Text Analysis: Exploratory Data Analysis for Introductory and Nonstatistical Learners

---

December 21, 2018

Abstract: Text analysis is the process of extracting information from text data. It has applications in fields ranging from statistics and computer science to linguistics and marketing. In this paper, we describe an introductory text analysis interface that we created and how it has benefits for developing statistical reasoning and exploratory data analysis skills. We start with a brief overview of basic text analysis terms and principles for introductory, and perhaps nonstatistical, audiences as well as its competency as a method of exploratory data analysis. We explore the benefits of an interactive interface within statistical education and place our interface within the PPDAC statistical education framework, specifically as it relates to exploratory data analysis for an introductory audience.

Note for Draft 1: There are a few places in the paper denoted in RED that simply aren't true about the module yet. These include things like adding lists of questions to each tab for the user to consider, further edits on making interface descriptions better and providing more examples, and including built in data sets. I include them in the draft because I don't want to add them in to the paper once they've been fixed.

Chapter 1: Introduction

At its core, text analysis is the process of extracting information from a body of text. This could be from novels, movie reviews, or any other body of text you could imagine. For instance, an English literature student want to see the results of an analysis of Tolstoy's *War and Peace*. This project is a user interactive text analysis interface. The interface guides the user through the process of uploading data, modifying the data, and producing plots that show information about the most frequently used words, sentiments throughout the text, and relationships between the words. The user can pick a built-in dataset or elect to use their own. The interface follows a specific flow – there are tabs that separate different concepts of text analysis, with each tab becoming slightly more complicated than the previous. There are descriptions and questions along the way that prompt the user to think about the actions they are making and what effect those may have on the displayed results.

The world of statistical education is constantly evolving and developing new techniques and approaches to teaching. There is a continual push to improve teaching materials and to encourage effective learning for students at an introductory level of statistics as well as for non-statistics students across the board. Part of the issue of learning statistics is that students are asked to apply statistical methodology before they have been taught about statistical reasoning (Bradstreet 1996). There are many questions that should be asked of a statistical project before the process of answering it with statistical models and algorithms begins. The user should ask themselves where the data came from, what types of variables are included, what limitations and issues there may be with the data collection and sampling. They should also have an exploratory look at all of the data before diving in to building models. Exploratory data analysis (EDA) is crucial for examining relationships in the data and for breaking preconceptions the user has about the data. Through exploration, the user often finds relationships and values in the data that they were not expecting to see. These revelations may prompt them to go back and modify the data until it has the structure and attributes that the user was hoping for.

This interface emphasizes the basics of text analysis by creating the building blocks that are necessary for understanding the data and the main ideas and attitudes in the text. Each of the plots presented in the interface tells us something unique that is crucial to conducting a thoughtful exploratory analysis of the chosen text data. These plots give people exposure to an area of statistics that they might not otherwise see but it presents it in a manner that is understandable with limited to no background. We want people to be able to extract meaning from text data without having to learn how to code or understand the fine details behind the results presented in the interface. One of the common problems I have encountered in introductory statistics courses is that the topics of the provided examples are not of interest to the learner. One of the primary goals of statistics is to learn through addressing real world problems (Bradstreet, 1996). This project allows the user to choose their own data source, so they can always choose a topic that is of interest to them. This will encourage more interaction with the project and should produce deeper levels of learning.

The interface is a surface level look at what can be done with text data. It is meant to be simple and easy to understand for all introductory audiences. In particular, students from disciplines in the humanities, such as Linguistics and English, may find this interface useful for summarizing their texts. It is also appropriate for introductory data science and statistics students who are learning about the importance of EDA in addressing statistical questions. Beyond the scope of students in the classroom, this interface could be useful to professionals in the marketing field who depend on the customer's sentiments when making business decisions.

The paper proceeds as follows:

- **Chapter 2**: A basic introduction to common terms and data used in text analysis. Additionally, I introduce the connection between text analysis and exploratory data analysis.
- **Chapter 3**: A discussion of the merits of an interactive user interface. This is followed by an example demonstration of the interface using J.M. Barrie's *Peter Pan.*
- **Chapter 4**: The main discussion of the role this interface plays in the dynamic field of statistical education. This includes an analysis of the project within a specific statistical education framework.
- **Chapter 5**: A final discussion of the paper, limitations of the project, and a look at possibilities for future work.

.

Chapter 2: Text Analysis Methods

Section 2.1: Text Analysis Basics

Text analysis is the process of extracting information from a body of text. This could be summarizing main themes by finding the most frequently used words, locating spots in the text where sad events occur by looking at the sentiments of the words being used, or finding strongly related words by looking at how often they occur together.

Before I explore the importance of this project within the larger realm of statistical education, I will give a brief overview of the interface structure as well as an overview of some of the more common terms used in introductory text analysis.

2.1.1: Data - what can we use and how is it processed?

The descriptions that follow detail the process of choosing and modifying the data. For reference to the interface, these steps occur in the “Data Upload” and “Token Variable and Cleaning” tabs that can be found in the left hand sidebar.

Text analysis can be done on anything from a collection of the lyrics from every Beatles song to your PhD dissertation. I have designed the interface to accept plain text and CSV files. This means that I cannot process Microsoft Word documents or PDF files, but the text from those can be easily transferred to a plain text file by copying the text and pasting it in a text file (which can be created in Notepad for windows or TextEdit on Macs).

The first step in working with any data set is to have a look at the raw data. A user should gather some information about the variables (if there are any) and get a sense of what is contained in the dataset. After giving that some thought, unnecessary data can be filtered out. For instance, as will be demonstrated in the example later on, I remove the table of contents when looking at *Peter Pan* because it does not hold any unique information that is not contained later on in the text. Another common portion of the text to remove is the copyright information. Once the user chooses which lines to remove, the modified data is displayed and they are asked to look over it once more. If they are not satisfied with the result, they are welcome to go back a step and try removing different lines.

