Title: Named Entity Recognition for Open and Deep Web Data - GOST Crawl

Keywords: Machine learning, data science, named entity recognition, statistical modeling, entity extraction, deep networks, natural language processing.

About Us: At Giant Oak we make the world a safe place by applying behavioral and computer science in big data environments to identify illicit behaviors, actors, and networks. We build software to help our customers effectively and efficiently reduce crime, fraud, coercion, and violence.

Our main product is GOST, which indexes the open and deep web to create custom information domains that include standard list-based data feeds, proprietary data sources, and up-to-date publicly available information on the internet. One of our main methods for building up a deep corpus of web data is through GOST Crawl, which iteratively "crawls" through the web and stores content that contains vital information on nefarious actors.

Problem Statement: After Giant Oak has retrieved publicly available open-source data, it is indexed in ElasticSearch according to a list of text-extracted attributes, such as names, addresses, and links.

The student's goal is develop and explore methods in Named Entity Recognition (NER) and Text Quality and Content Extraction, both of which are open problems in Natural Language Processing literature. For NER, the students will build upon the results established by the previous Clemson capstone team and Giant Oak Science members, whereas for Text Quality and Content Extraction, the student's goal will be to implement and recommend a state-of-the-art model architecture from literature.

Resources: The resources available for this project include but are not limited to the following.

1. Access to publicly available open-source data previously generated by Giant Oak.
2. GPU training using the Clemson University Palmetto Cluster.
3. Results of Name Entity Recognition experiments from the previous capstone team and essential results from the Giant Oak Science team.
4. Opportunities to meet regularly with the members of the Giant Oak Science team to discuss methods and results while diving into the culture of Giant Oak.

Responsibilities: You will be responsible for building upon the entity extraction and entity recognition methods used for GOST Crawl, gauging success by the following metrics.

1. Ability of the NER model's ability to extract entities from text.
2. Ability of the text quality and content extraction model's to distinguish NER-parsable text from noisy open web data.
3. Ability to scale in a cost effective way, both model computation and entity indexing
4. Effectiveness of the extracted entities in a simulated production environment

Your work will contribute to the advancement of knowledge in the Natural Language Processing field and will be made publicly available.

About You: You are an aspiring data scientist, machine learning engineer, or researcher in a related area. You are interested in expanding the ideas of what's possible with data, and you aren't afraid to get your hands dirty in feature engineering or unwieldy model training. You are very interested in the intersection of natural language processing and machine learning, and want to use these tools to help make the world a better place.