Project Proposal for a
Final Year Project of
Computer Science
In Partial Fulfillment of the Requirements
For a Bachelor of Engineering Degree (Computer Science)

The University of Hong Kong
Department of Computer Science

Yeung Tsz Lok (3035366788), Yu Tung Chuen(3035377127), Yang Lingqin (3035330404)
Supervisor: Kao Benjamin

Proposed Date: 25 September 2019

1 Tentative Project Title:

Information Extraction from Hong Kong Court Cases

2 Abstract

Studying past rulings by tribunals plays an essential role in the work of legal professionals. However, such task usually consumes a fair amount of effort for comprehension and is considered a tedious job. To remedy such an issue, this project aims to apply Natural Language Processing (NLP) technologies to automatically extract relevant information from Hong Kong court cases, and thus reducing the prohibitive reviewing cost. Beyond this, the project group believes that there is a demand for tools across various disciplines to harness the exhaustively parsed court case data, either in visualization, prediction, or classification. Nevertheless, the development of such tools is tentative, as the main determinant still lies in the quality of parsing data obtained from the previous stage.

3 Project Description

This chapter covers the project background, a review and comparison to previous work in information extraction from court cases, and an architecture design of our project.

3.1 Specific Aims

This project aims to develop a system that extracts pertinent information from Hong Kong court cases, especially with regard to Drug Trafficking cases. The system might be dealing with two aspects of data, the low-level basic information and judgment-related factors, respectively.

3.2 Background

The law of Hong Kong is based on the amalgamation of English common law and local regulations codified in the ***Laws of Hong Kong*** (“Law of Hong Kong,” n.d.). Thus, stemming from the dogma of stare decisis, which demands similar decisions for similar facts with regard to principled rules, it is crucial for legal professionals to study precedential rulings of relevant courts and synthesize standards applicable to the current facts(“Common law,” n.d.). Whereas, such routine is usually very cumbersome and time-consuming as legal practitioners have to abstract the apposite information from a fair amount of precedent cases. Therefore, the common practice is only reviewing a limited number of court cases but hopefully captures all possible scenarios. Thus, the project will seek to derive AI tools to solve the problem.

The advancement of Artificial Intelligence in the field of Natural Language Processing (NLP) in the past three decades is astonishing. The subfield of linguistics, computer science, information engineering, and artificial intelligence mainly deals with interactions between computers and human (natural) languages, in particular how to program machines to extract, understand and synthesize information from a variety of text sources(“Natural language processing,” n.d.). In recent years, there have been many breakthroughs in the field and it could already perform tasks like name entity recognition, machine translation, and automatic summarization etc. Therefore, our team is convinced that present technologies have reached a level that is capable of solving a significant portion of the issue mentioned above.

From what has been illustrated above, we believe that automating the information extraction process would liberate legal professionals from previously low-end paperwork, and their endeavors could thus be primarily focused on comprehension and synthesis of the extracted information, which could enhance the whole work efficiency a lot. Beyond such, legal practitioners could gain better insight of the full picture, from the full landscape of court case, instead of only a selected group of past rulings.

3.3 Prior Work

Information extraction from court cases is not a field that has received significant recognition, hence, only a limited number of journals is found to be directly related to such topic (Cheng, Cua, Tan, Yao, & Roxas, 2009). Nevertheless, several researchers worked on applying NLP technologies in the realm of law (Chalkidis & Androutsopoulos, 2017; Dragoni, Villata, Rizzi, & Governatori, 2016; Kanapala, Pal, & Pamula, 2017), and these journal articles relate to the topic in various forms, but with different focuses. Aside from researches that are cross-discipline between NLP and law, the research group also reviewed various NLP papers and textbook(Chen, Fisch, Weston, & Bordes, 2017; Clark & Gardner, 2017; Devlin, Chang, Lee, & Toutanova, 2018; Fan et al., 2019; Martin & Jurafsky, 2009), for the reason that most of the project would be based on existing NLP technologies.