Text data is full of common words such as “like”, “because”, and “the”. These words do not give us much context about the unique topics in the text, so we often choose to remove them before conducting any text analysis. We call these “stop words”. The interface is designed to remove stop words by default, but there is an option to keep them in if the user would like to see how they affect the results. In fact, we encourage the user to compare the results both with stop words left in and with them removed. This is a good exercise for understanding how different ways of filtering and subsetting data can give dramatically different results. The interface removes stop words from a dataframe within the tidytext package that is a compilation of English stop words from three different sources (Silge & Robinson, 2017). The user is also given the option to remove words that they would like to add to the stop word list. This pushes

them to think about what words they may not want to see in their analyses for one reason or another. The user is free to return to this step later on in the process if they see or remember words that they want to remove retroactively.

The last important part of preprocessing the data before it can be visualized is to choose a token variable. A token is a unit of text that the computer will use to run the analysis (Silge & Robinson, 2017). For the computer to create tokens, the user must specify which variable they want the tokens to come from - in other words, the user must specify which column of their data contains the text. The process of splitting the text up into tokens is called “tokenization” (Silge & Robinson, 2017). This project considers a single word to be the token of interest (with the exception of a few plots that use two word tokens called bigrams). This is the most important variable in the analysis because frequency, sentiment, and network plots are all created using tokens.

2.1.2: What are the main ideas in the text?

Once the data has been processed into a format that is easy to work with, we can begin to explore the data with the aid of visualizations. The plots in the interface start off very simple and increasingly become more complex as the user progresses. The plots that are described in the following descriptions can be found under the “Frequency Plots” tab.

The first two plots in the interface present the user with the most common words in the text. These plots offer a broad overview of what the text is about. The most common words can be indicative of what the text is about and perhaps who it focuses on. From my test runs of the interface, I have found that the names of the main characters almost always appear as very frequently used words when analyzing books.

The first is a classic frequency plot where the most common words appear at the top and go in descending order as you move down the plot. The second plot is a word cloud that also shows the most frequently used words. The larger a word appears in the cloud, the more common it is in the text. The word cloud may be more appealing for people who are more spatial and visual learners.

Both of these plots have a slider inputs that filter the minimum frequency of words that appear in the plot. The slider gives the user freedom to limit or broaden their scope as much as they would like. They may only want to focus on the 10 most common words or they may want to learn about some of the words that are used somewhat often in the book but maybe do not fall under the main topic. These sliders appear beneath many of the plots in the interface.

2.1.3: How do the words used in the document(s) convey emotion?

Sentiment analysis is often used to determine an author’s attitude towards a specific topic. It is conducted by using sentiment lexicons, which, for our purposes, are word banks that categorize words that carry an attached attitude/emotion. There are many different sentiment lexicons that can be used in text analysis but we will only use two in this interface. The first is called the AFINN lexicon and it ranks words on a score of -5 to 5 with -5 having the most negative

emotion/attitude and +5 having the most positive (Silge & Robinson, 2017). The second lexicon used is the Bing lexicon and it simply categorizes words as having positive or negative sentiment connotations (Silge & Robinson, 2017). AFINN and Bing are just two examples of the different ways sentiment lexicons can be scored/categorized. There is a lot of variety between lexicons in terms of scoring and subject matter. The lexicons used here are general, but there are specific lists for Economics and other fields.

We can work with the data in different ways when we use these lexicons because they score words in slightly different ways. For instance, when using the AFINN lexicon, we can multiply a specific word's frequency in the text by its score on the lexicon and this gives us a "word score". This gives us a sense of the strongest sentiment words in the text. For news articles this could give us a sense of the author's opinion on the subject matter; for books, this may indicate happy births or sad deaths that occurred in the text.

We can use sentiment analysis to trace the plot of a text document (usually a book). The user can group the data by chapter or a certain number of lines and calculate an overall sentiment score for that chunk of text. These can then be plotted in chronological order and the plot of the text can be traced by interpreting the sentiment scores.

While this interface considers text analysis through a bag of words lens in which we ignore grammar and the order of words, there are more nuanced relationships we can examine by looking at two words next to each other instead of one word by itself. The importance of looking at multiple words in sentiment analysis appears when you come across phrases such as "not happy". Taken alone, "happy" would be categorized as positive and would likely be given a high score. When you consider "not happy", it has a negative emotion. This project brings up this issue in a plot that shows the most strongly negated words that occur relatively frequently and carry strong emotional association. This is one instance where bigrams (two words) are used as the token instead of the default single word token in this project.

Sentiment analysis is a great tool for exploring the data. Humans are stimulated by strongly emotional writing and events, so it can be useful to get a sense the words and topics in your text that are associated with strong emotion.

2.1.4: How are the words in the text connected?

While it can be useful to look at which specific words are contributing the most to the text in terms of frequency and emotion, there are complex relationships between the words themselves that can provide insight into what is going on in a text.

One way to look at relationships between words is to measure how often they occur close to each other. This may mean appearing directly next to each other or it may simply be appearing within a few lines of each other. There are two ways that we have presented word co-occurrence in this project - tables and network graphs.

Another way to assess relationships is to look at the correlation between two words. This is a measure of how often they occur together as well as how often they do not occur together. It

captures more meaning than cooccurrence, which only considers how often words appear together. A strong correlation means that the two words occur together most of the time and do not appear on independent of each other very often.

This interface is geared towards introductory students, and for that reason, we hope to keep it simple to work through and truly understand. Co-occurrence and correlation are a bit more advanced than the rest of what is done within this project, but they are a good taste of what more advanced text analysis can begin to look at.

Section 2.2: Connection to Exploratory Data Analysis

John Tukey (1977), who originally coined the term “exploratory data analysis" details EDA in depth in his book titled *Exploratory Data Analysis*. His definition is not set in stone, but it includes being flexible when looking at the data and using graphical displays to explore distributions and relationships within the data. For our purposes, EDA means modifying data, displaying and interpreting outputs, summarizing results, and re-expressing data in a process that can be repeated as many times as are necessary.

Basic text analysis concepts can be used as a way to expose students to EDA. The structure of this project guides the user through looking deeply at their data to visualizing the main elements of it. Exploring the data is important for getting a sense of what facets of the data may prove most interesting or useful in future analyses. Data structure is crucial in determining which analyses are appropriate for any given data set.

