**Meeting objectives:**

**1. Go over each of these points and discuss the best way to start creating these features**

* **I want to have a solid grasp about what features we want the program to have so that I can get them up and running as well as test them out with a bunch of different data to see what doesn’t work**

**2. Make a meeting with Daniel Barnett or someone else who has gone through this process**

* **See how he did his AND ask questions about issues that I know I’ll run into**

**3. Go over things I may have forgotten about or problems that I haven’t thought of**

* **More of these will certainly come when I start creating the interface**
* **Want to make a meeting with someone who has made one of these modules before in order to see how they take user input and use it within the code that has already been written**
  + **Email Daniel Barnett?**
* User input:
  + They choose a file from their computer and somehow my function determines which file type it is and then reads it in accordingly
    - When they choose a file to upload, does R receive the whole file pathway or how does that work? Does uploading the file upload it to the R program?
    - Ask them to specify whether they are reading in one file or multiple files?
      * Just one example of there needing to be a differentiation: if you only have one txt file to read in, you need to specify a header argument whereas the header argument throws an error if you have multiple txt files (a corpus)
    - What type of file is it? Is it a book where we’d want to create a line number variable?
      * Are there other types of documents where we’d want to preserve some of the metadata?
        + ex: may want to break down by the date a news article was published if you have a bunch over the course of some period of time

Same thing goes for tweets → how does word usage change over time?

* + - Think about having some feature where they can enter their Twitter username and it will pull their tweets
      * I’m not sure if this is possible with the API but something like it could be possible with the twitter google docs page that I found (the one that is set up to automatically pull tweets every hour for me)
    - Web scraping: should we have a place where they can paste a URL and we draw in the file from there?
      * Questions about what text to take from a webpage → just take text in html paragraph elements?
        + That leaves out bulleted lists and other things that may be important but it is certainly the easiest way to do it
  + Display raw data to user
    - To make tokens from a csv file (or really any file with more than one column of data), the user needs to know which column to tokenize
    - Are there words that we want to add to the stop word list (have them removed) OR that we want to keep (and would have been removed by the stop words)?
      * Will they be able to see the stop word list beforehand?
  + Actual visualizations: which ones do we want?
    - For books: Would we want to be able to break it up by chapter?
      * Have some feature where you can choose what to split it up by (lines or chapters) so that we can create graphs like the Peter Pan sentiment graphs
    - Word frequencies, word clouds, sentiment analysis, network graphs for concurrence and correlation, topic modeling
  + Sentiment analysis
    - Do we want to have a couple of options of lexicons for the user to choose from (and compare between)? Do we want to use just the ones I’ve been using or should I find some more specialized ones to offer as well (finance lexicons, etc)
    - Have an option to specify words to exclude (like Darling in Peter Pan)
      * Will they be able to see the lexicon beforehand?
      * Should there be some prompting for the user to actually look at the output and see that the words make sense?
  + What types of explanations should I have throughout the document?
    - A description of text analysis as a whole and how it can be useful to us
    - Maybe things you should know about your data before you begin?
      * What do you want to tokenize?
      * Do you have variables? Or is your file plain text?
      * Are there specific words that you should be weary of throughout your analysis?
    - Sentiment analysis
      * Will need to explain sentiment lexicons
      * Some description or information about breaking documents down into sections for analysis (by chapter or groups of lines)
    - Tf-idf score
      * May want to look back at Michael Barlow resources and see which statistics they use to measure a term’s “importance”
    - Further things you can do with text analysis and its application to research/academics
  + Which features are most important to include?
    - Word frequency types of things: bar charts, scatterplots, wordclouds
    - Identifying themes throughout the document: can show clustering through bar charts and boxplots
    - Sentiment analysis: tons of different types of boxplots
    - Tf-idf scores across documents: bar charts
    - Correlation and cooccurrence: bar charts and network graphs

Notes to myself:

* In order to use the unnest\_tokens function, we have to know the name of the column to extract the tokens from
  + This would have to be user input?
    - Prompt them to say which column they want to get tokens from? So should they be able to see the raw data immediately when they upload it so that they can be making these types of decisions?
* Working directory and file pathway concerns —> how do we ensure that R is able to read the file from the correct place?
  + Do they upload a file and then that is saved as a file in R and that’s how we do it?
    - This is the kind of thing that I’ve never done before
* Web scraping —> read in using html
  + Can get all of the text designated within a paragraph
  + What about other text we may want such as lists and things?
    - These are defined as different html elements
* Get whatever variable is tokenized to be named “word” in all cases so that all of the code I’ve already written works