Text Analysis: Exploratory Data Analysis for Introductory and Nonstatistical Students

---

2018-12-12

Abstract: Text analysis is the process of extracting information from text data. It has applications in fields ranging from statistics and computer science to marketing, linguistics, and digital humanities. In this paper, we describe an introductory text analysis interface that we created and how it benefits 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 (Problem, Plab, Data, Analysis, Conclusion) statistical education framework, specifically as it relates to exploratory data analysis for an introductory audience.

Note for Draft: There are a few places in the paper denoted in RED that simply aren't true about the interface 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 (and I'll have time to fix them on my way home from school before the submission date).

Section 1: Introduction

Text analysis is the process of extracting information from a body of text. Text could be from novels, movie reviews, or any other body of text imaginable. For instance, an English literature student might explore questions related to word use in Tolstoy's *War and Peace* or narrative timeline in Jane Austen's *Sense and Sensibility*.

This project describes a user interactive text analysis interface implemented as a web application using Shiny (R Studio, 2017). The interface guides the student through the process of uploading and modifying data, producing plots that show the most frequently used words, sentiments throughout the text, and relationships between the words. The student 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 building on the previous. There are descriptions and questions along the way that prompt the student to think about the actions they are making and what effect those have on the results.

Statistical educators are constantly 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. The process of solving a problem begins with statistical thinking. There are many questions to ask of a project before choosing and running statistical models and algorithms. The student should ask 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 a preliminary look at the data before building models. Exploratory Data Analysis (EDA) is crucial for examining relationships in the data and for breaking preconceptions the student has about the data. Through exploration, the student often finds relationships and values in the data that they were not expecting. These revelations may prompt them to go back and modify the data until it has the structure and attributes that they need.

We use the PPDAC (problem, plan, data, analysis, conclusion) framework to introduce ways to approach a statistical problem (MacKay & Oldford, 1994; Wild & Pfannkuch, 1999). This framework emphasizes the statistical reasoning and exploratory steps of data analysis. Only one of five steps - analysis - focuses on doing statistical analyses. Each of the other steps centers on thinking about the problem at hand, appropriate approaches to address it, and data exploration. Text analysis provides an opportunity to do all of these.

This interface demonstrates the building blocks for understanding the main ideas and attitudes in the text. Each of the plots presented in the interface tells us something unique about the text data. We want students to be able to extract meaning from text data without having to learn how to code or understand the details behind the results presented in the interface. One common problem in introductory statistics courses is that the provided examples may feel mundane to the student (Gould, 2010). One of the primary goals of statistics is to learn through addressing real world problems (Bradstreet, 1996). This project allows the student to choose their own data source, so they can always choose a topic that is of interest to them. This encourages more

interaction with the project and should produce deeper levels of learning. We designed it to be fun to play with.

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 and social sciences, 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 others, such as professionals in the marketing field who are interested in the customer's sentiments when making business decisions.

The paper proceeds as follows:

- **Section 2**: A basic introduction to common terms and data used in text analysis. Additionally, we introduce the connection between text analysis and EDA.
- **Section 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.*
- **Section 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.
- **Section 5**: A final discussion of the paper, limitations of the project, and a look at possibilities for future work.

Section 2: Text Analysis Methods

2.1: Text Analysis Basics

Text analysis is the process of extracting information from a body of text (Silge & Robinson, 2017). This includes finding the most frequently used words, locating spots in the text where sad events occur or finding strongly related words by looking at how often they occur together.

Before we explore this project within the larger realm of statistical education, we 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 be used and how is it processed?

We begin by choosing and initially modifying the data. In the interface, these steps occur in the "Data Upload" and "Data Wrangling" tabs are in the left hand sidebar (see Figure 1).

Text analysis can be done on anything from Beatles song lyrics to PhD dissertations (Silge & Robinson, 2017). We designed the interface to accept plain text and CSV files.

The first step in working with any data set is to look at the raw data. For *Peter Pan,* the first few lines include information about the author, title, and copyright information (see Figure 2). An analyst should gather some information about the variables (if there are any) and get a sense of what is contained in the dataset. After thinking that over, unnecessary data can be filtered out. For instance, as is demonstrated in the example later on, we 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 student 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 can go back a step and try removing different lines.

