requirements, as well as connection to real-world work.

Figure 9. Word cloud of major sentiments by "Would you pick Etown again?"

Those who responded "Probably" concentration on "focus", "experience", "work", "help", and "like." Students centered on the experience and focus of coursework that relates to their work upon graduation, as well as the help received from professors.

Those who responded "Definitely" focused on "Dr.", "writing", "understand", "course", and "teaching". Students within this category provided feedback related to the teaching they received in their courses, and well as the understanding achieved through professor help.

There is a difference in the sentiments of those who say they would not choose Etown again and those who say they would. Those who would not choose Etown tend to focus on changes within departments, lack of opportunities, and the curriculum requirements. On the other hand, those who say that they would choose Etown again tend to focus on the professors, specifically the understanding and help they received through the teaching.

Table 2. Top sentiment groupings from LDA of "Would you pick Etown again?" from one-year out survey.

pick_etown	Top sentiments
Definitely Not	Department, curriculum, changed, opportunities, applications
Probably Not	Career, urban, department, work, education, college
Not Sure	Marketing, world, graduate, liked, required, concentration
Probably	Focus, work, help, course, skills, practice, emphasis
Definitely	Dr, writing, senior, understand, course, teaching

The data organized by response to the Etown selection question was fed into a LDA model in order to receive the most prevalent topics in each of the feedbacks.

The results from this model analysis are similar to the word clouds that were generated. The top 5 groupings can be seen within Table 2. As stated with the word cloud, the categories of "Definitely Not", "Probably Not", and "Not Sure" focused on required curriculum, and departmental changes. The categories of "Probably" and "Definitely" focus on skills and understanding built through their courses and professors.

4. FURTHER APPLICATIONS

Continued application of the survey analysis could be important for the college as there is an abundance of information that can be derived from graduating seniors and alumni. This could involve inserting new targeted questions into the surveys, and tracking the responses throughout the senior survey, one-year out, five-year out, and ten-year out surveys.

There is a large variety of models and information that can be drawn from the alumni data. This will most likely be driven by what the college is particularly looking for within the Department of Institutional Research reporting and the information that the Office of Alumni Relations is interested in.

One example of a potential model is a prediction of whether a student will get a job. This can be achieved by feeding in pulled data concerning alumni within a student's particular major and their success in finding a job. Not necessarily within this project, but for the college they can combine this alumni program information with particular students' academic success, GPA, test scores, and other aspects in order to predict if a student will get a job based on success of prior alumni in these same aspects.

In addition to being able to get a job, whether an alumnus goes into academia versus industry can be investigated. This would be an interesting statistic for many departments and could shift the focus of certain classes, or lead to the offering of different or additional classes.

In the Department of Institutional Research, we currently maintain Tableau workbooks containing admissions, enrollment, course information, and other college information. However, currently we are very limited in the alumni data that is presented to the college administrators in an interactive form. Therefore, some work could be done in Tableau using the data collected in order to inform and provide relevant and interesting information. The alumni information is pertinent to many departments not only the Office of Alumni Relations. This includes the admissions and marketing teams, as alumni numbers can be used to promote the college, attracting prospective students.

If regression models are built on numbers such as graduation rates or number of percent of alumni getting a job upon graduation, then measures of accuracy such as R-squared would be most applicable. R-squared values, mean squared error, and mean absolute error are three values that can used to evaluate models. We would be able to use the data drawn from the data collected of past years to feed the model, and then compare these project values from the regression model to numbers that the college have calculated in the past. Therefore, this gives a basis for the validity of the model predictions. The accuracy measure can be used to prune the model, comparing between models to select the one that is the best fit for the data.

If classification models are built on groupings such as academia versus industry, which industry sector alumni end up working in, or what is their highest level of education, classification accuracy measures are most applicable. The classification report and confusion matrix will provide information that will also drive the success of the model. Comparison of precision, f1 scores and other accuracy metrics will determine which model is the best fit for

classifying the data. Metrics such as accuracy and loss can be used to compare models. For the most part, models with the highest accuracy and lowest loss will be the best fit for the data. However, other metrics such as the confusion matrix must be consulted in making that determination.

5. CONSIDERATIONS

Expansion into web scraping from LinkedIn can hopefully still be investigated in the future. Outlined below are existing solutions that involve LinkedIn information retrieval, as well as the shortcomings of those solutions. The development of the web scraping and data collection techniques would be extensive.

