SPII Guard: Securing Personal Identifiable Information in Receipts Using Regular Expressions and Named Entity Recognition

Katoh, Erina (School: Booker T. Washington High School)

With the need for collecting and analyzing large-scale data dramatically increasing in the big data era, researchers and companies have begun to rely on crowdsourcing platform workers to annotate user data. With around 15.1% of all tasks on Amazon Mechanical Turk (MTurk), a popular crowdsourcing platform by Amazon, being the transcription of receipts from third parties, the information of third-party customers' personal identifiable information is at risk of falling into ill-intentioned people. Receipts that are uploaded onto crowdsourcing sites may contain a variety of privacy-risk information, such as times, dates, and third-party users' last 4 digits of a credit card, and more. Therefore, in this research, a novel algorithm is created, detecting sensitive information within receipts to protect third party users' information before crowdsourcing workers transcribe the receipts.   To begin identifying sensitive information in receipts, PyTesseract optical character recognition, which assists in converting text in pictures to machine readable format, is uniquely implemented. To detect personal identifiable information, Regular Expressions and Named Entity Recognition are utilized. This novel tool can be implemented by those who assign tasks on crowdsourcing platforms so that receipts with sensitive information are not assigned as tasks.   There is currently no other algorithm created to directly address this urgent issue of protecting sensitive information on crowdsourcing platforms. Therefore, in the research, a novel solution is created to detect privacy-threatening information, specifically dates, times, names, and last 4 digits of credit card, in pictures of receipts to protect individual privacy on crowdsourcing platforms.