Text data is full of common words such as “like”, “because”, and “the”. These extremely common words do not give us much information about the topics in the text, so we often remove them before conducting any text analysis. We call these “stop words”. The interface identifies 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 interface is designed to remove stop words by default, but there is an option to keep them in if the student would like to see how they affect the results. In fact, we encourage the student 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 student is also given the option to add to the stop word list. This pushes them to think about what words are most or least relevant to their analyses. The student is free to return to this step if they see or remember words that they want to remove later.

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 uses to run the analysis (Silge & Robinson, 2017). To create tokens, the student must specify which variable they want the tokens to come from (which column of their data contains the words to be analyzed). The process of splitting the text into tokens is called “tokenization” (Silge & Robinson, 2017). This project considers a single word as the token of interest (with the exception of a few plots that use two word tokens called bigrams). For example, the first sentence of Chapter 1 of *Peter Pan* reads “All children, except one, grow up” (Barrie, 1904). This sentence is tokenized into the words “all”, “children”, “except”, “one”, “grow”, and “up” as individual units of text (tokens). This is the most important variable 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 workable format we can begin to explore the data with simple visualizations. The interface starts off very simply and becomes more complex as the student progresses. The plots that are described here can be found under the “Frequency Plots” tab (see Figure 3 for an example).

The first two plots in the interface present the student with the most common words in the text. These plots offer a broad overview of what the text is about. Not surprisingly, the names of the main characters almost always appear as frequently used words when analyzing books. For *Peter Pan* this includes Peter, Wendy, and Captain Hook.

The first visualization is a frequency plot where the most common words appear at the top and go in descending order as you move down the plot. The second 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 to spatial and visual learners.

Both of these plots have slider inputs that filter the minimum frequency of words that appear in the plot: the user can limit or broaden their scope with these. They may only want to focus on the 10 most common words or identify what words that are used less often in the book but do not fall under the main topic. Sliders appear beneath many of the interface plots as a way to facilitate exploration.

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

Humans are stimulated by 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.

Sentiment analysis is often used to determine an author's attitude towards a specific topic (Silge & Robinson, 2017). 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 sentiment lexicons that can be used in text analysis but two are implemented 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 sentiments can be scored/categorized. The lexicons used here are general, but there are specific lists for fields (e.g. Economics and Medicine).

Different lexicons score words in 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 produces 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 news event; for books, this may indicate happy or sad events occurring in the text. For instance, we could see which emotion words are used to express Peter's sadness at Wendy and her family having to leave Neverland to go back to the real world.

We can use sentiment analysis to trace the plot of a text document (usually a book). The student can group the data (e.g. chapter or a certain number of lines) and calculate an overall sentiment score for that chunk of text. These scores can then be plotted in chronological order and the ebb and flow 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, more nuanced relationships can be examined by important linguistic factors such as 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. We see 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) used as the token instead of the default single word token in this project can provide different insights.

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. Words may appear directly next to each other or they may appear within a few lines of each other. We present tables and network graphs as two ways to explore word co-occurrence.

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 co-occurrence, 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.

2.2: Connection to Exploratory Data Analysis and the statistical analysis process

John Tukey, who originally coined the term “exploratory data analysis" details EDA in his book *Exploratory Data Analysis* (1977). His definition suggests being open-minded when looking at data and using graphical displays to explore distributions and relationships within the data. For our project, EDA means modifying text data, displaying and interpreting outputs, summarizing results, and re-expressing data in a process that can be repeated as often as necessary. Nolan and Perrett (2016) mention that learning to make and analyze graphs is as important to quantitative reasoning as reading and writing are for effective communication.

EDA is crucial to addressing each step of the PPDAC framework briefly mentioned in the introduction section of this paper. The process begins with identifying the “problem” and and determining how it can be answered. The “plan” step identifies what tools we need to answer the question - which types variables do we need? What plots can be used to demonstrate EDA? - and how to design the layout and process. The “data” step is where the student determines a suitable text data source and begins the data cleaning process. This is followed by the

“analysis” step where the student continues to arrange their data, look for patterns and surprising information in exploratory plots, and adjust plots to see different perspectives. Finally, the interface ends with a “conclusion” step. Here is where the student reflects on what they saw as well as what did not appear in the plots. They should summarize what they learned and think about future work with this data or things they would have liked to see but were not included in the design.

