Nafisa Ali Amir (namir2@jhu.edu)
Avais Pagarkar (apagark1@jhu.edu)

26 April 2019

Legal Information Retrieval: Precedence Retrieval using Catchphrase extraction

Legal profession involves dealing with large amount of text. With the evolution in natural language processing, the amount of digitized textual data is increasing as well. This provides an opportunity to use the tools of machine learning and statistical information retrieval for the benefit of the people in the legal profession. We chose the task of precedence retrieval using catchphrase extraction since it encompasses many subproblems within.

Precedence retrieval essentially involves looking up prior cases and retrieving the most relevant ones with respect to a case under consideration. This is one of the most common use cases for lawyers and judges. FIRE 2017 IRLeD Dataset [1] provides two annotated datasets containing Indian Supreme Court cases. The first dataset is for catchphrase extraction containing 100 training samples and 300 testing samples and the second one for precedence retrieval contains 200 query cases and 1000 prior cases. We are open to other datasets for the task if available.

We plan to approach the problem by creating sub-problems and tackling them separately. Sub-problems include parts-of-speech (POS) tagging, candidate catchphrase extraction based on POS, vector embedding for phrases (multiple words), classification of catchphrase from the candidates using machine learning model, computing similarity measures of cases using the catchphrases extracted, rank the relevant precedent cases, evaluate the model based on precision and recall measures specifically mean average precision and mean reciprocal rank.

Works Cited

[1] Mandal, Arpan, et al. "Overview of the FIRE 2017 IRLeD Track: Information Retrieval from Legal Documents." *FIRE (Working Notes)*. 2017.