3.4 Methodology

3.4.1 Development Principle

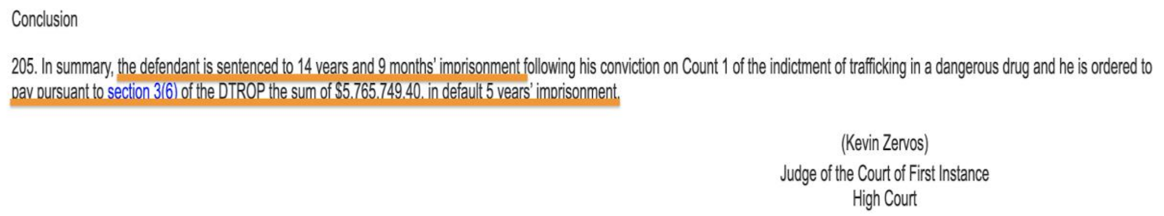
There are two major cornerstones in the development philosophy, namely Test-Driven Development (TDD), and statistical methods. The field of NLP is relatively new in terms of intersecting with legal professionals. Moreover, the nature of the dataset of legal court cases has little documentation provided, more experiments are needed to determine its inherent structure and characteristics. As a result, the research group expects a short development cycle throughout the project, frequent adjustments and experimental developments to adapt to the latest feedback collected during development time. Under such assumption, codebase, models and data structure will be built with the highest degree of flexibility when possible to tackle with uncertainties lying ahead. Also, the lack of knowledge in the legal discipline and the structure, nature of the data basically precludes the option of solely relying on knowledge engineering, in contrary, methodologies are inclined to statistical methods — methods that inference rules and relations out of the dataset.

3.4.2 Architecture Design

The features designated to be extracted will fall under one of the systems, each described in the details of the extraction flow.

<u>Static System</u>

Static system in this project is referring to rule based models incorporated with knowledge engineering to extract features with specific pattern and structure in legal judgements.



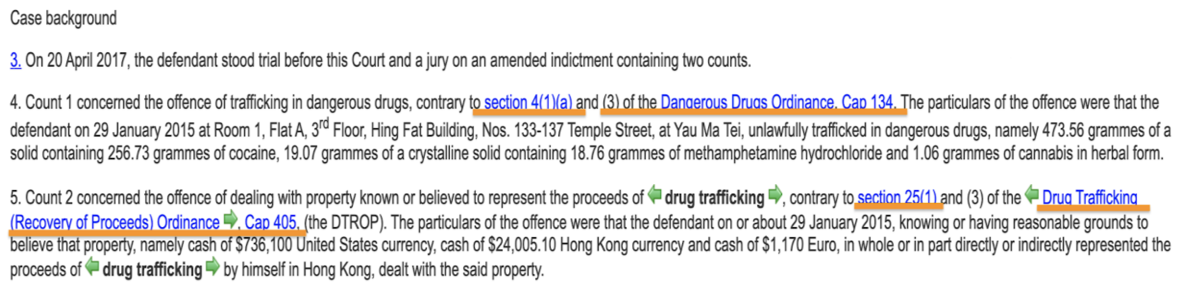
Conclusion

205. In summary, the defendant is sentenced to 14 years and 9 months' imprisonment following his conviction on Count 1 of the indictment of trafficking in a dangerous drug and he is ordered to pay pursuant to section 3(6) of the DTROP the sum of $5,765,749.40, in default 5 years' imprisonment.

(Kevin Zervos)
Judge of the Court of First Instance
High Court

Figure 1: *Conclusion of one Court case containing the charge to the defendant.*

Charge extraction: Charge is a piece of well-structured information (see Figure 1), with a finite domain. Thus, charges could be obtained at a fixed location and could be extracted through keyword matching. Nevertheless, there are some minor variations and noise in the data, for instance, misspellings, numbering. This might hinder the performance of extraction. As a countermeasure, the research group might deploy tokenization, fuzzy search to alleviate such a problem.



Case background

3. On 20 April 2017, the defendant stood trial before this Court and a jury on an amended indictment containing two counts.