Basic text analysis concepts can be used to expose students to EDA. This project takes the student from looking deeply at their data to visualizing its main elements. Exploring the data is important for getting a sense of what may prove most interesting or useful in future analyses. Structure is crucial in determining which analyses are appropriate for any given data set (Nolan & Perrett, 2016).

This project focuses on simple exploratory aspects of text analysis. We are not running any complex analyses such as topic modeling or ranking algorithms (Silge & Robinson, 2017). We are cleaning and processing data so that it can be visualized and explored. Visualizations help us to make sense of what is happening in a document by engaging in statistical reasoning (Nolan & Perrett, 2016; GAISE, 2016). This project should give students a good sense of their text data. They are 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 inform future analyses that incorporate more advanced and sophisticated statistical methods. Most importantly, this interface engages the student in statistical reasoning by asking questions and creating exploratory graphs and does not force them to jump directly into complex analyses without having sufficient knowledge of the data.

Section 3: Why use an interactive Interface?

This section introduces the benefits of using a dynamic and interactive interface as a replacement or supplement to more classical statistical learning methods. This interface was created with an R Shiny dynamic web application (R Studio, 2017). We discuss the benefits of Shiny for creating dynamic user interfaces and provide an overview of the interface.

3.1: Benefits of creating a user interactive interface

Creating an interactive interface has three potential advantages over more classical worksheets and computer lab exercises: 1) it excites students to learn about their own data, 2) it hides the fine details of the process so that the student can focus on broad concepts and results, and it 3) engages the student in more active learning.

When students are able to actively engage with a platform instead of viewing a static file, they have freedom to explore data that is interesting to them. In turn, this motivates deeper engagement and hopefully more understanding of the process, insights higher levels of curiosity in the subject matter, and contributes to more imaginative and attentive thoughts for longer periods of engagement (Wild & Pfannkuch, 1999; Nolan & Perrett, 2016, GAISE, 2016). Many

traditional assignments in introductory statistics classes are static: worksheets and computer labs are designed to work with a specific data set.

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 (Gould, 2010). This leads to going through the motions mechanically with 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 we grasp what the data is about but do not understand how it is structured.

This is where the interactive interface may be better suited for introductory students. We want students to be excited to see the results of analyzing their favorite book or reviews from their favorite movie. The excitement may translate into more active and effective learning (GAISE, 2016).

This interactive interface outlines the process of elementary text analysis for the student. For example, the student 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 student is not writing the code to filter and subset the data themselves, they are aware that *preprocessing* 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 students.

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 students grasp a large, overarching idea before they dissect the details. Students are more likely to continue practicing and learning if they do not feel bogged down and nervous about the minute details. Bradstreet (1996) describes "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 students feel comfortable when engaging with this project. This attitude should be emphasized in any tool used for introductory statistics (Bradstreet, 1996). The student 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 student feel comfortable is to break the process down into small, digestible steps. Our interface is broken into tabs, which each have a specific topic and purpose. The content in each tab builds on foundations formed in the previous sections. Within each tab, each plot falls under the concept of that tab but also under a larger umbrella of introductory text analysis, particularly as it pertains to EDA. This teaching style 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 student 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 student is asked questions that check for understanding at each major step of the project. These engage the student 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 student. It is clearly laid out and each step leads naturally to the next. The interface asks the student to do two tasks: 1) make decisions and 2) reflect on how those decisions tie into the larger picture. This process can repeat as many times as the student chooses. The code behind the interface does the majority of the work, but the student makes 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 how they apply to introductory text analysis. We stimulate by having the student make decisions and reflect on them.

This project is designed so that the student engages with the interface at every step of the way. There are updates to the data or filtering and subsetting options provided for nearly every output. This requires active decision making but scaffolds the process to avoid cognitive overload. The student first has to think about what they are being asked to do and then think about how they can successfully approach the problem.

We provide the background and tools to give the student confidence in making decisions. 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 student 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 student 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 is to figure out what stop words are. Then they might think about *why* they might want to keep or remove stop words. We provide a *prescription* above this decision that should make answering both of these questions easy. Another example of active decision making in the interface is a step in the sentiment analysis section where the student 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 a positive word. We 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, a band of characters who appear in multiple works by J.M. Barrie. 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 student 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 student confidence to make similar decisions.

