Jose Ignacio Velarde
Fall 2016
Professor Regina Barzilay
October 30, 2016

My UROP project for this term is "Extraction of Relevant Metrics from Medical Articles". I am working with the Bayes Mendel group from Harvard/MGH to find a way to automatically extract relevant information from cancer research articles. The group is working on a web application used by doctors that calculates a patient's risk for certain types of cancers based on their genetic mutations. In order to determine the risk associated with each genetic mutation, these researchers review previously published articles that describe links between certain mutations and certain cancers. The group then determines what the risk is based on the information from the articles. One of the biggest problems the group faces is determining the reliability of each article. Many of these articles offer conflicting information about the correlation of certain gene mutations and certain types of cancer. To account for this, the group reads every article and rates each one on its reliability.

An article's reliability is determined using the following metrics: the number of participants in the studies and the number of studies conducted. We are finding a way to automatically extract these two quantities from a medical research paper. Then, we hope to expand the scope of the extraction to include more information that would be helpful to the researchers, such as the genetic mutation specified in the paper, the type of cancer, gender of participants, study type, etc. We are approaching this task as a classification task. We are trying to classify words in a paper as either describing the information we want, or as describing information that is not relevant.

Thus far, under the guidance of Professor Barzilay, I have created a system that converts PDFs of research papers into text files, converts each word in the text files into a vector, and then uses this vector representation to classify each word as either relevant or not relevant. I am currently improving the system's accuracy in classifying words. I am also testing different models for classification, such as logistic regression, support vector machines, and conditional random fields.

I am interested in this project because I care a lot about the impact that my work has on the world. I hope to use the skills that MIT has given me to help others, and I think this project is a good way to do it. Finding a way to automate the extraction of these quantities can save researchers a lot of time when they are determining the reliability of an article. This will allow them to review more articles and expand the scope of their risk calculator, which could help

thousands of doctors nationwide. In addition to this, I find the problems posed by this project to be challenging and relevant. The internet has greatly increased the sharing of information amongst people. Developing good ways to extract relevant information could benefit researchers all over the world and advance everyone's knowledge. I am extremely excited to be working on this project and further develop my skills in processing text.