SentNet

Next Generation Document Scoring Using Advanced Graph Analytics

Executive Summary

The volatility of the world in the twenty-first century necessitates that our nation's leaders have access to detailed and timely intelligence products that allow them to make informed decisions in a rapid manner. However, the analysts that produce these products are increasingly being asked to collect, interpret and synthesize insights from ever growing volumes of data stretching their time and existing resources. Given this inherent tension between the IC customers' needs and analysts' limited time and resources, it is clear that ODNI and other IC agencies could benefit from new tools and methodologies to assist analysts in drafting accurate and timely intelligence products. While there are numerous existing Natural Language Processing (NLP) tools and methods, such as Automatic Essay Scoring (AES), none fully meet the requirements that ODNI outlined for a robust intelligence product scoring solution.

In response to this need, we have developed and implemented an analytical prototype called SentNet. In short, SentNet is an advanced document scoring tool designed for the Intelligence Community that allows agencies to define and develop highly accurate models to automatically score intelligence products providing analysts with instant feedback about the content and quality of their reports. Combining methods from academia, the commercial sector, and the open-source community, SentNet is able to identify more nuanced linguistic features that can be used to predict how an intelligence product will score against establish ODNI or agency specific criteria. In contrast to existing NLP methods SentNet offers analysts and agencies the following capabilities:

1. **Advanced Feature Extraction:** SentNet uses cutting edge machine learning and linguistic methods including WordNet, Doc2Vec, linguistic network analysis, Louvain clustering, and perceptual hashing to develop 12 classes of features for each document. When combined, these features allow SentNet to "see" more nuanced structures within documents compared to other NLP methods.
2. **Model Chaining:** Within SentNet document level predictions are dependent on not one, but multiple models, each tasked with a different function. SentNet then constructs a final model to ensemble (combine) outputs from these multiple input models and features resulting in more accurate and stable predictions.

In this paper, we discuss how we implemented a prototype of SentNet using standard open-source software libraries in conjunction with customized code. To test its performance, we developed 43 different models using SentNet and compared the predicted document scores from those models to scores assigned by professional human graders. These essays covered a

variety of topics and essays were scored according to a range of different criteria. Initial results from these models are promising. SentNet correctly classified 61% of essays, producing a score identical to that given by a human reviewer with an overall Root Mean Square Error (RMSE) of only 0.6. This means that, for most models, 98% of SentNet's document score predictions fall within one point of the professionally assigned score. Given the relatively limited amount of time allocated to model tuning/adjustment, this performance demonstrates that SentNet's methodology is fully capable of rapidly evaluating and numerically scoring brief analytic reports with no human intervention.

Additionally, given our team's decade of work in the intelligence industry and the data science domain, including the delivery of tools and solutions involving graph-based text analytics, we firmly believe that SentNet has the ability to transform the IC analytic toolset. Our unique solution, combined with industry-specific experience, offers a creative and robust solution to IC analysts and agencies.