Data Science @ Westat

Natural Language Processing

Using both traditional and cutting-edge deep learning techniques, Westat data scientists implement natural language processing (NLP), aimed at training computers to understand and process human languages. Whether faced with written or spoken language, our NLP experts foster contextual understanding of content to help clients effectively process data and develop appropriate solutions to the challenges they face.

Using Technology to Find Solutions

Westat uses a variety of innovative technologies and supportive tools to fit specific client requirements. We have implemented support vector machine methods, lexicon-based approaches, NLP in neural network models, and more. Those solutions include Python, R, and C++ routines supported by libraries such as TensorFlow, Flask, and Marshmallow, as well as AFFIN, Bing, and NRC lexicons, and word-embedding algorithms, including GloVe.

Abstracts/Reports • Disparate Data Sources • Open-Ended Text • Verbal Sentiments • Form Entry Mismatches • **Westat NLP Expertise** • Information Leakage Protection • Data Harmonization • Coding Efficiency Assessment • Quality Control & Catgorization • Improved Matching Algorithms & Forms

Illustrative Projects

- For the **National Institutes of Health (NIH) Office of Disease Prevention**, Westat has performed exploratory testing of NLP approaches on submitted abstracts, including the use of NLP to classify and code NIH abstracts to augment human review. We have also developed Java-based programs to recognize names, locations, and institutions to assess and prevent information leakage on submitted abstracts.

- For the **Centers for Disease Control and Prevention's National Center for Health Statistics**, Westat created a custom-built text classification system using support vector machine methods. This tool promoted categorization and standardization of medical payer names contained in a data set that integrated data arriving in multiple structures, including national standard billing claims, patient clinical documents, and custom extracts from hospitals and health care organizations. The program also categorized cases according to a set of final standardized codes, reducing the need for manual coding.

- For the **Substance Abuse and Mental Health Services Administration's Drug Abuse Warning Network (DAWN)**, Westat applies neural network machine learning models that incorporate NLP, text analytics, and deep learning techniques on drug-related emergency department visit data collected by medical record abstractors. The models score how well human abstractors apply DAWN case definition rules in their reviews and whether they assign correct analytic group categorizations. By using these models, all data are subject to a high quality control standard while reducing the need for expert data review by staff.

- To assess the efficacy of the **From Here to There**, an interactive touch-based application that allows students to interact with numbers and expressions, Westat conducted a sentiment analysis to understand students' experiences with the learning technologies used. Our data scientists used AFFIN, Bing, and NRC lexicons to examine students' word-level and sentence-level responses. We conducted a methodological study that compared two approaches for processing and analyzing qualitative data and compared the coding findings from traditional qualitative coding and topic modeling, unsupervised machine learning method. During the preprocessing phase, we applied the NLP to annotate the test (i.e., assigning part of speech) and extracted nouns and verbs. Latent Dirichlet Allocation (LDA) modeling was used to classify the underlying topics in the data.

Our Experts

- **Atsushi Miyaoka** brings extensive knowledge of statistical applications and analytic solutions, including data mining, NLP, and text analytics to diverse research projects in education and immigration.

- **Eric Pan** is a senior physician informaticist and a leader in health informatics research, currently doing exploratory work in applying NLP to complex abstract submissions.

- **Marcelo Simas** leads design, development, implementation, and operations of data processing and analytic systems, including developing systems architecture and implementing data science techniques.

CONTACT

Ting Yan
Vice President and Associate Director
(301) 251-4220
tingyan@westat.com