4. Count 1 concerned the offence of trafficking in dangerous drugs, contrary to section 4(1)(a) and (3) of the Dangerous Drugs Ordinance, Cap 134. The particulars of the offence were that the defendant on 29 January 2015 at Room 1, Flat A, 3rd Floor, Hing Fat Building, Nos. 133-137 Temple Street, at Yau Ma Tei, unlawfully trafficked in dangerous drugs, namely 473.56 grammes of a solid containing 256.73 grammes of cocaine, 19.07 grammes of a crystalline solid containing 18.76 grammes of methamphetamine hydrochloride and 1.06 grammes of cannabis in herbal form.

5. Count 2 concerned the offence of dealing with property known or believed to represent the proceeds of **drug trafficking**, contrary to section 25(1) and (3) of the Drug Trafficking (Recovery of Proceeds) Ordinance, Cap 405, (the DTROP). The particulars of the offence were that the defendant on or about 29 January 2015, knowing or having reasonable grounds to believe that property, namely cash of $736,100 United States currency, cash of $24,005.10 Hong Kong currency and cash of $1,170 Euro, in whole or in part directly or indirectly represented the proceeds of **drug trafficking** by himself in Hong Kong, dealt with the said property.

Figure 2: *Specific structure of Ordinance reference.*

Ordinance extraction: Ordinance is a more complicated piece of information to extract, for the reason that, ordinances might be embedded in the main corpus. Beyond such, inference from charges might be needed to obtain the related ordinance to court case. According to preliminary observation, ordinances embedded in text follow a specific structure (see Figure 2), hence, there is evidence to believe that keyword matching could be sufficient to parse such information in text. Charges also has a one-to-one relationship with a specific section of ordinance. Completing such a knowledge graph could potentially be solution to extract such a field.

Type of drugs: There is a finite set of domains of type of drugs, hence, keyword matching might be sufficient for obtaining the involvement of type of drugs. Nevertheless, the correspondence of the type of drugs to the defendant might be a multi-to-multi relationship, might require high-level of coreference resolution. The responsibility of extracting such relationship falls upon the dynamic system.

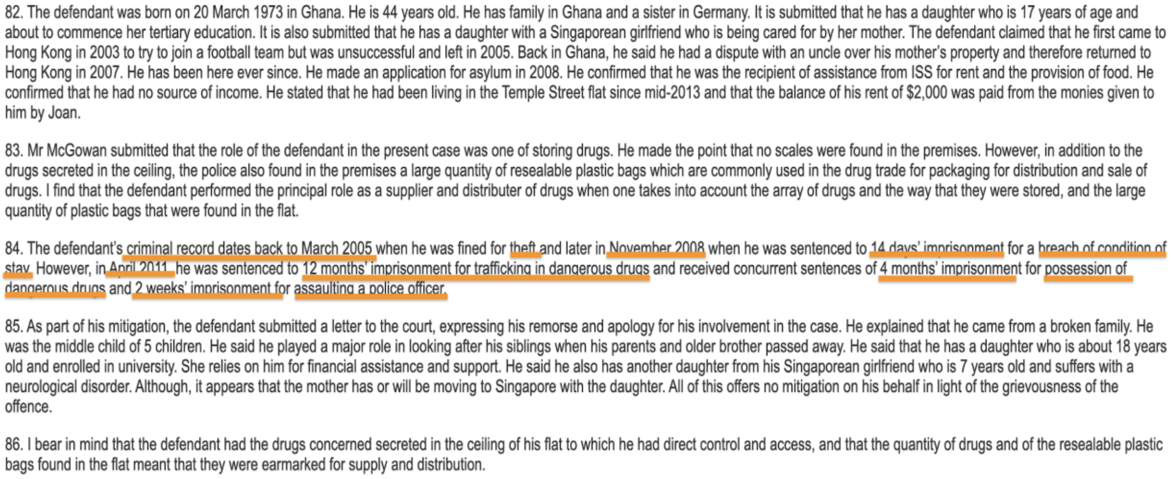
Dynamic System

Dynamic system mainly consists of state-of-the-art NLP models, in theory could circumvent some of the pitfalls of traditional rule based methods and extract information that is not possible with traditional methods.