Most statistical methods have a set of assumptions and guidelines that must be met in order to use them. These assumptions can include proper methods of data collection and sampling, correct variable distributions, and having equal spread throughout your variables. This is just the tip of the iceberg in the world of statistical assumptions. Here is where learning statistical reasoning becomes very important. To boil it down very simply, statistical reasoning is how we make sense of the data we have. This includes looking at the raw data, making modification to give it the desired structure, and visualizing it to examine relationships and distributions.

This project focuses on the simple exploratory aspects of text analysis. We are not running any complex analyses such as topic modeling or ranking algorithms. We are cleaning and processing data so that it can be visualized. Visualizations help us to make sense of what is happening in a document. This project should give students a good sense of their text data. They will be able to identify the main themes, the words with the strongest positive and negative attitudes, and a few simple relationships between words that appear.

This exploratory analysis could then potentially be used to inform future analyses that incorporate more advanced statistical methods. Most importantly, this interface engages the user in statistical reasoning by asking questions and creating exploratory graphs and does not allow them to jump directly to complex analyses without having sufficient knowledge of the data.

Chapter 3: Why use an Interactive Interface?

This chapter will introduce the benefits of using a dynamic and interactive interface as a replacement or supplement to more classical statistical learning methods. This project interface was created with an R Shiny app, so there will be a discussion of the benefits of Shiny for creating dynamic user interfaces. Finally, a demonstration of the interface is included.

Section 3.1: Benefits of creating a user interactive interface

We hope that creating an interactive interface will have three primary advantages over more classical worksheets and computer lab exercises: it excites people to learn about their own data, it hides the fine details of the process so that the user can focus on broad concepts and results, and it engages the user in more active learning.

3.1.1: How can an interface make people more excited to learn?

When a user is able to actively engage with a platform instead of reading a static PDF, it gives them the freedom to explore data that is interesting to them. In turn, this increases engagement with the interface and can encourage deeper understanding of the process and content. Higher levels of curiosity in the subject matter and contribute to more imaginative and attentive thoughts that the user is will entertain for longer than they would have otherwise (Wild & Pfannkuch, 1999). Many of the classic types of work assigned in introductory statistics classes are static -- worksheets and computer labs are designed to work with one or two specific data sets.

From my personal experience, I have found that working with prespecified data sets carries a risk of being unproductive in two different ways. Sometimes we are simply uninterested in the subject chosen and don't care to learn about it. This leads to going through the motions mechanically and produces very little active engagement with the activity. Secondly, and maybe more importantly, sometimes we do not understand what is in the data. It may be from a field that we have little knowledge in, such as protein folding in biology. It may also be that you grasp what the data is about but you do not understand how it is structured.

This is where the interactive interface may be better suited for introductory learners. As Wild and Pfannkuch mention, users have increased awareness of processes and results when they are interested in the topic. This can lead to more nuanced revelations that they may have missed had they not taken a specific interest in the data. We want users to be excited to see the results of analyzing their favorite book or reviews from their favorite movie. The excitement will translate into more engaged and effective learning.

3.1.2: How can an interface be less intimidating than other learning tools?

This interface outlines the process of elementary text analysis for the user. For example, the user is prompted from the very beginning to look at the structure of their data. They are asked to think about whether some lines should be removed and if there are words that they might not want to include in their analysis. Although the user is not writing the code to filter and subset the

data themselves, they are aware that there is some preprocessing that must occur before the data can be visualized and analyzed. This is just one instance of how an interface conveys broad concepts but removes the coding and intricate details that can be intimidating for some users.

Some anxiety about learning a new topic can be removed by “hiding” some of the finer details of what goes into producing the results. It is important that learners grasp a large, overarching idea before they begin to dissect the details. Users will be more likely to continue practicing and learning if they do not feel bogged down and nervous about the minute details. Bradstreet (1996) mentions the term “statistical anxiety” in introductory students as a fear that they won’t be able to do the math or understand what is happening. We want to reduce statistical anxiety and make users feel comfortable when engaging with this project. This attitude should be emphasized in any tool used for introductory statistics (Bradstreet, 1996). The user should be challenged to think about and explore the new concepts, but they should not be so pressured that they do not want to do it again.

One way to make the user feel comfortable is to break the process down into small, digestible steps. Our interface is broken into tabs, each of which has a specific topic and purpose. The content in each of the tabs builds on the foundations formed in the previous sections. Within each tab, each plot falls under the concept of the tab but also under the larger umbrella of introductory text analysis, particularly as it pertains to EDA. This structure of learning emphasizes each step as a part of a larger process but simultaneously hides the finer details (Wild & Pfannkuch, 1999).

Beyond breaking the process down into small steps, it is important to check that the user is getting the main takeaways of each section. At the top of each tab there is a short description of the main insights that can be learned from the decision being made and the tables and plots outputted. The user is asked questions that check for understanding at each major step of the project. These will engage the user and prompt them to reflect on what they should have learned from the results in comparison to what they actually learned.

The goal of this interface is to do most of the legwork for the user. It is presented in a clearly laid out manner, where each step leads naturally to the next. The user is guided through the decisions they will make that will affect the output. The module asks the user to do two tasks: make decisions that can be adjusted and reflect on how those decisions tie into the larger picture. This process can repeat as many times as the user would like. The code behind the interface does the majority of the work, but the user engages by making broad decisions. In summary, “systemize what you can, stimulate what you cannot” (Wild & Pfannkuch, 1999, p. 243). We systemize by choosing which plots to produce and where they fall into introductory text analysis. We stimulate by having the user make decisions and reflect on them.

3.1.3: How does an interface keep the user more engaged in active learning?

This project is designed so that the user engages with the interface at every step of the way. There are updates to the data or filtering and subsetting options for nearly every output. This

requires active decision making, which requires having thought about how to make an informed decision in the first place. The user first has to think about what they are being asked to do and then think about how they can successfully approach the problem.

