1-1-2011

Text Mining for the Social Sciences

Walter Morris
Dyson College of Arts and Sciences, Pace University

Text Mining for the Social Sciences--Final Report

1. Text Mining for the Social Sciences
2. Thinkfinity Cornerstone #3
3. Dr. Walter Morris Dyson College of Arts and Sciences Economics Department
4. January 22, 2012

The original proposal identified Text Mining as a technique using unstructured text documents that can be transformed into measurable values for later use in a statistical analysis. The procedures usually attempt to identify the presence or absence of typical words or phases in a wide range of documents that ultimately can be used for purposes of classification. For example, one such application of Text Mining was a statistical analysis of the most commonly reported phrases the candidates used during the final two months of campaigning in the 2008 presidential election. Using data bases collected from news reports and blogs several common phrases were identified and a statistical analysis was conducted to determine the relative frequency of each recorded response. The results indicated that the most frequently reported phase through these media outlets was 'You can't paint lipstick on a pig'; the second was 'It's the economy stupid'; and so on.

What was of interest to me was the existence of several computer programs that easily make the transition from text to data analysis in order to evaluate information contained in written documents. One such program is SAS Text Miner (a component of SAS Enterprise Miner 6.1). Together with my Research Assistant, Robert Hamilton, we have successfully installed the Text Mining node into SAS Enterprise Miner and have researched several applications in the Social Science area.

For purposes of this grant, I was able to secure the entire set of 85 Federalist Papers (in a SAS file) which were analyzed from a quantitative Text Mining perspective. As you may know 50 of the 85 original Federalist papers are attributed to Alexander Hamilton, 17 to James Madison, and 3 to John Jay—leaving 15 unaccounted for and author(s) presumably unknown. Since my grant is to illustrate exactly how Text Mining can be utilized for the Social Sciences, authenticating the documents through a combination of Text/Data Mining is one of the applications pursued. Stylographic authenticity of this sort could have many applications for Political Science as well as History majors.

On the basis of vocabulary usage and styling, we used Text Mining for the purpose of classifying these 15 unknown works. To achieve this objective, we initially took a close look at the 70 known writings and estimated a quantitative model based on the text documents converted into a data structure conducive to statistical analysis. Using these 70 known Federalists papers, we first trained a quantitative model through a combination of logistic regression, classification trees, neural networks, and memory-based reasoning. After determining a high degree of statistical accuracy on these known writings using the above statistical techniques, we selected

an ensemble model in an attempt to 'forecast' the authors responsible for the 15 unknown works. In essence we were able to estimate probabilistic values that a certain writer produced a particular document. For instance, using the first unknown document, transforming the text into a data format, and 'running' the transformed document through our trained statistical model, we estimated a 96% probability that Madison wrote the document, a 32% chance Hamilton produced it, and a 5% likelihood that Jay was the author. We then proceeded to classify the remaining 14 documents in similar fashion.

The results were quite interesting. According to our statistically based analysis that computed conditional probabilities for each document, the analysis concluded that 14 were authored by Madison, 1 was attributed to Hamilton, and none were written by Madison. Separately, a canvas of academic Historian's and Political Scientist's revealed a consensus that all of the 15 papers were written by Madison. With the exception of a single document, it's quite noteworthy that our statistical analysis closely corresponds to the Historians/Political Scientists views. Maybe these scholars may wish to re-visit and review their findings concerning the 'errant' document in question?

With regards to disseminating this information to the University community, I have incorporated Text Mining into several of my classes beginning these Fall/Spring semesters. My Economics 240 (Quantitative Analysis and Forecasting) course already included Data Mining and will add a Text Mining component beginning this Fall Semester. Three sections of the Eco 240 course were offered this Fall 2011 semester with approximately 50 students enrolled to date. My Economics 385 (Econometrics) and Finance 325 (Data Analysis for Finance) courses will include both Text/Data Mining components in the 2012 Spring semester—expected enrollment is in the 25 student range. In addition, this Spring semester will include another Eco 240 class with a 25 student enrollment. I've attached the syllabi for both courses and may I direct your attention to Topic #6 for the Econometrics (Eco 385) course and Topic # 7 for the Quantitative Analysis and Forecasting (Eco 240) course. Finally, together with the mathematics department, I am presently developing a new course, 'Statistical Methods for the Social Sciences' (Math 143), and Text Mining will be an integral part of this curriculum. The Federalist Papers previously alluded to above will have particular relevance to the Mat 143 course.

Also since returning from my sabbatical I have had opportunities to engage faculty with regards to the Text Mining project. In the Spring '11 semester, I'll be advising and mentoring my research assistant, Mr. Robert Hamilton (no relationship to Alexander) who is a double major in Economics and Mathematics, and will be presenting a paper on the topic of 'Text Mining for the Social Sciences' at the Dyson Society of Fellows annual meeting held on March 3, 2012. In essence Mr. Hamilton will be presenting the research results described above with regards to stylistic authorship of historical documents as well as other outcomes from the research grant. We are expecting a large turnout of both faculty and students for this event.

To date I have completed the following computer related seminars that related to the Text Mining for the Social Science proposal.

THE FOLLOWING ARE A LIST OF THE SAS COURSES COMPLETED WITH THE CORRESPONDING DATES.

1. 4-DAY ON-LINE SAS WORKSHOP 'PREDICTIVE MODELING USING LOGISTIC REGRESSION', FEB 22-25.
2. 4-DAY ON-LINE SAS WORKSHOP 'TEXT MINING AND SAS ANALYTICS' MAR 10, 11, 14-16 WITH ROBERT HAMILTON.
3. 4-DAY ON-LINE SAS WORKSHOP 'Applied Clustering Techniques', Mar 22-25.
4. 1-DAY ON-LINE SAS WORKSHOP 'Stationarity Testing and Other Time Series Topics', May 13

THE FOLLOWING ARE A LIST OF THE ON-LINE STATISTICS.COM COURSES COMPLETED.

1. 'LOGISTIC REGRESSION', MAR 11-APR 8.
2. 'ADVANCED LOGISTIC REGRESSION', APR 15-MAY 13
3. 'TEXT MINING', JUNE 15-JULY 15