裁判文书金额实体提取软件 V1.0

- 设计说明书

版本	1.0
文档状态:	完成
作者:	刘奕鑫
负责人:	刘奕鑫
创建日期:	2021 年 4 月 02 日
更新日期:	2021 年 4 月 02 日

目录

1 简介

1.1 开发目的

实际法律业务需求中，通常需要将裁判决文书中的金额进行结构化的提取，金额类别有比较多种，通常会包括：贷款本金、本金余额、合计利息、利息余额、罚息余额、复利余额、保证金额、抵押物最高抵押金额。这部分工作目前大多数都是由人工完成，费时费力。本项目结合自然语言处理和人工智能技术，开发了裁判文书金额实体提取软件。

1.2 产品范围

本软件主要专注于为法律领域裁判文书的金额字段提取，核心功能是**自动化提取文本中的十类金额实体**，包括贷款本金、本金余额、合计利息、利息余额、本息合计、罚息余额、复利余额、保证金额、抵押物最高抵押金额、其他金额，辅助功能是**提取并匹配金额相关的三类字段**：利息金额截止日、保证金额对应人、抵押金额对应抵押物。

1.3 模块描述

软件主要包括如下模块：

1）判决段落提取模块：

负责裁判文书中判决文本的定位和提取，根据关键词进行关键段落的定位和提取。

2）金额实体提取与分类模块：

负责提取与分类金额实体，包括一个判定金额存在性的分类模型、提取金额实体的正则提取器、分类金额实体类别的分类模型。

3）金额关联相关实体提取与匹配模块：

负责提取金额相关的三类实体，并根据最近原则进行匹配。

4）Flask 后端模块：

负责响应软件外部的 POST 请求，调用内部金额提取算法。

2 设计说明

2.1 总体流程

软件总体工作流程概述如图 2-1 所示，主要分为三部分：提取判决段落、提取并分类 10 类金额实体、提取并匹配三类关联实体。

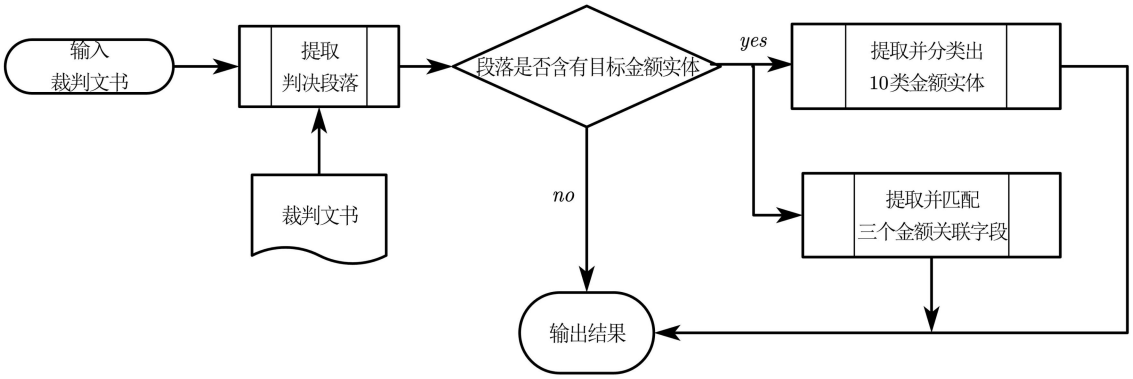

图 2-1 总体流程图

软件不含交互界面，通过 API 请求进行调用。输入为一篇裁判文书的文本，格式为纯文本。下面针对三部分进行详述。

提取判决段落

由于目标金额实体一般都在判决段落，所以在软件得到输入的裁判文书后，会调用**判决段落提取模块**，完成判决关键段落的提取。这部分的提取流程见流程图 2-2 所示。首先，程序根据前缀关键词定位关键段落的开头，然后再根据后缀关键词定位结尾，最后提取得到关键的段落。具体的定位关键词见表 2-1 所示。在完成判决文本段落的提取后，我们会得裁判文书中的判决段落，这部分含有我们需要提取的金额实体。

