CCSU

DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Wednesday, December 4
1:00 – 2:00 PM
Copernicus Hall (NC), Room 20102

IMPROVING WORKPLACE ACCIDENT FATALITY CLASSIFICATION MODELS WITH TEXT MINING AND ENSEMBLE METHODS

THOMAS WILK

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: Using publically available OSHA accident investigation data and open source Python scientific computing tools this applied research project asserts the value of exploiting all types of data available when constructing predictive models. The goal of this project was to build the best classification model of accident outcome, either fatal or non-fatal, using data available for catastrophic accident investigations conducted by OSHA over the last few decades. Multiple feature sets were engineered using a variety of techniques that leveraged all types of data available, including structured, semi-structured and unstructured accident attributes. This thesis proposes that features mined from each accident's text-based attributes will capture concepts and information that are not present in each accident's structured data attributes and that this infusion of new information will enable classification algorithms to better discriminate between fatal and non-fatal accidents, thereby improving model accuracy. Baseline classification models of accident outcome were trained on a feature set created from the structured accident attributes only. With the goal of improving baseline classification accuracy a variety of text mining and data mining techniques were employed to create statistics-based and linguistics-based feature sets from accident keywords, descriptions and summaries. These efforts resulted in a measurable improvement in model accuracy.

For further information:

gotchevi@ccsu.edu 860-832-2839
http://www.math.ccsu.edu/gotchev/colloquium/