Active learning can be guided by starting broad and gradually narrowing in. The interface starts with data upload and 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 of conceptual understanding and output interpretation.

3.2: Case Study: Peter Pan

We provide a few screenshots of the interface to give a basic sense of the overall structure. We include one graph (including a description) from each section. The interface is available at: http://usresp-student.shinyapps.io/text_analysis.

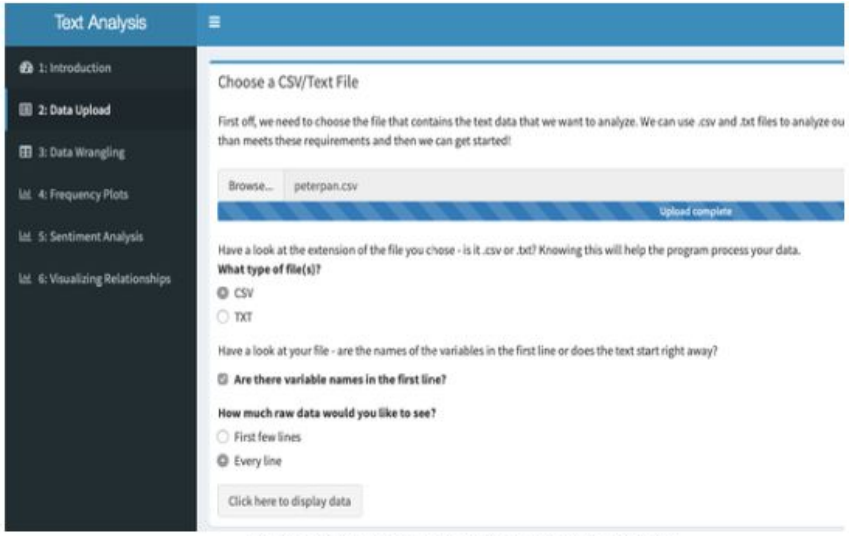

Figure 1: Interface structure and data upload

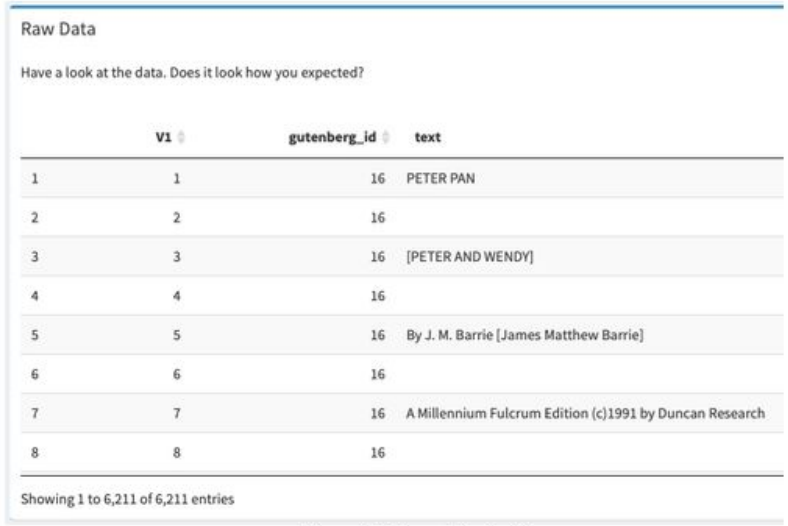

Figure 2: Raw data table

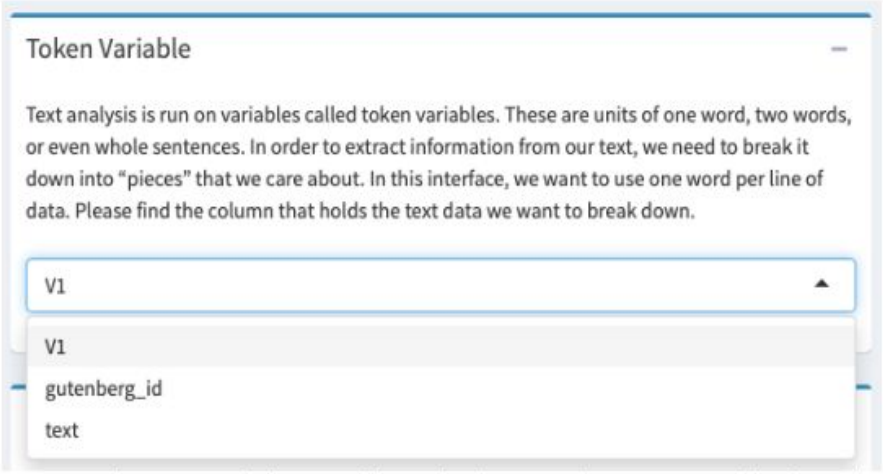

Figure 3: A description of token variables and subsequent decision required by the student