We hope to provide the background and tools to give the user confidence in making decisions. In their paper on statistical thinking, Wild and Pfannkuch propose that one way to prompt active learning is to provide prescriptions instead of descriptions. Prescriptions describe procedures in enough detail that the user can understand and implement parts of the procedures themselves. Descriptions may identify and define the procedure but they lack the extra information necessary to instruct someone else how to carry the procedure out (Wild & Pfannkuch, 1999).

For example, after reading the introduction to the project and choosing what data they will use, the user is asked a series of questions about the structure of their data (how many files they have, if there are variables in the first line, etc). One of the decisions they have to make is whether or not to remove stop words. The first step of this decision is to figure out what stop words are. Then they might think about *why* they might want to keep or remove stop words. We have provided a prescription above this decision that should make answering both of these questions easy. We have provided the framework for the user to be informed about whether or not they would like to remove stop words.

Another example of active decision making in the interface is a step in the sentiment analysis section where the user is asked if they would like to filter any words from their data for the sentiment analysis portion. For instance, when working with *Peter Pan*, we saw the word “darling” come up as an extremely positive word. We thought about it and finally remembered that Darling is the family name in *Peter Pan*, and therefore, we didn’t want to include it in our sentiment analysis. The word “darling” did not have the positive connotation that it normally carries. We did the same with “lost” because of The Lost Boys. This is not something we would have realized had we not actively been examining the results and thinking about whether or not they make sense in the context of our data. To help the user think about these sorts of issues, we used this as an example of when they might want to remove their own words. By providing a real life example, we can give the user confidence for making similar decisions.

Active learning can be guided by starting broad and narrowing in as the interface proceeds. The interface starts with data upload and then continues with frequency plots, sentiment analysis, and more advanced plots showing relationships between words. Each of these sections, denoted by a tab on the interface, increases in complexity in terms of conceptual understanding and output interpretation.

Section 3.2: Peter Pan Examples

Note to Nick: I am having a tough time with these screenshots and the format of this section. I’m not sure which ones would be best to pick. Also, they get really grainy and the text is too small. The ones here are essentially just placeholders. Let's discuss this on Thursday.

This project was built using an R Shiny application. Shiny allows statisticians and data scientists to present their work in dynamic and interactive interfaces. It encourages user engagement and

leaves as little or as much room for flexibility as wanted. There are places where the user is required to engage with the interface in every tab -- that might include uploading data, choosing variables, removing words, and toggling slider inputs for graphs.

We will show a few screenshots of the interface to give a basic sense of the overall structure. We will include one graph (including a description) from each section but you can get a much better sense of the interface by using it yourself at: http://usresp-student.shinyapps.io/text_analysis.

3.2.1: Data Upload

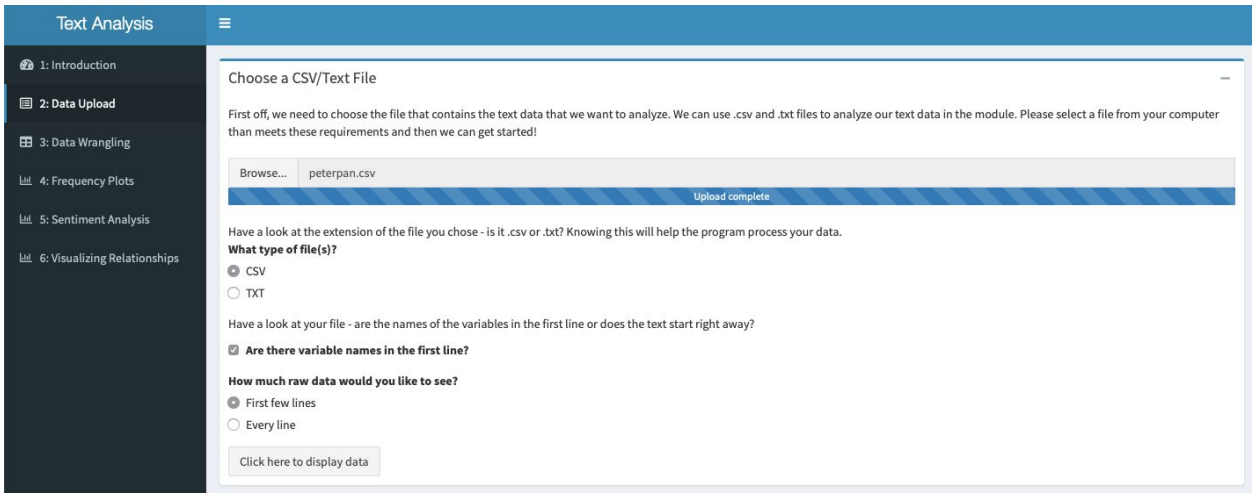

Figure 1: Dashboard structure and Data Upload

Figure 1 demonstrates the overall structure of our Shiny dashboard. The numbered tabs on the left hand side show the flow of the project -- as the user moves from 1 to 6, the content builds on each previous step. Additionally, you can get a sense of the questions asked during the initial data upload step.

3.2.2: Data Wrangling

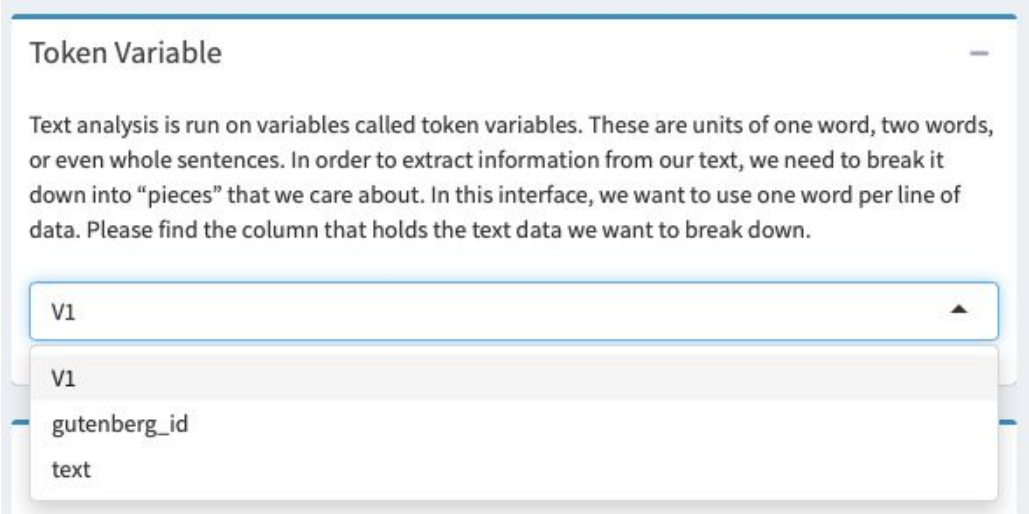

Figure 2: An example description of token variables and subsequent decision required by the user

Figure 2 shows a brief description of what a token variable is and how we might identify it in our own data. The user is asked to find the column of their data that holds the text that they want to analyze. The user might go back to the data tables in the previous tab to see which variable this might be. For *Peter Pan*, it happens to be “text”.

