Pace University
DigitalCommons@Pace

Cornerstone 3 Reports : Interdisciplinary Informatics

The Thinkfinity Center for Innovative Teaching, Technology and Research

5-1-2011

# Text Mining for the Social Sciences

Walter Morris (Principal Investigator)
*Dyson College of Arts and Science, Pace University*

# Mid-Term Status Report

Project Title: Text Mining for the Social Sciences

Cornerstone # 3

Principle Investigator: Dr Walter Morris Dyson School of ?Arts and Sciences

Date: May 30, 2011

The original proposal identified Text Mining as using unstructured text and documents that transforms them into measured values such as the presence or absence of words. There exist sophisticated computer programs that make the transition from text to analysis in order to evaluate information contained in written documents. One such program is SAS Text Miner (a component of SAS Enterprise Miner). I, together with Robert Hamilton my Research Assistant, have successfully installed the Text Mining node into SAS Enterprise Miner.

For purposes of this grant, I was able to secure the entire set of 85 Federalist Papers (in a SAS file) which will be analyzed from a quantitative Text Mining perspective. As you may know 50 of the 85 original Federalist papers are attributed to Hamilton, 17 to Madison, and 3 to Jay—leaving 15 unaccounted for and author(s) presumably unknown. On the basis of vocabulary usage and styling, Text Mining can be used for the purpose of classifying 15 of the unknown works. 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 I'm currently pursuing. Stylograpic authenticity of this sort could have many applications for Political Science as well History majors.

The Text Mining component of Data Mining will be incorporated into several of my classes beginning this Fall/Spring semesters. My Economics 240 (Quantitative Analysis and Forecasting) course already includes Data Mining and will comprise Text Mining beginning this Fall Semester. Three sections of the Eco 240 course are offered this Fall 2011 semester with approximately 50 students already enrolled to date. My Economics 385 (Econometrics) course will include both Text/Data Mining components in the 2012 Spring semester—expected enrollment is in the 25 student range. Finally I am developing a new course, 'Quantitative Methods for the Social Sciences', and Text Mining will be an integral part of this curriculum. The Federalist Papers previously alluded to will have particular relevance.

Since I've been on sabbatical I haven't an opportunity to engage faculty with regards to the Text Mining project but fully expect to do so upon my return in the Fall 2011 semester.

To date I have taken the following computer related seminars that relate to Text Mining.

THE FOLLOWING ARE A LIST OF THE SAS COURSES ALREADY TAKEN.

1. 4-DAY ON-LINE SAS WORKSHOP ‘PREDICTIVE MODELING USING LOGISTIC REGRESSION’, FEB 22-25.
2. 4-DAY ON-LINE SAS WORKSHOP ‘SAS ANALYTICS USING SAS TEXT MINER’, MAR 10, 11, 14-16.
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 STATISTICS.COM COURSES ALREADY TAKEN.

1. ON LINE ‘LOGISTIC REGRESSION’, MAR 11-APR 8.
2. ON LINE ‘ADVANCED LOGISTIC REGRESSION’, APR 15-MAY 13