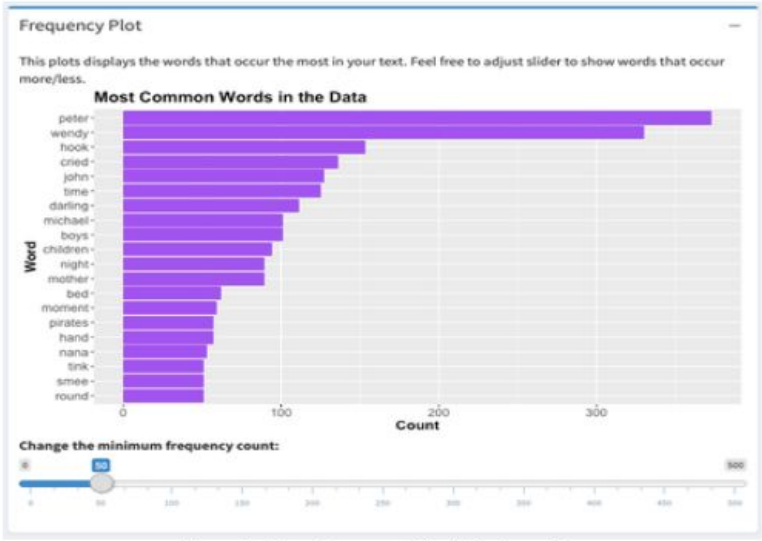

Figure 4: Most Common Words in Peter Pan

Figure 1 displays our Shiny interface. The numbered tabs on the left hand side show the flow of the project - as the student moves from 1 to 6, we build on each previous step. To begin, the student can get a sense of the questions asked during the initial data upload step. Once the student clicks the display data button, they are presented with the raw data presented in Figure 2.

Figure 3 shows a brief description of what a token variable is and how we might identify it in our own data. The student is asked to find the column of their data that holds the text that they want to analyze. The student might go back to the data tables in the previous tab to see which variable this might be. For *Peter Pan*, it is “text”.

Figure 4 is an example of the simple frequency plot that the student sees directly after “Data Upload” and “Data Wrangling”. The student can toggle the slider input the graph will upload each time.

Figure 5: Sentiment analysis by chapter

Figure 5 demonstrates one way that you can use sentiment analysis to show a narrative arc. It shows sentiment score by each chapter by showing scores for every sentiment word in the chapter. Chapters 8 and 15 appear to be the most negative, which makes sense because both chapters detail harrowing confrontations on the pirate ship (Barrie 2008).

Figure 6 (which follows below) is helpful for students to understand what words might be highly correlated with a word of the student’s choosing. For instance, we 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. Peter is also highly correlated

with “Pan”: this is an instance where we might go back and remove it from the analysis. Pirate seems to be capture words about Captain Hook and his right hand man, Smee.

