Emoveo: Application of Natural Language Processing to Document Redaction

The proposal would enable efficiency, performance, and cost-effectiveness of implementing systematic, secure redaction solutions with the application of automated machine learning based systems. The project is grouped into three parts: (1) content extraction, (2) user flexibility, and (3) maintaining readability. Content extraction will be achieved through syntax tagging and keyword tree. The end goal is enabling users with reliable and consistent methods allowing for end-users to use context extraction to perform semantic and pragmatic analysis on the document.

Information access control is the key player in document redaction. While significant portions of concerns are damage and privilege controls, traditional methods over redact the documents leaving but a few words on the page. Moreover, security measures are deemed useless when documents are improperly and under-redacted in word documents and adobe reveal sensitive meta-data and the text under the redaction boxes.

Algorithms used consists of syntax taggers, keyword generators, and a concordance search engine. The syntax tagger allows for the program to identify nouns, noun phrases, and other phrases reducing the overhead costs of performing Named Entity Recognition (NER). Instead of having to rely on large databases and building exceptions, the entire process is reduced to automatic search as well as custom search through the concordance engine to help the user make decisions at the semantic and pragmatic level. This allows the user to redact important information while keeping the topic and meaning of the document generally the same. The keyword generator relies on word frequency and the removal of words that do not contribute meaning to give users ideas what bodies in the document are describing allowing users to identify sections of text at a time. The keyword generator is used on text and internet resources as a way to generate low memory cost word-relation trees to identify sentences and phrases pertaining to that key word.

To increase the effectiveness of the application, the user is allowed to view every step of the process and loop back to correct false positives and manipulate the data retrieval tools of provided solutions. In this situation, the user is effectively able to work with the machine to minimize the error rate that might occur.

Intellectual Merit: The research proposal provides a new scheme of improving work flow efficiency in document redaction. The potential of the program is for end users to develop computerized problem solving aptness over time and improving overall quality of redaction. Interaction with effective and simple instructions paired with user flexibility allows for the program to augment novel solutions in real time and with more accuracy than traditional methods of redaction.

Broader Impact: This work has the potential to improve document redaction and security mechanisms across the board. The use of text files eliminates the threat of leaking metadata and limits information manipulated to computer memory. With proper review over the redaction methods, users are then able to effectively and quickly process heaps of information quickly and with the expectation of accuracy. The applications of effective algorithms in content analysis, information extraction, and language structure algorithms allows for effective manipulation of raw data.