3.2.3: Frequency Plots

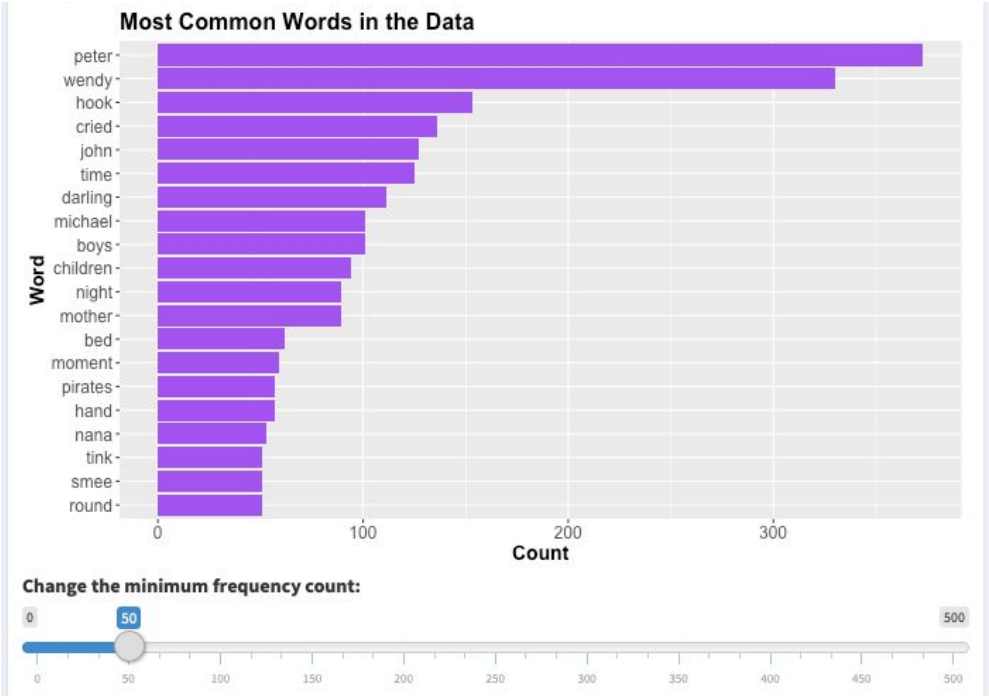

Figure 3: Most Common Words in Peter Pan

The figure above is an example of the simple frequency plot that the user sees directly after moving on from “Data Upload” and “Data Wrangling”. The user can toggle the slider input as much as they want and the graph will upload each time.

3.2.4: Sentiment Analysis

Figure 4: Sentiment Analysis by Chapter of Peter Pan

Figure 4 demonstrates one way that you can use sentiment analysis to trace attitudes throughout a text. It is showing sentiment score by each chapter by computing a summation of sentiment scores for every sentiment word in the chapter. Chapters 8 and 15 appear to be the most negative, which makes sense because both of these chapters detail large confrontations on the pirate ship (Barrie 2008).