Amount of drugs: First, Named-entity recognition (NER) will be applied to extract the number of grams appeared in text. Then machine comprehension model will be used to extract the amount of drugs corresponding to the specific amount of drugs. The NER obtained value will serve as a check to ensure the quality of extraction.

Penalty: Machine comprehension model has the potential to extract the penalty imposed when only one defendant in the case. Multi defendant case is still in design.

Features under observation



82. The defendant was born on 20 March 1973 in Ghana. He is 44 years old. He has family in Ghana and a sister in Germany. It is submitted that he has a daughter who is 17 years of age and about to commence her tertiary education. It is also submitted that he has a daughter with a Singaporean girlfriend who is being cared for by her mother. The defendant claimed that he first came to Hong Kong in 2003 to try to join a football team but was unsuccessful and left in 2005. Back in Ghana, he said he had a dispute with an uncle over his mother's property and therefore returned to Hong Kong in 2007. He has been here ever since. He made an application for asylum in 2008. He confirmed that he was the recipient of assistance from ISS for rent and the provision of food. He confirmed that he had no source of income. He stated that he had been living in the Temple Street flat since mid-2013 and that the balance of his rent of $2,000 was paid from the monies given to him by Joan.

83. Mr McGowan submitted that the role of the defendant in the present case was one of storing drugs. He made the point that no scales were found in the premises. However, in addition to the drugs secreted in the ceiling, the police also found in the premises a large quantity of resealable plastic bags which are commonly used in the drug trade for packaging for distribution and sale of drugs. I find that the defendant performed the principal role as a supplier and distributer of drugs when one takes into account the array of drugs and the way that they were stored, and the large quantity of plastic bags that were found in the flat.

84. The defendant's criminal record dates back to March 2005 when he was fined for theft and later in November 2008 when he was sentenced to 14 days' imprisonment for a breach of condition of stay. However, in April 2011, he was sentenced to 12 months' imprisonment for trafficking in dangerous drugs and received concurrent sentences of 4 months' imprisonment for possession of dangerous drugs and 2 weeks' imprisonment for assaulting a police officer.

85. As part of his mitigation, the defendant submitted a letter to the court, expressing his remorse and apology for his involvement in the case. He explained that he came from a broken family. He was the middle child of 5 children. He said he played a major role in looking after his siblings when his parents and older brother passed away. He said that he has a daughter who is about 18 years old and enrolled in university. She relies on him for financial assistance and support. He said he also has another daughter from his Singaporean girlfriend who is 7 years old and suffers with a neurological disorder. Although, it appears that the mother has or will be moving to Singapore with the daughter. All of this offers no mitigation on his behalf in light of the grievousness of the offence.

86. I bear in mind that the defendant had the drugs concerned secreted in the ceiling of his flat to which he had direct control and access, and that the quantity of drugs and of the resealable plastic bags found in the flat meant that they were earmarked for supply and distribution.

Figure 3: *Defendant background embedded. Criminal records highlighted.*

Mitigating and aggravating factors are deeply embedded in the context of court case (see Figure 3). The research group has little confidence in this stage to extract such features with high accuracy and recall. The main hurdle lies in the incomplete ontologies of such factors. Nevertheless, some results could be foreseen to be obtained through keyword matching and machine comprehension. The performance thought is still uncertain in each case.

Data Storage

There is no de facto standard of data storage solution in the field of NLP. Researchers adopt various measures to manage data, including plain text, csv, json, etc. One of the most advanced methods would be protocol buffers (protobuf)—— a method which serializes structured data. In comparison, the research team has adopted with non-conventional practices — MySQL and in consideration of Redis.

The mainstream research direction in computational linguistic community revolves in model training, hence, batch form, simple data storage solution will suit the needs of most projects. Nevertheless, this project has greater emphasizes on data analytics. Recognize patterns in the data set in order for setting up strategy for extraction. With different focuses, static files that fit the need of mainstream researchers have less value in our circumstance. Protocol buffers has the highest performance among all solution and is the only mechanism that support genuine mass scale parallel model training. One of its weaknesses is the level of difficulty, the effort of protobuf is enormous. PhDs encounter tremendous difficulty according to one of the researchers at the University of Maryland, College Park — Jordan Boyd-Graber. As a result, the research group believes that the costs outweigh the benefit of the installation of protobuf and static files alone could not meet the demand of this project.

