Project pipeline questionnaire

1. What is the question you are interested in exploring?
___________________________________________________________________________
___________________________________________________________________________
___________________________________________________________________________

2. Is the question, and therefore the analysis, exploratory or hypothesis-driven?
___________________________________________________________________________
Not sure? Consult ‘Analysis types’.

3. What is your potential data source?
___________________________________________________________________________
___________________________________________________________________________
___________________________________________________________________________
Not sure? Consult ‘Understanding data sources’.

4. Have you identified what to measure and if metadata is needed?
___________________________________________________________________________
___________________________________________________________________________
___________________________________________________________________________
Not sure? Consult ‘Measures and metadata’.

5. Have you decided to evaluate the measures qualitatively or quantitatively?
___________________________________________________________________________
___________________________________________________________________________
Not sure? Consult ‘Qualitative and quantitative research’.

6. Have you identified an appropriate analysis method and tool(s) to perform the analysis?
___________________________________________________________________________
___________________________________________________________________________
___________________________________________________________________________
Not sure? Consult ‘Types of analysis and tools to use’.
Analysis types
Exploratory research is conducted to gather information or explore relationships for which we have a less-than-adequate basis to make a claim. Exploratory analyses often lead to insight that can generate a hypothesis, and therefore is often called ‘hypothesis-generating’ research. Hypothesis-driven research is conducted to test a prediction or claim and seeks to provide an explanatory relationship.

Understanding data sources
There are two basic poles of textual data; structured and unstructured. Structured text has been curated, that is, it has been prepared so that it is ready to be used in computational-based data analysis. Examples include repositories such as Old Bailey online or North American Slave Narratives Collection and to a lesser extent Project Gutenberg. Some of these repositories provide search interfaces or an Application Programming Interface (API) to access data efficiently. The opposite extreme is text that has not been prepared for data analysis and therefore needs preparation before it can be used for successfully with computers. Preparation may include any number of things depending on the source of the data; digitizing hardcopies or PDF documents, web scraping, removing artifacts from Optical Character Recognition (OCR) software or HTML, labeling text with metadata information, inter alia.

Measures and metadata
Selecting a measure and the appropriate metadata to conduct an analysis is highly dependent on the question asked and characteristics of the data. Simple measures may include counts of explicit language structure such as characters, words, phrases, etc. and/or implicit language structure, for example part-of-speech categories. However, some more complex measures are derived from calculations, log-likelihood, mutual information, inter alia. Metadata refers to the characteristics of the text that often are used to contrast texts, e.g. Republican/ Democrat

Qualitative and quantitative research
In essence, qualitative research deals with types or kinds of data. That is, what appears or doesn’t—not how many times. On the other hand, quantitative research focuses on counts or frequencies to derive relative differences in usage.

Types of analysis and tools to use
Choosing an analysis strategy is contingent on the ‘analysis type’, ‘measures and metadata’, and whether the research is qualitative or quantitative. Basic analyses include frequency, dispersion, concordances, collocations, keyword identification, etc. Popular GUI tools for these types of analyses include: AntConc, Voyant-Tools, WordSeer, and WordSmith. Popular visualization tools include TagCrowd, Wordle, and ManyEyes. For more complex analyses and visualizations such as clustering/ topic modeling and classification Weka and MALLET are popular tools.¹

¹ NOTE: The advantages to GUI software are clear —intricate analyses and visualizations can be performed with a general idea of the underlying principles of the analysis and a mouse to click buttons. However, these tools come at a price (even if it is not monetary) —researchers are bound to available features and computer compatibility. Furthermore, given the relative easy to wield these tools without a fundamental understanding of the principles behind them can lead to misuse and incorrect interpretations of results. For these reasons many researchers that frequently work with text analysis and data mining migrate towards programming languages like Python or R and familiarity with UNIX (Bash shell). Both of these languages have large development communities and are extensively documented. Although learning to program is a time investment, ultimately the ability to reliably reproduce research and freedom from the constraints of downloadable GUI software is worth it for most.