3.2.5 Visualizing Relationships

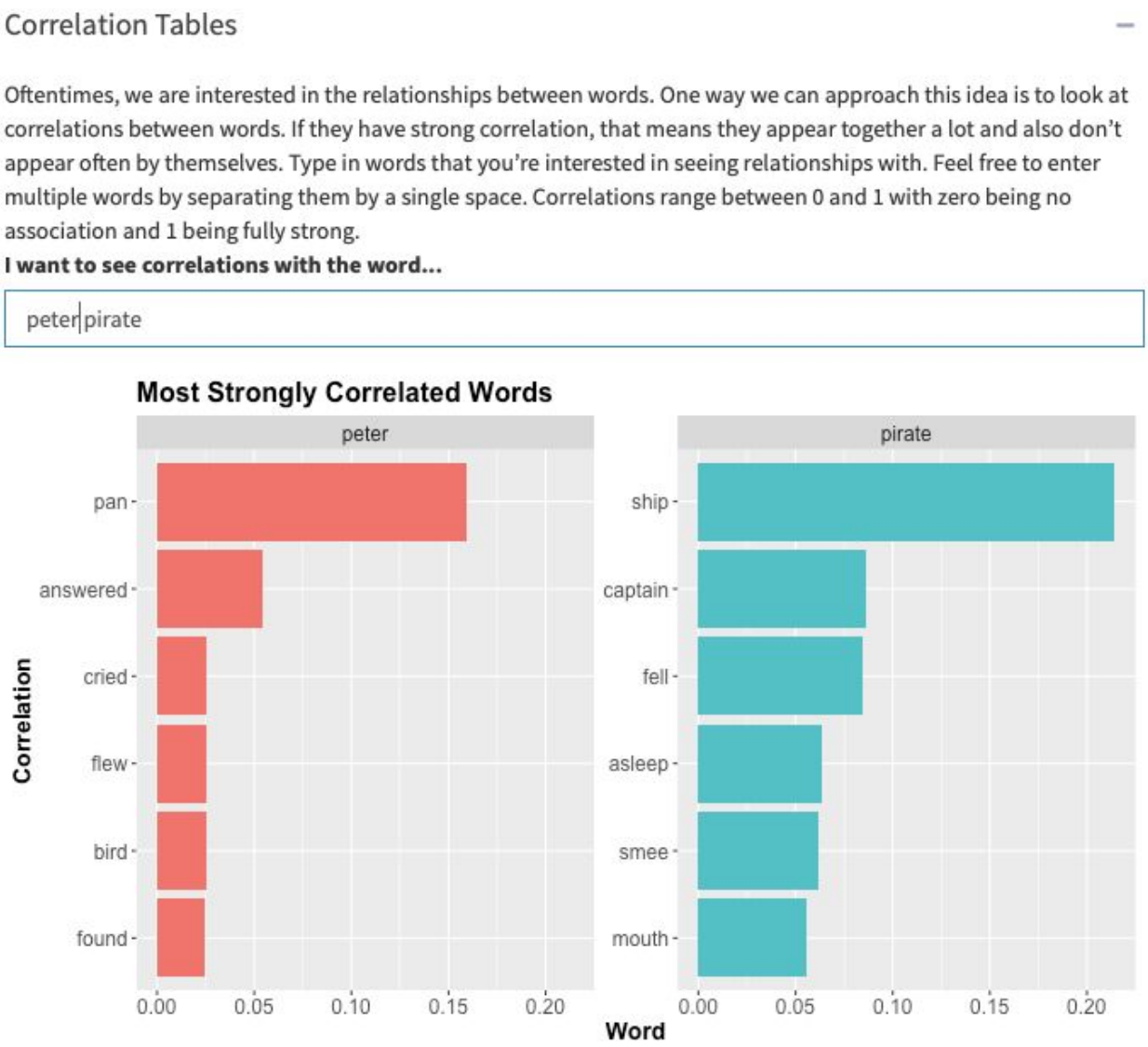

Figure 5: Words most correlated with "Peter" and "Pirate"

Figure 5 is good visualization for understanding what words might be highly correlated with a word of the user's choosing. For instance, I chose "peter" and "pirate". Peter seems to be correlated with action words, which makes sense because he is one of the main characters in the book and is involved in most of the action. Pirate seems to be capture words about Captain Hook and his right hand man, Smee.

Add a few sentences here to conclude.

Chapter 4: Importance to Statistical Education

Section 4.1: What are some advantages of this project?

4.1.1 Teaching users about Statistical Thinking

A lot of statistical thinking and considerations go into a statistical project before beginning any complex analysis. We must consider what information is available to us before we begin to work with it (Tukey, 1977). First off, we should be concerned with how data was collected and whether it came from a reliable source. Secondly, we should familiarize ourselves with the layout and distribution of the data.

A sampling of questions to consider in the statistical thinking stage:

- What does one observation represent?
- What are the variables included? What structure do they have?
- Based on my variable structures, what types of plots are useful for visualizing the data?
- What is the distribution of each variable?
- Is there any missing data? How is it stored? Will it cause problems in my analysis?
- What do the relationships between the variables look like?
- What are the variables of interest?
- Based on exploratory analysis, are there any variables of interest I didn't consider based on my previous knowledge?
- Are there any observations that don't seem to fit in with the rest of the data?
- Did any of the results of the EDA surprise me? Why?

This project allows the user to incorporate real-world applications that they are truly interested in by including a 'use your own data' section. When students are interested in the subject matter that is being used to teach a concept, they are more likely to actively develop their statistical thinking skills (Bradstreet, 1996). One goal of this interface is to get users to consider concepts and possible approaches before diving in to work on a problem.

We want to encourage reflection and thoughtfulness and for users to realize the importance of thinking before doing. Additionally, the user should recognize that their preconceptions about the data rarely include the full picture. They may have previous beliefs and attitudes about the data that should be checked and reflected on as they move through exploratory data analysis. We want users to consider how they are thinking and reflect on ways they could improve their thought process in order to engage more with the data (Wild & Pfannkuch 232). Reflections of this type will push the user to engage in statistical thinking and (hopefully) recognize that this is a crucial step in the data analysis process.

4.1.2 Make easy to use interface broadly accessible

Accessibility is a hot button issue in the modern statistical world. There are large swaths of people using data to produce research and software all over the world. Despite this, a lot of the products and research are not freely open for public use. Our project will be published online for

anyone to use for free. We hope to make it accessible to as many people as possible in part because it can be useful to such a broad audience.

In a different sense, a lot of current products are not accessible to introductory learners because they are written and produced above a level that an introductory student or user could understand. One way we are combatting this is by including in depth descriptions of what each step is doing. We are also asking reflective questions as well as questions that check for understanding. Our project assumes that the user has no statistical or analytical background, let alone any background in text analysis.

Every plot included in our interface is explained. We hope this helps the user to engage with and understand every step of the process instead of skipping over parts they may not understand right off the bat. All of the plots are well labeled and use colors and other visual cues to highlight differences or results that we want to emphasize.

Covering up the fine details of data wrangling that must be done in order to create these graphs is part of what makes this interface clear and easy to read for the introductory learner. This work is often tedious and unimportant to understanding the output, so we do not believe we are inhibiting any learning for our specific audience by leaving it out.

As a consequence of targeting this project at introductory learners, there is no novel or complicated text analysis presented in the module. This is simply an exploratory analysis of the the user's data. Therefore, this project does not develop any new methodology or gather any novel insights, but instead is a look at how interactive interfaces can be useful for introductory learners in relation to developing statistical education practices.

4.1.3 Dynamic outputs allow for more data exploration and fixing mistakes

One problem often encountered in introductory statistics coursework is the limited freedom to experiment and adjust repeatedly. There is usually a worksheet or lab that has a clear format and step by step process. There are not many opportunities to adjust the data or results and see how that changes the output.

While our project does follow a step by step structure, the user is always welcome to go back and adjust an earlier decision. Maybe they have decided they want to filter out some more lines of data or that they only want to see words that occur more than 200 times instead of just 100. There is no limit on the number of adjustments that can be made to the data. In fact, the user is encouraged to adjust the data multiple times. It can be very helpful to see how a slightly different setting can have a large impact on the results.

