Project Proposal
CSI5386 Natural Language Processing
By Michel Custeau 8658589 and Beril Borali 300036112

For the end of semester project, we have decided that we will be exploring the application of statistical NLP on legal documents. This area of research is continuously growing as it has important potential impacts on the real world and the field of law. Some of the main drivers for progress in this area are pretrained models on legal documents being released open source, such as LegalBert, and also competitions focusing on NLP tasks for legal documents.

For our project, we will be focusing on the second task of the COLIEE 2021 and 2022 competition offered by the University of Alberta. This task focuses on measuring entailment between legal texts. Specifically, it involves identifying a paragraph from an existing case that entails a new given case. These new cases and potential entailing paragraphs are given to us in the form of a separate training set and testing set.

The purpose of the project will be to perform an evaluation of different entailment measuring methods. We will be evaluating these methods on two datasets, the first one being the 2021 competition dataset, and the second the 2022 competition dataset. For these datasets, we will compare and evaluate the performance of both non neural network and neural network methods to see which one can better and consistently identify entailing legal paragraphs. These methods will include BM25, sBert, Legal Bert and T5.

We think this is an important project to work on, as it will touch on Michel's masters thesis, and also Beril's job at Statistics Canada. More broadly, finding entailment between texts is something we believe can be applied in a myriad of future opportunities in industry and academia as NLP becomes more prominent in our society.

Links to competition pages:
2021 competition: https://sites.ualberta.ca/~rabelo/COLIEE2021/
2022 competition: https://sites.ualberta.ca/~rabelo/COLIEE2022/