ARTIFICIAL
SOLUTIONS

AnonyMate: Artificial Solutions' Approach to Anonymizing Unstructured Text Data

Allison Adams

Large amounts of data are required for many NLP tasks, which poses a considerable challenge to companies who want to prioritize the data integrity and privacy of their clients while operating in this space. Personal identifying information (PII) is often present in human-computer dialog data, and as such, steps to remove sensitive information through anonymization are essential. To address this need, we at Artificial Solutions developed our anonymization tool, AnonyMate, with two main objectives in mind:

- To ensure that historical data stored for R&D purposes do not contain any PII data.
- To enable our platform to produce anonymized data.

In light of these objectives, our goal was to build a tool that can identify and classify types of PII data and apply different anonymization strategies on the detected PII types. We further sought to detect and annotate named entities beyond the scope of anonymization purposes as a part of this pipeline.

The scope of this project was broad and encompassed diverse tasks including: the development of a tag set of PII and named entity types with guidelines for annotation, a large-scale annotation effort in multiple languages, and the development and testing of Named Entity Recognition and language identification systems. We aim to present an overview of this project and our pipeline architecture, as well as to discuss the unique challenges we faced while developing this system for anonymization.