While it is fun for a user to keep adjusting inputs and watch how that affects outputs, there is real value in being able to correct things they may have missed upon first inspection. Going back to the previous example in which I discussed removing "darling" from the sentiment analysis, this ability to correct output proved useful. This interface makes updates and corrections as simple as doing it the first time. There is no frustration or tedious work that has to be done to make corrections.

Section 4.2: Application to PPDAC Framework

In the following sections, I will contextualize this project using the PPDAC (problem → plan → data → analysis → conclusion) statistical education framework as it has been proposed by MacKay and Oldford (1994) and adapted by Wild and Pfannkuch (1999).

4.2.1: Problem

A major challenge in creating learning tools for introductory students is considering aspects of the product design that will maximize engagement and learning. We want the end result to be clear and easy to understand. The user should be able to walk a clear path from start to finish, with each step having a definite purpose. Not only should the process be unambiguous, but it should be engaging and fun to use. This can be achieved by giving the user freedom to choose a topic of interest for completing the activity. A dynamic learning tool can achieve this far better than a worksheet. Worksheets and pre-designed labs are standard in many classrooms around the world. By making this project accessible to anyone with the link, we hope that this will open opportunities for more interactive learning for anyone who may want it.

EDA is a cornerstone of any data-oriented project. Text analysis is a great way to do explore data because it requires the user to think about the data structure and what components and themes of their data stand out. From anecdotal conversations with people who work with data professionally, many say that most of their time at work is spent wrangling and exploring data, with only a small percentage of time spent actually running analyses. We hope to show the user just how important EDA is to the world of data analysis.

4.2.2: Plan

The planning portion of this project includes considerations about which decisions to leave up to the user, which plots to include in the output, and how to get users to reflect on what they are seeing. Each of these factors plays a role in addressing the overarching goal of using the basics to text analysis as an exposure to EDA for any audience that takes an interest in digital humanities or statistics.

We want users to iteratively reflect on their decisions and how different perspectives of looking at the data and analysis may show new insights. Wild and Pfannkuch (1999) propose a process called “transnumeration” that focuses changing perspectives and considerations about how different viewpoints may offer different ways of understanding the data (p. 227). For example, after the user has decided which data to keep and which to remove, they are encouraged to look at their new data and reflect on how it turned out. If it was not satisfactory, they have the opportunity to revise it before moving on to the next step. This can be thought of as switching between thinking statistically - *how* do I remove the lines of interest? - and thinking contextually - *why* would I want to remove these lines? - in order to gain new insights (Wild & Pfannkuch, 1999). We hope that users are contemplating these types of questions because they are what deep and innovative learning stem from. In fact, “questions are more important than answers”, especially with regard to EDA (Wild & Pfannkuch, 1999, p. 233).

Spending time on exploration and reflection in the initial stages of working with data can save a substantial amount of time in the long run. This project is designed to encourage the user to think about what they want to do *before* they jump straight in. This begins with considering what is already known about the data and why it might be suitable for text analysis. Gathering context information is crucial for engaging in statistical thinking (Wild & Pfannkuch, 1999).

Beyond engaging the user in statistical thinking at the beginning, this interface should keep them engaged with the material at every step of the process. The design was planned so that a simple explanation of a concept, and possibly an example, appears before each subsection. Then the user is asked to apply what was described by making a decision or answering a concept check question. This decision could involve choosing which variables they might need to use for a certain visualization or considering which relationships they might want to explore given their context knowledge of the data.

Beyond reflecting on what the user wants before making a decision, we also want students to review the results of their decisions. This may include questions like, "Do the results I have make sense based on what I know about the data?", "Is there anything new or surprising that I didn't expect to see?", or "How could these results inform future work with this data?". These open-ended questions can further the user's understanding of their data and the processes in place to give results. A certain level of openness when considering the results can reveal information that may be unexpected or is antithetical to assumptions made about the data and results (Wild & Pfannkuch, 1999). Asking questions can provide answers about the data, some of which will confirm previous assumptions, and some of which may show differences in preconceptions and reality.

4.2.3: Data

All of the data used in this project must contain some element of text. We hope the text has some overarching theme - whether it be a mystery novel, magazine article, or final course paper - that unites the data.

The interface specifically accepts CSV and plain text files, which are easy to create if you don't already have them. Any text within a Word document or PDF can be copy-pasted into a plain text file for use in this interface. This makes getting acceptable data easy and is one way to ensure that the interface is accessible to anyone who may want to use it. One example of an amazing source of text data is Project Gutenberg, which has thousands of classic books available for download [5].

After picking and uploading the data, most of the processing happens behind the scenes. One objective of the interface is to hide the code from the user because it can be intimidating and isn't necessary for understanding basic text analysis. We want students to enjoy the broad concept without having to learn the fine technical details required to run the analysis.

Although the user doesn't see the specific data preprocessing code, they are prompted to think about ways they might amend the data prior to creating visualizations. There are options to remove full lines of data (sections deemed unnecessary such as copyrights or

acknowledgements) as well as specific words that they do not want to appear in the analysis. By the end of the “Data Upload” step, the user should have a good idea of what data is appropriate to use and how they might want to modify it for their own use.

4.2.4: Analysis

The first step in analysis is choosing data and, if the user elects to use their own, uploading it properly. The first step of the interface that requires user input is the “Choose a CSV/Text File” box within the “Data Upload” tab. Here is where the analysis for the user begins - this box has inputs that require looking at the data file and deciding which options to choose for reading in and doing basic processing on the data.

Guiding questions are in place to assist the user in the decision making process. This starts with the data upload and continues throughout. For example, when the user has to choose their token variable, they are asked questions that might help them figure out which variable they will choose. More generally, there are questions -- for example, “What words do you expect to appear the most based on what you know about the data?” -- that attempt to engage the user to think broadly about their data.

Each of the plots displayed was picked for a specific purpose. Some, such as the frequency plots, are useful summaries of the most prominent features of the text. Others, such as the AFINN and Bing sentiment plots, demonstrate how results differ based on how you process the data. Wild and Pfannkuch (1999) mention the use of “trigger questions” to serve as aids in structuring the process of statistical thinking that we want the user to engage in. In this case, we want the user to think about *why* each plot was included and *how* the information each plot displays relates to the broader picture of text analysis.

