WHITE PAPER

Natural Language Processing

Natural Language processing is a key component of Kingfisher’s VARYSS platform

Prepared By:

Kingfisher Systems, Inc.
3110 Fairview Park Drive, Suite 1250
Falls Church, VA 22042
Phone: (703) 635-2951
Fax: (703) 820-7976
Email: varyss@kingfishersys.com

Kingfisher has deep expertise interfacing Natural Language Processing (NLP) pipelines with core algorithms in Artificial Intelligence, including deep neural networks. Our focus has been extracting actionable quantitative intelligence from open source media. Larger datasets and enhanced tools for topic discovery allow us to discover complex and emergent topics, and the combination of high frequency time series analysis and automated entity and concept discovery provides increased understanding of causes and effects.

Both academic and commercial NLP tools have been used in text processing pipelines for over 20 years, but without enhancement these tools are poorly optimized for large-scale quantitative analysis. Academic tools provide moderate fidelity but scale poorly due to high computational cost. Purpose-built tools for intelligence work were initially hampered by the lack of tagged corpora, leading to overfitting and poor out of sample performance. Recent web scale statistical approaches massively increase performance across a range of NLP tasks, and many of these tools have open APIs, but those we have looked at are optimized for tasks found in the commercial market - particularly sentiment analysis. In short, common NLP tools are not sufficient for quantitative politics or social media exploitation, an effect primarily due to the loss of salience and context when entity and sentiment are the objective.

We recognized these deficiencies and spent over 10 years enhancing existing and new algorithms specifically for quantitative intelligence. While we are centered on entity, relationship, and concept extraction, our expertise extends beyond proficiency in standard NLP (*e.g.* topic discovery using stochastic generative models, or tagging, chunking, and shallow parsing). Our platform was first developed using open source news reporting, and and later extended to other domains such as social media and technical information. We currently leverage a database of structured information obtained from over 400 million documents, which allows us to address the lack of tagged data using semi-supervised and unsupervised approaches.