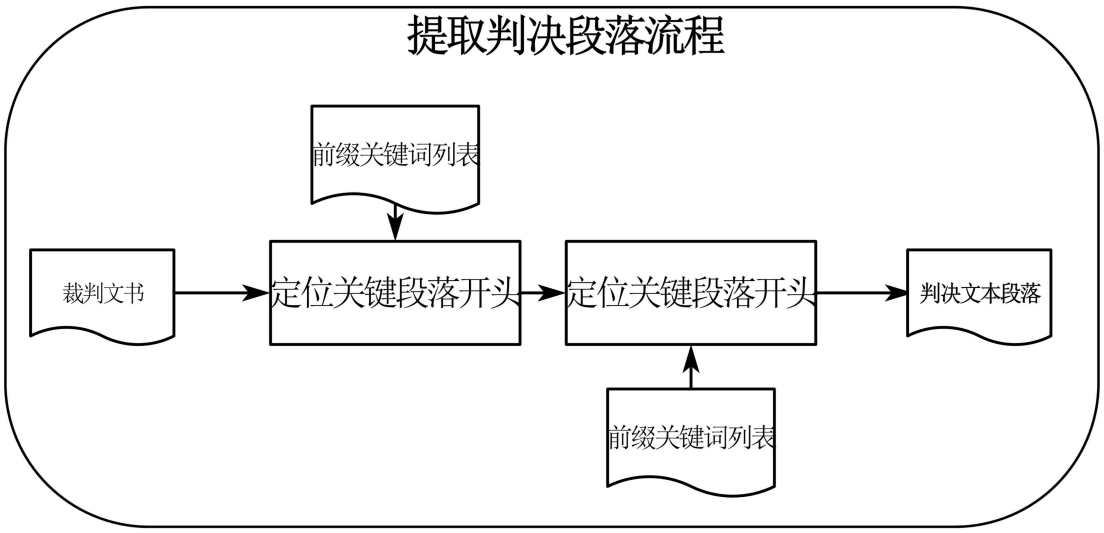

图 2-2 判决段落提取流程

定位类型	定位关键词
判决段落前缀	['裁定如下','判决如下','裁判主文','特发出如下支付令','如下调解协议','本院分析如下','判决结果']

判决段落后缀	['审判长','审判员','如不服','受理费','书记员']

表 2-1 判决段落前后缀的关键词匹配列表

提取并分类金额实体

在得到判决段落后，软件首先调用金额存在性分类模型。该模型已经在标注数据上训练完成，输入判决文本段落后，能够分类出其是否存在目标的金额实体。在判定文本包含目标金额实体后，模块继续调用正则提取函数来提取其中包含的所有金额实体。对于每个金额实体，我们通过其前后文构建其分类特征，并且将该特征输入已经训练好的金额十分类模型，我们可以得到该实体的所属金额类别。我们在图 2-3 总结了这部分的流程。这部分结束后，会得到如图 2-4 所示的金额提取结果。