Once the user has moved on to looking at the visualizations, encourage them to interpret the results. Here are some prompts to consider:

- What are the title and axis labels? What is the scale of each axis?
- Were there user inputs for this plot? How did your choice of input affect what is shown?
- Are each of the plots unique? If they appear to show the same information in different formats, can you identify why including each plot might be beneficial?
- Provide a one sentence summary of the output displayed in this plot.
- What do these results mean in the context of your data?

They should be able to deduce the main idea of each plot and be able to adjust the data to look at different results. This encourages exploring the data – making changes based on what they saw in the first round of plots, evaluating the updated results, and repeating the process. While it can be useful for the user to think about what stood out to them in the visualizations, it is just as important for them to think about what *did not* show up. They may want to think about how and why this happened and whether there are changes to make to address the discrepancy between what they expected and what they observed.

A five step “interrogative cycle” detailed by Wild and Pfannkuch (1999) summarizes the above analysis process quite well. The user first *generates* ideas, *seeks* information from previous knowledge or outside sources, *interprets* and subsequently *criticizes* results, and *judges* the next decision to be made. The analysis detailed above covers each of these steps for the user. Prompting the user with questions is an attempt to engage the user in each of the five steps of analysis. These questions encourage idea generation, reflection, revision, and repetition.

4.2.5: Conclusion

This interface is designed to display the main ideas of a text through data wrangling and exploratory analysis. This is largely aided by placing descriptions and probing questions in places where the user encounters a new concept or plot that we want to emphasize. We want these learning aids to be detailed enough for the user to understand the plots and what information can be learned from basic text analysis.

One of the previously stated goals of the interface is to show how basic text analysis serves as a method of EDA. One way to measure this is to ask the user to summarize three things they learned about their data by using the interface.

Ideally, the user’s original thoughts and questions about their data have been answered. If they were not, the user is prompted to think about what could be done to answer them. They might brainstorm different ways of wrangling the data or plots that provide new insights.

The user should reflect on the results of their text analysis and compare them to any expectations they may have had when beginning the process. They may have thought of new questions or ideas along the way. They may want to see more advanced plot and analyses completed. A good final exercise for the user would be to generate their own trigger questions and making an attempt at outlining an approach to answering each of them.

Chapter 5: Conclusion

5.1 Overview

Our project sought to create an interactive user interface that is useful to a larger audience than our typical statistics students. Text analysis can serve as a link between statisticians and people who work with text “data” (though an author may not think of their book as data, it is). This interface is a brief glimpse of how text data can be processed and summarized through interactive visual displays that encourage active learning.

We want it to be accessible to everyone who might want to use it. For us, accessibility means making the interface publically available, using data types that are widely available, and keeping concepts simple and well-explained.

Basic text analysis is a wonderful tool that can be used to illustrate exploratory data analysis. This project summarizes the main facets of the data - simple word frequencies, attitudes in the text, and simple relationships between words - using data visualizations. We ask the user to do a lot of reflection on each step - why it is important, what information it tells us about the data, how they might change it - so that they can get a sense of just how much this small number of plots can tell us about the data.

Finally, this paper details the our project’s application to the statistical education theory. Users are guided through the process of PPDAC to analyze their data. Wild and Pfannkuch (1999) mention that “the elements of this model should be self explanatory to statisticians” (p. 225). As a statistics student, I am aware of the statistical thinking processes that must happen before a data analysis project can be in full swing. What we hope is that the project provides that same knowledge for all of its users, no matter what background they come from. Projects like this one can demonstrate how engage in critical thinking about decision-making and the results that follow from your decisions.

5.2 Limitations

The audience for this interface is introductory students in the humanities as well as statistics courses. It is assumed that the user does not have any background in text analysis or, more generally, any statistical analysis.

One limitation with this design is that more advanced users might want to see the code behind the interface. A possible revision for future work is to add an option to see the code that corresponds with each plot and input. Many of the examples in the Shiny Gallery and in resources around the internet include a sidebar that has the code to produce the app. Creating an option to see the code would increase the reproducibility of this project.

Additionally, a more advanced user might want to see more advanced text analysis plots and methods used. These methods might include topic modeling and clustering (Silge & Robinson, 2017). Advanced users might find themselves bored with the methods and visualizations used in the interface.

5.3 Future Work

There is a lot of room for more development on the interface. First and foremost, we would like to add more capabilities that gather insights from multiple files. There are more complex analyses such as calculating TF-IDF score and creating topic models that can give users a taste of some of the cooler things you can do with text analysis. These plots could be stored in an “Advanced” tab so that the introductory users recognize that these are moving away from the “basic” topics we covered to more in depth text analysis methods.

The applications to statistical education theory can be expanded and improved. At this point, the interface does not explicitly walk through each step of PPDAC. In the future, we might want to introduce the PPDAC framework at the beginning of the interface and show how it meets the criteria throughout the process.

References

Barrie, J. M. (2008). *Peter Pan*. New York, NY: Modern Pub.

Bradstreet, T. E. (1996). Teaching Introductory Statistics Courses so That Nonstatisticians Experience Statistical Reasoning. *The American Statistician, 50*(1), 69-78.

How Kids Learn – the statistical enquiry cycle. (2017, September 20). Retrieved from http://new.censusatschool.org.nz/resource/how-kids-learn-the-statistical-enquiry-cycle/

MacKay, R. J., & Oldford, R. W. (2000). Scientific Method, Statistical Method and the Speed of Light. *Statistical Science, 15*(3), 254-278.

MacKay, R.J. & Oldford, W. (1994). *Stat 231 Course Notes Full 1994.* Waterloo: University of Waterloo.

Project Gutenberg. (n.d.). Retrieved January 21, 2018, from www.gutenberg.org.

Gallery (2017) Retrieved September 8, 2018, from https://shiny.rstudio.com/gallery/

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. Beijing: OReilly.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Pearson.

Wild, C.J., and Pfannkuch,M (1999) Statistical Thinking in Empirical Enquiry, *International Statistical Review/Revue Internationale de Statistique*, Vol. 67, No. 3, pp 223-248.