Description of Project: I am seeking a history or other humanities major to take part in text-mining research, which is a form of the digital humanities applying techniques often described as involving Big Data. The project involves collaboration with faculty and graduate students in the Department of Computer Science, and a doctoral candidate in the Department of Geography who works full-time in industry as a computer programmer.

The humanities are traditionally taught as emphasizing the close reading of a limited number of documents. Text mining emphasizes the broader analysis - sometimes called "distant reading" - of bodies of text so large that no individual could read them in a lifetime. Scholarship in the study of literature has proven fruitful in identifying patterns of language within very large sets of text that previous work has not detected. I propose to use them in the study of intellectual and political history.

My project explores very large bodies of digitized/searchable text documents, including the Congressional Record and the American Periodicals data set (which includes virtually every non-academic journal or magazine published in the United States through the mid-twentieth century). I use text-mining technology developed by collaborating programmers (database design and scripts written in the R programming language) to review these materials for data shedding light on how Americans in different periods of the nation's history discussed the role of the federal government in a variety of fields. These include the protective tariff in the nineteenth century, a subject on which I have recently published an article and am preparing additional publications. One historian has identified the tariff as the most important political issue of the nineteenth century, save slavery. Although the tariff establish rates of taxation on imported goods, many Americans in this period attached additional significance to it. Arguing that the tariff produced higher wages than competition with low-paid foreign labor would allow, the measure's advocates presented it as a social policy - in effect, an early expression of what we would now broadly call the welfare state. My own research has identified individuals and texts elaborating forms of this argument for a specific contemporary audience. Text-mining technology will allow me to explore the extent to which these arguments emerged in the larger discourses of Congress and the general public.

I use a variety of text-mining techniques, but emphasize topic modeling most heavily. Topic modeling work uses algorithms to identify individual words that occur in close proximity to each other in a body of text, and track how these relationships between words change over time. This allows researchers to gain insight into what themes, or topics, appear most commonly in the materials.
In the case of the protective tariff, Americans often discussed it as a form of "free labor" allowing workers paid high wages to accumulate enough money to open their own small businesses. Thus references to "proprietor" and social mobility often accompanied discussions of the tariff. At the same time, tariff proponents also discussed the policy as a measure that represented an advanced civilization characterized by sympathy for the less fortunate, in which the state took care to protect the weak (defined as wage laborers) against the vagaries of a market economy. In this discourse, words like "civilization," "civilized," "sympathy," "benevolence," and references to Christianity often appeared in discussions of the tariff. Historians and political scientists have emphasized the former interpretation in publications for many years. I document the existence and development of the second discourse in my publications.

The above is a single example in what will be a broader research project exploring several ways that Americans imagined and discussed the role of the federal government in the period roughly 1820-1920.

I would hope to find a humanities major interested in learning about this variety of the digital humanities. I would ask a student to read several important works in the field, which explain text mining work and its present applications. I would also ask her/him to help me in the preparation of texts for analysis, which can involve the use of basic scripting languages learned from the collaboration programmers. Finally, a student would assist me in the analysis of the topic models that the technology produces. This involves thinking about the sets of words identified as often occurring near one another in the context of their period's social, political and intellectual life, and discussing what their frequent co-occurrence might suggest. It also involves digging into specific instances in which these words accompanied one another in order to ascertain exactly how they related to each other in that case.

In addition to a general familiarity with text mining, a student working on this project would gain knowledge of the specific formats used to store and deliver large data sets; the metadata standards associated with them; and the processes that together make up a text-mining work flow. I want to be clear; it is very unlikely that this student will be able to take the skills s/he learns and find full-time employment in the field of data science (unless they already possess considerable technical skills). Rather, I hope to introduce a student to the analysis of data at scale, a general activity that has recently become increasingly important in many fields.

Although advanced technical skills are not required for the student's work, a general familiarity with and interest in digital and information technology would be very helpful.