The requirements of data solution of this project are a flexible data scheme that could adapt to changes, analysis is an essential part, hence, the capability of query is also necessary, beyond doubt, decent performance is also crucial. MySQL is one of the most widely test solutions existing, the capability of SQL query enables the research team to integrate it with existing analytics tool, reducing the cost of understanding our data. One of the main drawbacks of MySQL, is the lower speed of retrieval when large amount of text or blob is involved. Redis is a key-value store database, with speed faster in order of magnitude even with large volume of data. Deploying both MySQL and Redis, allows the research team to perform analysis with a high degree of freedom, and in the meantime, train project models with selected data dynamically. Hence, this states the factors considered and principle applied in the process of data storage selection.

4 Work Plan

4.1 Limitations, Assumptions, and Alternatives

The assumptions for the project rely on the belief that current technologies have reached a level of adequacy to identify and extract the features interested.

However, there still exist risks that the court case documentation is too messy and intricated to be analyzed accurately and it might be too difficult to correctly decipher the correlation between desired items, for example, the correct match of drug name and weight, which would result in imprecise information extraction.

Possible alternatives for such complication would be extracting a whole sentence (context) of where the wanted information lies, or involving human tagging and extraction during the process.

Current proposal is still at a preliminary stage, and advanced feature extraction is still under design. Multi-to-multi relationship extraction resolution is in research stage without a blueprint of the extraction workflow. Aggravating factors and Mitigating Factors extraction are still under study, primarily comprehending the legal court case data. The research team is searching for pattern that could consistently works as clues for models to extract such aforementioned features. The proposal will be updated in accordance to any progress in experiment or theoretical study.

4.2 Preliminary Schedule

Milestone	Duration(weeks)	Description	Expected Completion Date

First Deliverable	2	Project plan and website.	30 September 2019
Preliminary Study	4	Literature review.	October 2019
Set up	4	System and environment setup.	December 2019
First Presentation	4	Interim report.	13-17 January 2020
Phase 2 deliverables	4	Interim report.	2 February 2020
Implementation	8	Implementation.	February - April
Phase 3 deliverables	3	Final report.	19 April 2020
Final presentation	1	Finish.	20-24 April 2020

References

Chalkidis, I., & Androutsopoulos, I. (2017, December). A Deep Learning Approach to Contract Element Extraction. In *JURIX* (pp. 155-164).

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Cheng, T. T., Cua, J. L., Tan, M. D., Yao, K. G., & Roxas, R. E. (2009, October). Information extraction from legal documents. In *2009 Eighth International Symposium on Natural Language Processing* (pp. 157-162). IEEE.

Clark, C., & Gardner, M. (2017). Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dragoni, M., Villata, S., Rizzi, W., & Governatori, G. (2016, December). *Combining NLP Approaches for Rule Extraction from Legal Documents*.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., & Auli, M. (2019). Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Kanapala, A., Pal, S., & Pamula, R. (2017). Text summarization from legal documents: a survey. *Artificial Intelligence Review*, *51(3)*, 371-402.

Martin, J. H., & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Pearson/Prentice Hall

Wikipedia contributors. (2019, September 29). Common law. In *Wikipedia, The Free Encyclopedia*. Retrieved 13:15, September 29, 2019, from https://en.wikipedia.org/w/index.php?title=Common_law&oldid=918517905

Wikipedia contributors. (2019, September 17). Law of Hong Kong. In *Wikipedia, The Free Encyclopedia*. Retrieved 13:13, September 29, 2019, from https://en.wikipedia.org/w/index.php?title=Law_of_Hong_Kong&oldid=916247549

Wikipedia contributors. (2019, September 15). Natural language processing. In *Wikipedia, The Free Encyclopedia*. Retrieved 13:16, September 29, 2019, from https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid=915734112