Joonpyo Jun (joonpyo3@illinois.edu)

Professor Zhai

CS 410: Text Information Systems

2 November 2021

**Industrial Usage of Apache**

            As the 4th industrial revolution is ongoing in various fields in technology, the Natural Language Processing (NLP) became one of the most important fields of study. Good NLP provides a better quality of media, resources, and other technology-related services to people who need them, especially for the data analysts and text processing software providers. Among various NLP sets of tools, I was intrigued by the Apache OpenNLP toolkit that provides several core NLP tools such as Tokenization, POS Tagging, Parsing, Natural Language Generation, Feedback Analysis, and more. Because the text processing service is most likely to be introduced in the industry sooner or later, I started to search for the project or research that utilized the Apache OpenNLP. From IEEE, the Automatic Approach for Discovering and Geocoding Locations in Domain-Specific Web Data attracted me. So, I decided to dive into this research paper and explain how the core functionalities of Apache OpenNLP can be implemented in the industry and potential research.

            Briefly, the Apache OpenNLP is an open-source Java library to process natural language text, providing services to handle raw text efficiently. The API and CLI are available for free via its website, <http://opennlp.apache.org/>, while individual models are also available for downloads. For example, you can specifically download Tokenization (as a .bin file) without other functionalities. Specifically, this project (Geocoding Locations) only utilized the location entity extraction model (en-ner-location.bin from the most recent manual). I could conclude that each OpenNLP model is perfectly individual for a specific use without implementing the entire API. So, what does the location entity extraction model do? This model provides a class called NameFinder that takes the input stream (tokens) and trained location model and identifies the locations in the input. After that, these identified locations are taken into general NLP steps, Tokenization, stop-word removal, and stemming.