Crime and Punishment

A machine for federal crime sentences

By Carlos Grandet & Hector Salvador

Objective

Make an automatic question-answering program that finds criminal and administrative sentences (e.g. jail years, fines, political disablement) in Mexican federal laws.

Data sources

- Active federal laws (~300, http://www.diputados.gob.mx/LeyesBiblio/index.htm). These are available as word files, and can be easily scraped from the website.
- We will explore using sources such as Yahoo! Answers, online newspapers/news blogs, and other websites with relevant information (e.g. specialized websites on law)

Expected original work

According to Jurafsky and Martin (2016), there are three stages for question answering: question processing (decomposing the question into a query that can be then "asked" to the documents), passage retrieval (take que query and the corpora, find relevant documents, find relevant passages, and pass these to the next step), and answer processing (returning the retrieved passage into a human-readable result).

Thus our project will do the following:

1. *Data generation*
    - Scraper for the Congress website. A program that processes documents into pure text (e.g. remove headers, remove old decrees, remove page footnotes and numbering).

2. *Query processing*
    - We will create a function that transforms a question into a query to apply Information Retrieval algorithms. We will limit it to questions related to fines and jail times and if there are no matches, suggest an alternative query.
    - Question classification should be very similar for all our queries. There exist some answer type taxonomies that can be built semi-automatically (Wordnet) but they can also be built manually.

3. *Passage retrieval*
    - Use information retrieval techniques (e.g. DF-IDF) to obtain the relevant sections, articles, and/or paragraphs.
    - Thesaurus generation - Clustering words that are close enough in the text.

4. *Answer processing*

    - Function that uses regular expressions and a trained model of answers to indicate if an answer is the right type of answer (year, fees, etc)

Evaluation of results

We will use *Precision/Recall* curves and Mean Average Precision to evaluate if an answer is the type of answer we are expecting and if it is correct. To do so, we will create some questions with answers to test our machine.

We will also use the *Mean Reciprocal Rank*, which is commonly used to evaluate the relevance of ranked results (according to the predicted relevance).

Timeline

- Week 5:
    - Generate scraper and store data (Carlos)
    - Design IR methodology for passage retrieval (Hector)
    - Design algorithms for query formulation and answer processing (Hector)

- Week 6 :
    - Generate thesaurus of related concepts to penalties and jail sentences (Carlos)
    - Implement and test function for question processing (Hector)

- Mid-quarter check-in
    - Show working functions for question processing, show corpus of concepts related to penalties in the Mexican law. (Carlos & Hector)

- Week 7 - 8 :
    - Use IR algorithms to retrieve queries from Mexican laws related to penalties and sentences (Carlos)
    - Implement and test function that processes answers (Hector)
    - Generate testing set of questions. (Carlos & Hector)

- Week 9 - 10
    - Test and evaluate model and tune in through extensions (Carlos & Hector)
    - User Interface* (if time permits)

- Final presentation
    - Present a program that asks a series of predetermined and non-predetermined questions and provides an answer. (Carlos & Hector)

Expected libraries that will be used

- Shell scripts
- BeautifulSoup + Requests (for scraping laws)
- Pandas (Feature preprocessing)
- Scikit-learn (for model cross-validation and other out-of-the-box ML tools)

- Possibly others for nlp (e.g. nltk)