The existing problem in the alumni data collection space, particularly in LinkedIn is the accessibility of profile information. LinkedIn has put many regulations and guidelines in place to stop companies from scraping profile data for commercial use. Many companies had been scraping profile data for targeted marketing, which has led to an increase in regulations.

In addition, at Elizabethtown, the problem that exists is the lack of expansion of the collection procedures that are currently in place. There is a want for a more automated system of collecting the alumni data that will limit the amount of manual work that is associated with this undertaking. However, there is a lack of time in exploring additional ways of collecting data; therefore, this project is going to address this problem and hopefully help in providing a system for collecting this data with less manual work.

There are existing solutions that exist to this solution. However, a lot of these solutions are not published and if they are published, they are more general in their description. For the most part, private institutions or colleges utilizing data scraping techniques for data collection are not publishing their findings; therefore, the projects that are able to be found are smaller scale. Another shortcoming of the existing solutions is that the projects that were published are at least a couple of years old. This was before the additional regulations were put into place, and therefore, information may have been a little easier to access. Although, the general uses of the data collected from these methods are still applicable and useful in the case of Elizabethtown's alumni collection.

There are some existing solutions on scraping LinkedIn for data collection.

The first solution [2] is posted on LinkedIn itself. The contents of the article outline how to use selenium and python to scrape LinkedIn profiles. This solution is more on the technical side walking through the initial step of necessary downloads and automating LinkedIn login. Then it moves into searching LinkedIn profiles on google in which for loops are used to unpack the values for each element in a list of URLs. After these initial steps are completed, you are then able to move into scraping the data where the Inspect Element on the webpage is used to locate the needed information to correctly extract each data point. After these steps are completed, you are able to print the information including one's name, job title, company, location, and URL.

This solution is very technical; therefore, its shortcoming is that there is no application of the technique to a problem or actual data source. The solutions that focus on some applications of data scraping do not include the technical information, therefore I think this source is particularly important in providing some basic information of how to go about the collection of data within LinkedIn.

Another existing solution [3] revolves around extracting alumni information from social networking websites. The websites that are included in this report are Facebook and LinkedIn, with a heavy emphasis on Facebook. The purpose of this project was to extract profiles of alumni and retrieve relevant information. From Facebook, the study worked to fetch the first name, last name, gender, work, Facebook ID, and URL. This solution utilized HTML Parsing in order to detect templates of information sources, extract the information, and translate it into a relational form, or scraper. Graph API was the primary way that data was retrieved from Facebook.

One shortcoming of this solution is the basic information that is being retrieved from profiles. There is a lack of complexity in the information that is being pulled and stored. The solution that Elizabethtown is looking for is definitely going to contain more complication and larger fields than the existing solution. Therefore, it will have to be investigated if any of the techniques used in this paper would be applicable to the current problem. This solution mainly focused on Facebook, where my project will focus solely on LinkedIn.

The existing solution [1] focuses on a smaller program within a mid-sized university in the southeastern USA. It focused on the LinkedIn profiles of 175 graduates of an Information Systems program. The information that was pulled from the profiles included the graduation year, gender, additional education, job titles after 1, 5, 10, and 15 years after graduation, and the current hob title and employer. The problem that they ran into was that often times the profiles of alumni were incomplete. In particular, many alumni only included their job 10 or 15 years after graduation, or in many instances only included their current job. Therefore, there was a lack of information on the first jobs that the alumni held right out of college. From the data that was collected, the program was able to calculate the median graduation year, gender distribution, highest level of education, and frequency of job titles. All of these statistics can not only be studied within the scope of the college, but against state or national averages.

One major shortfall of this particular study was that they were only able to collect data one individual's profiles that accepted an invitation to an alumni group within LinkedIn. They found that many individuals did not accept the group invitation, and therefore they were unable to collect data from their profile This is a severe limitation for the fullness and representation within the data collected. Along this line, the study also saw many more females within the alumni group in proportion to the actual make-up of the graduates. Therefore, based on who had created a LinkedIn and then accepted the group invitation, the data was skewed from the actual composition of the full body of alumni. Another shortcoming of this study was the intense focus on the information system program. Because of this focus, the researchers were able to do more manual work or make assumptions that could not have been done with a larger group of alumni.

While there are some existing solutions to using data scraping to collect information on alumni, there are some shortcomings of these previous works. One of the solutions was strictly technical, which provided useful information; however, there was no application to a real-world problem. In addition, one of the studies was very focused on a small number of students in a particular program; therefore, the overall technique will most likely be more difficult with a much larger scale of students. Within that study as well, accepted invitations to a LinkedIn group were required which is definitely a limitation of this technique. This not only led to a

smaller number of students that they were able to collect data from, but the data also consisted of a greater number of females than there was proportionally in the graduating classes themselves.