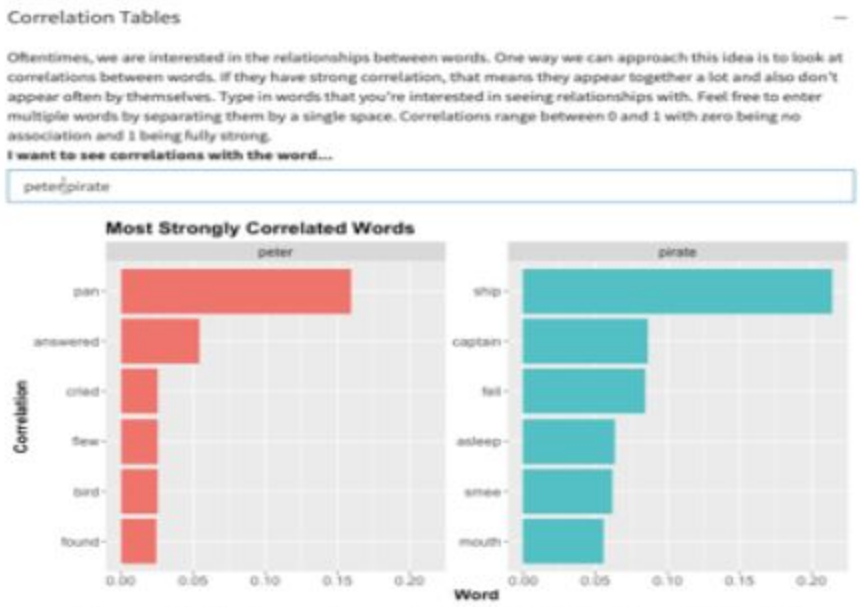

Figure 6: Words most correlated with “Peter” and “Pirate”

Each plot demonstrates a specific concept to the student without requiring computation. Statistics such as correlation can be hard for an introductory student to compute. Describing correlation and how it might provide information about specific words gives the student an understanding of the term without having to ask them to compute it. This applies to many of the concepts displayed throughout the interface. One goal of the interface is to introduce statistical concepts and EDA to students in an easily understandable way, such as the correlation table in Figure 6.

Section 4: Teaching Text Analytics as part of broader statistics

4.1: What are the advantages of this project?

4.1.1 Teaching students 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, we should be concerned with how data was collected and whether it came from a reliable source. Second, we should familiarize ourselves with the layout and distribution of the data. Students might want to engage in “one minute revelations” where they consider the purpose and important facets of each graph (Nolan & Perrett, 2016).

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 student to incorporate data 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 students to consider concepts and possible approaches before diving in to work on a problem.

We want to encourage reflection and thoughtfulness and for students to realize the importance of *thinking* before *doing*. Additionally, the student 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 EDA. We want students 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 199; GAISE, 2016). Reflections of this type may push the student to engage in statistical thinking and help them to recognize this crucial step in the data analysis process.

4.1.2 Make easy to use interface broadly accessible

Accessibility of data and software is a common topic of discussion in that statistical world. Many products and data sources are not freely open for public use. Our project is published online under a creative commons open source license for anyone to use for free. We make it accessible to facilitate as many students and instructors to use it as possible.

In a different sense of accessibility, many products are not accessible to introductory students because they are written and produced above a level that an introductory student or student can understand. We describe in-depth of what each step is doing, ask reflective questions and questions that check for understanding. We assume that the student has no statistical or analytical background, let alone any background in text analysis.

Every plot included in our interface is explained. This helps the student engage with and understand every step of the process instead of skipping over parts they may not understand. 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 and coding 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 student. 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 we target this project to introductory students, there is no novel or complicated text analysis presented in the interface. This is simply an exploratory analysis of the the student's data. Therefore, this project does not develop any new methodology or gather any novel insights, but instead illustrates how interactive interfaces can be useful for introductory students in relation to developing statistical education practices (GAISE, 2016).

Further research is needed to assess how effective this project is for use by novices. This might include usability tests and feedback through focus groups, video taping users, and interviews. We can use the results from these tests to further refine the text analysis interface.

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 student is always able to reconsider and revise 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 student 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. For instance, with *Peter Pan*, we could decide to remove "darling", "lost", and "pan" retrospectively.

While it is fun for a student to keep adjusting inputs and watch how that affects outputs, there is value in being able to correct things they may have missed upon first inspection. This interface makes updates and corrections as simple as doing it the first time.

4.2: Application to Statistics Education

In the following sections, we further 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). This structure loosely follows an example demonstrated on the New Zealand Census at School website (http://new.censusatschool.org.nz/resource/how-kids-learn-the-statistical-enquiry-cycle/**).**

4.2.1: Problem

EDA is a cornerstone of any data-oriented project. Text analysis is a great way to explore data because it requires the student 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. The goal is to show the student 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 student, which plots to include in the output, and how to get students 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 students 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 student 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 students are contemplating these types of questions because they are what deep and innovative learning stems 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 student 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 student in statistical thinking at the beginning, this interface should keep them engaged with the material at every step of the process. Each subsection is designed to include a simple explanation of a concept and, in some cases, an example. Then the student 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 student 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 student’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 which may have some overarching theme - whether it be a mystery novel, a set of magazine articles, or a set of poems - that unites the data.

