Proposed Tasks

The main target of this paperwork analysis project is 40,000 judgments, mainly related to civil and criminal cases such as private lending, sales and purchase contract disputes, theft and robbery.

Data Source: https://cuhko365-my.sharepoint.com/:f:/g/personal/221022026_link_cuhk_edu_cn/EuxcB5v_rEtHqefgNvGk45UBoc6hM-OkRZ65dkge0N58Pw?e=jjDOei

[ATTENTION] As this dataset is a record of real court judgments and involves a great deal of personal privacy, it is important that you **do not disclose these legal documents to the public** and that you do not mention the specific contents of the documents (especially highly private information such as the names of the parties) to others in public, as this could lead to unnecessary legal problems. If found we may have to impose appropriate penalties ! ! !

Task 1: Information Extraction and Mining
- Case **type** extraction (criminal or civil case)
- Extraction of the **cause** of the case (civil lending dispute, sale and purchase contract dispute, theft, robbery)
- Date of Judgment
- Basis of judgment (e.g. Contract Law of the People's Republic of China; Article 60(1); Article 60(2).....)
- Any other information you think is valuable (province/region, win/loss of the case, whether a lawyer intervened, education level of the parties .....)

Task 2: Clustering Problems
Cluster analysis using classical text clustering methods such as *k*-means and DBSCAN
- Clustering the cause and the type of case.
- Clustering the key information involved in these cases, possibly in the form of a **word cloud.**

Task 3: Text Similarity
Set your *student id* as random seed, sample 1000 cases from 40,000 judgments.
- Following task 1, regard the corresponding type/cause of your sampled cases as natural labels.
- After proper vectorization of the text, choose cosine similarity and Jaccard similarity as measurements of text similarities. Calculate and use proper methods to visualize the similarities among your sampled cases.
- Under the reference of natural labels, try to find out which two cases are the nearest, which two are the least similar? And does there exist more patterns?

You can extract the required information from the JSON file into a data format you are familiar with, and try to do the following task which is actually not too technical but require some analytical thinking skills.

Task 4: Anomaly Analysis
- Split out the case types of contract cases and private lending cases and extract the amount of the underlying case. You can analyse the amounts and try to find some anomalies and analyse the reasons for them.
- Cases where the case type is private lending are split out and the interest rate of the case is extracted. To distinguish them from anomalies, use the following flags: -1 is the default value of the PBC concurrent rate; -2 is the PBC concurrent rate for a period of less than 1 year; -3 is the PBC concurrent rate for a period of less than 5 years; -4 is the PBC concurrent rate for a period of more than 5 years. Interest rates can be analysed to find some unusual rates and try to analyse the reasons for them.