In addition, many of the published works about this topic are a couple of years old, in which the restrictions within LinkedIn have strengthened, which will most likely make this task more difficult. In addition, technology shifts very quickly; therefore, some of the techniques or application that were used may be considered outdated nowadays.

One can anticipate running into several of the issues that were brought up in the previous studies including the access to user profiles. In addition, the information that is pulled from LinkedIn is dependent on alumni creating and maintaining a LinkedIn profile. If there is a low number of alumni that have profiles, the effectiveness of this type of data collection becomes nonexistent.

In addition to having a LinkedIn profile, the information is put in directly by the user. Therefore, we are dependent upon the truthfulness and correctness of the information the alumni are inputting. If there is incorrect data this could lead to an issue with the data analysis, where we are perceiving information as true, when in fact it is not. This could throw counts, projections, and percentages of alumni off, as well as causing models trained on this data to be not fully correct or true.

On the other hand, if incorrect is caught at the data preprocessing level, then it can be adjusted but this could create manual work for those upkeeping this program, when less manual work and easier collection of data is a main goal of this project. However, there is definitely data issues when data is manually entered into the system; therefore, at least some automation to the data entry process will hopefully lower the amount of corrections that need to be made.

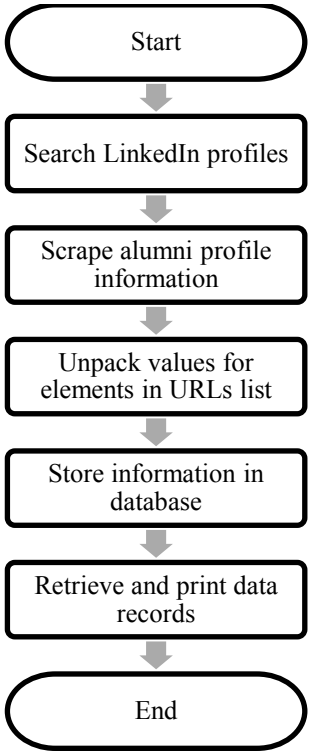

Figure 10. Data pipeline for data scraping.

The programming languages can be used would be determined when going through the project and seeing what fits best. The API the definite component, with something that most likely contain HTML parsing. Previous solutions used Selenium as their API, so that is a definite choice, and then that was used with Python.

The data scraping procedure will be the most difficult aspect of the development design. In addition, the information that can be retrieved from individual's profiles in limited.

6. RECOMMENDATIONS

For easier analysis and organization of data, both the senior survey and the one-year out survey should have consistent questions and key identifiers of students. There were four different ways of categorizing a students' major department within the surveys:

- major codes and major department codes
- major description and major department codes
- only major department codes
- major codes and major department descriptions.

The analysis and aggregation of the alumni survey data would be greatly benefited from having consistent questions and collection methods for this information. In order for all of the surveys to contain matching fields, manual importation of information would need to be completed. Therefore, these key fields need to be consistent on the side of the survey questions.

These questions should be pre-determined selections and not allow students to type their majors, as has been seen in past surveys. This alone causes manual identification and rework of errors. In addition to spelling errors, there could be inconsistencies in student responses.

In addition, student id should be included on all of the surveys as this is the key identifying marker for every student in the college database systems. Including student id allows the data to be joined on this is, allowing for each tracking of responses across alumni surveys. Containing the id number on the survey eliminates the manual work that comes with brining in these id numbers later.

The college should take an invested interest in analyzing and understanding the alumni responses. This alumni information can lead to changes within the college in relation to majors, and keeping up with changing industries.

Also, if a student states that they would not pick Elizabethtown again, the college should be seeking to understand why this is the case, and hopefully make adjustments in order to improve the college experience of current and prospective students.

7. REFERENCES

[1] Case, T. et al., 2012. A LinkedIn Analysis of Career Paths of Information Systems Alumni. *Journal of the Southern Association for Information Systems, Volume , Number 1, 2012.*

[2] Craven, D. 2018. Use selenium & python to scrape linkedin profiles. *Linkedin.com* (Oct 2018). DOI= https://www.linkedin.com/pulse/how-easy-scraping-data-from-linkedin-profiles-david-craven/

[3] Wagh, M. et al., 2015. A Web Scraper For Extracting Alumni Information From Social Networking Websites. DOI= https://www.ijset.in/wp-content/uploads/2015/05/042015.1948_445-448.pdf