The interface accepts CSV and plain text files. Any text within a Word document or PDF can be copy-pasted into a plain text file for use in this interface. This makes data easy and is one way to facilitate access. One rich source of text data is Project Gutenberg (https://www.gutenberg.org), which has thousands of classic books available for download (no longer under copyright restriction).

After identifying and uploading the data, most of the processing happens behind the scenes. One objective of the interface is to hide the code from the student 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 student doesn't see the 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 that appear in Project Gutenberg files) as well as specific words that they do not want to appear in the analysis. By the end of the "Data Upload" step, the student 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 student elects to use their own, uploading it properly. The first step of the interface that requires student input is the "Choose a CSV/Text File" box within the "Data Upload" tab. Here is where the analysis for the student 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 student in the decision making process. For example, when the student 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 student to think broadly about their data.

Each of the plots displayed was chosen 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 student to engage in. In this case, we want the student 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 student has moved on to looking at the visualizations, they should next be asked 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 student 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 (GAISE, 2016). This encourages an investigative nature of 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 student 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 (GAISE, 2016).

A five step “interrogative cycle” detailed by Wild and Pfannkuch (1999) summarizes the above analysis process quite well. The student 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 student. Prompting the student with questions is an attempt to engage the student 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 introduce the main ideas of a text analysis through data wrangling and exploratory analysis and visualization. This is largely aided by placing descriptions and probing questions in places where the student encounters a new concept or plot that we want to emphasize. We want these learning aids to be detailed enough for the student 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 assess this is to ask the student to summarize three things they learned about their data by using the interface.

Ideally, the student's original questions about their data have been answered. If they were not, the student 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 student 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 student would be to generate their own trigger questions and make an attempt at outlining an approach to answering each of them.

Section 5: Conclusion

5.1 Overview

Our project sought to create an interactive user interface that is useful to a broad audience (e.g. K-12 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 the web application to be accessible to everyone who might want to use it. For us, accessibility means 1) making the interface publically available, 2) using data types that are widely available, 3) keeping concepts simple and clear, and 4) minimizing cognitive load.

5.2 Limitations

Our intended audience for this interface is introductory students in the humanities and social sciences (high school or postsecondary) as well as statistics courses. It is assumed that the student does not have any background in text analysis or, more generally, any statistical analysis.

One limitation with this design is that more advanced students 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 (R Studio, 2017). Creating an option to see the code would increase the reproducibility of this project.

Additionally, a more advanced student might want to see more advanced text analysis plots and methods used. These methods might include topic modeling and clustering (Silge & Robinson, 2017). Advanced students 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 students 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 students 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.

We developed this project as an existence proof of an introductory text analysis tool that demonstrates EDA within statistical education framework (namely PPDAC). We made many speculations about accessibility and how effective a tool like this can be for an introductory audience. We hope to improve this project in future iterations by videotaping students using the interface, running focus groups, and interviewing people who tested the interface out. Collecting feedback from the targeted audience is the next step in assuring that this interface achieves the goal of accessibility and a clear introduction to EDA via text analysis for introductory audiences.

Basic text analysis can be a tool to illustrate EDA. This project summarizes the main facets of the data - simple word frequencies, attitudes in the text, and relationships between words - using data visualizations. We ask the student 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 extract insights about the data.

Finally, this paper details our project’s application to the statistical education theory. Students 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 statisticians, we are aware of the statistical thinking processes that must happen before a data analysis project can begin. We hope that the project provides new knowledge to all of its students, no matter their background.

References

Barrie, J. M. (1904). *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.

Census at School. (2017). Retrieved from http://new.censusatschool.org.nz/resource/how-kids-learn-the-statistical-enquiry-cycle/

GAISE College Report ASA Revision Committee, "Guidelines for Assessment and Instruction in Statistics Education College Report 2016," http://www.amstat.org/education/gaise

Gould, R. (2010). Statistics and the Modern Student. *International Statistical Review, 78*(2), 297-315.

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

Nolan, D., & Perrett, J. (2016). Teaching and Learning Data Visualization: Ideas and Assignments. *The American Statistician, 70*(3), 260-269.

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

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

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

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, PA: Addison Wesley.

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.