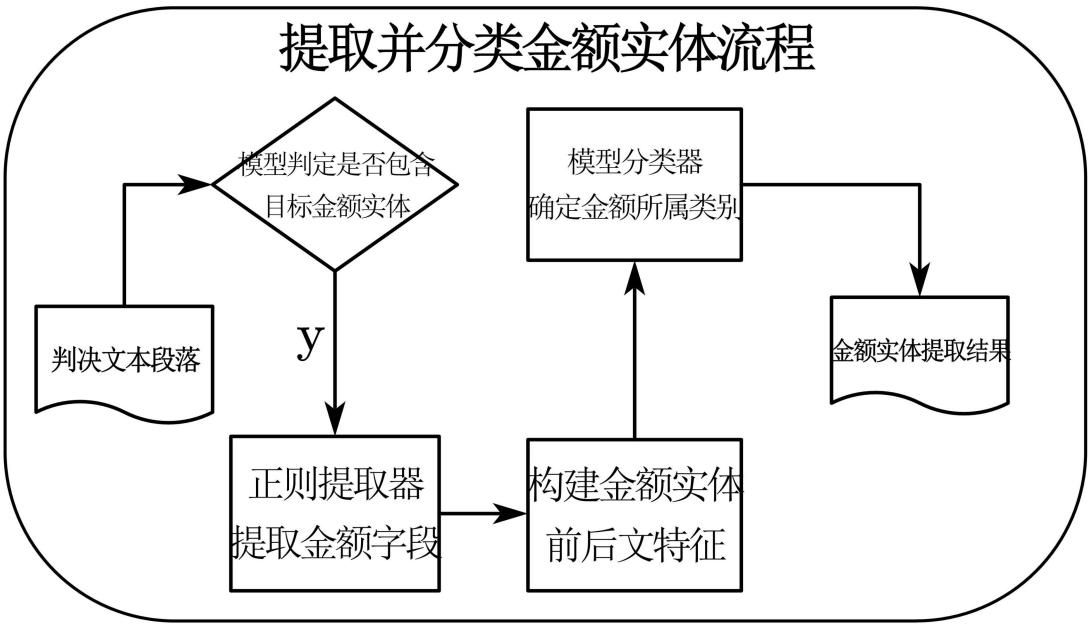

图 2-3 提取并分类 10 类金额实体流程


```
{"提取金额": {
    "复利余额": ["1831.2元"],
    "贷款金额": ["200000.0元"],
    "利息余额": ["20033.34元"],
    "其他金额": ["20033.34元"],
    "罚息金额": ["18449.89元"],
    "本金余额": ["199998.91元"]
    }
}
```


图 2-4 提取并分类 10 类金额实体结果示例(部分)

提取并匹配三个金额关联字段

对于保证金额、抵押物最高金额、利息余额这三个金额字段，我们的软件提取并关联了其相关的三个字段：保证人、抵押物、利息截止日期。首先，我们根据关键词定位到实体所在的句子，将句子作为提取相关字段模块的输入。我们采用百度的 Lexical Analysis of Chinese（LAC）工具提取保证人，实体类别包括了公司和人名；针对抵押物的提取，我们采用了自主训练的基于 Bert 的地址标注模型进行提取；针对利息截止日期，我们先同样采用 LAC 模型提取日期字段，然后根据关键词筛选出符合要求的日期实体。在提取完三类相关字段后，我们采用“最近关联”的规则，来完成其与金额实体的关联。我们在图 2-5 总结了这部分的提取流程，流程执行完可以得到如图 2-6 的提取匹配结果。

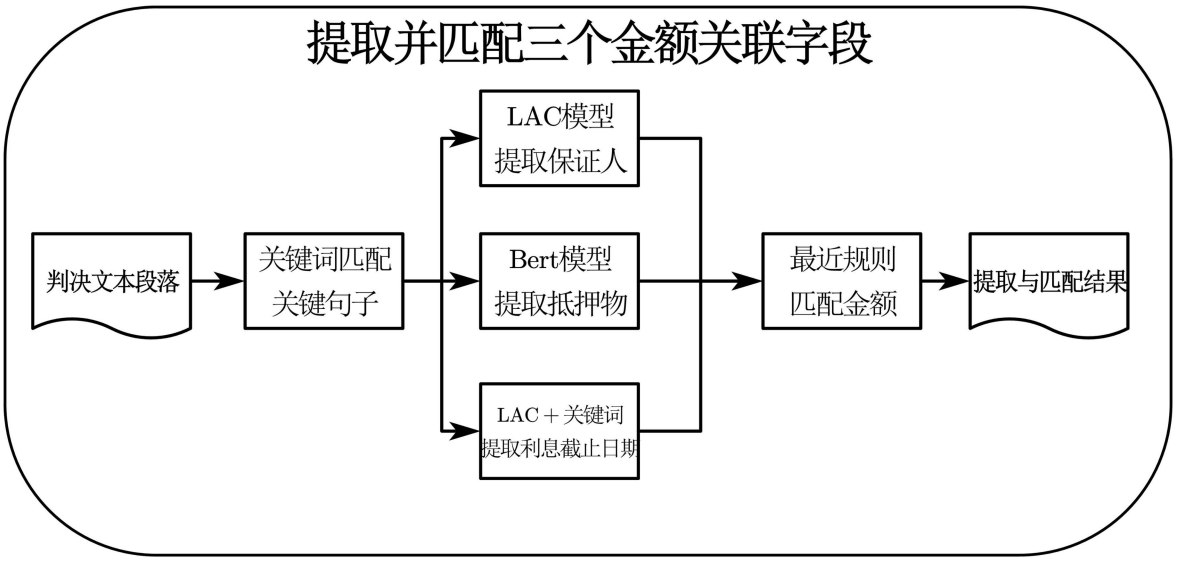

图 2-5 提取并匹配三个金额关联字段流程图

```
{"利息截止日期": ["2015年6月20日"],
"保证人-保证金额": [{
                "保证人": [
                    "XXX公司"
                ],
                "最大保证金额": "22230000元"
            },
            { "保证人": [
                    "XXX"
                ],
                "最大保证金额": "223300000元"
            }],
        "抵押物-最大抵押金额": [{
                "最大抵押金额": "22230000元",
                "抵押物": [
                    "XXXXXXXXXXXXXXXXXXXXXX",
                    "XXXXXXXXXXXXXXXXXXXXXX",
                    "XXXXXXXXXXXXXXXXXXXXXX"
                ]}]
}
```

图 2-6 金额关联字段提取匹配结果示例

将两部分提取的结果进行拼接，就得到了我们软件最终的输出。至此，软件完成了从输入的裁判文书到十类金额实体、三类金额关联实体的提取。

2.2 模型原理

在上一节中，我们对软件工作的流程进行了详细的叙述，并未对其中涉及到的一些模型的设计原理进行更进一步的说明。本节，我们对软件涉及的几个重要模型进行原理上的说明：**基于 TF-IDF 和朴素贝叶斯的金额存在性判定模型、基于正则表达式的金额实体提取器、基于 Bert 与逻辑回归的金额十分类模型、基于 BERT 和双指针的抵押物字段提取模型**。

基于 TF-IDF 和多项式朴素贝叶斯的金额存在性判定模型

由于裁判文书的种类繁多，输入文书不一定含有我们要提取的金额实体，所以为了提高处理效率，我们引入了一个预分类的模型，来判定文书中是否包含我们要提取的对象。模型基于 TF-IDF 与多项式朴素贝叶斯模型，用 TF-IDF 特征来表征文档的每个词的数值特征，使用多项式朴素贝叶斯来进行类别的归类预测。

给定文档$d \in D$，与其包含的所有词$w_i \in d$，基于朴素假设我们可以计算文档属于类别c_j的未归一化概率如下：

$$P(c_j|d) = \frac{P(d|c_j)P(c_j)}{P(d)} = \frac{P(c_j)\prod_i P(w_i|c_j)}{P(d)}$$

其中，$P(c)$表示类别 c 下的文档总数与所有类别下的文档总数的比值，$P(w_i|c)$表示单词w_i的文档在类别 c 中出现的次数+1 与类别 c 下的文档总数+C 的比值，其中 C 为类别总数，$P(d)$为文档出现的频率。由于$P(d)$对所有类别均一致，我们可以得到以下的归一化概率：

$$P(c_j|d) = \frac{P(d|c_j)P(c_j)}{\sum_k P(d|c_k)P(c_k)} = \frac{P(c_j)\prod_i P(w_i|c_j)}{\sum_k P(c_k)\prod_i P(w_i|c_k)}$$

在文档d上，其每个词w_i的统计数值特征$P(w_i|d)$由 TF-IDF 方法进行计算：

$$P(w_i|d) = TF \times IDF = count(w_i,d) * \left(1+\ln\left(\frac{1+D}{1+|d \in D: w_i \in d|}\right)\right)$$

其中，$count(w_i,d)$表示词w_i在文档d出现的频率，$|d \in D: w_i \in d|$为文档集合D中出现了词w_i的文档个数。

给定标注数据后，由极大似然估计的方法，我们可以得到每个类别c_j下词w_i的概率特征$P(w_i|c_j)$：

$$P(w_i|c_j) = \frac{\sum_d P(w_i|d)\mathbb{I}(d \in c_j)}{\sum_k \sum_d P(w_i|d)\mathbb{I}(d \in c_k)}$$

在推理阶段，我们根据以下的最优化指标选取预测类别c：

$$c = \underset{c_j}{argmax}(P(c_j)\prod_i P(w_i|c_j))$$

基于正则表达式的金额实体提取器

该部分，我们提取了两类金额实体，一部分是数字的金额，另一部分的中文的金额。提取的正则表达式如表 2-2 所示

表 2-2 金额提取的正则表达式

提取字段类型	正则表达式
数字金额	(([1-9]\\d*[\\d,,]*\\.?\\d*)|(0\\.[0-9]+))(元|百万|万元|亿元|万|亿|人民币|美元|美金)
中文金额	(一|二|三|四|五|六|七|八|九|十)+([一|二|三|四|五|六|七|八|九|十|百|千|万|亿]+)(元|美金|人民币)

基于 Bert 与逻辑回归的金额十分类模型

Bidirectional Encoder Representations from Transformers，简称 BERT，是近期非常热门的一个自然语言处理领域的预训练模型，简洁架构却有着非常突出的效果，受到了工业界和学术界的宠爱。在我们的金额实体分类中，也运用到了该模型来词特征提取。我们运用的 BERT 模型由 12 层的 transformers 层构成，输入为句子的词编号特征、掩码特征、位置编码特征，输出为嵌入维度 312 的词向量。示例的输入输出如图 2-7 所示：

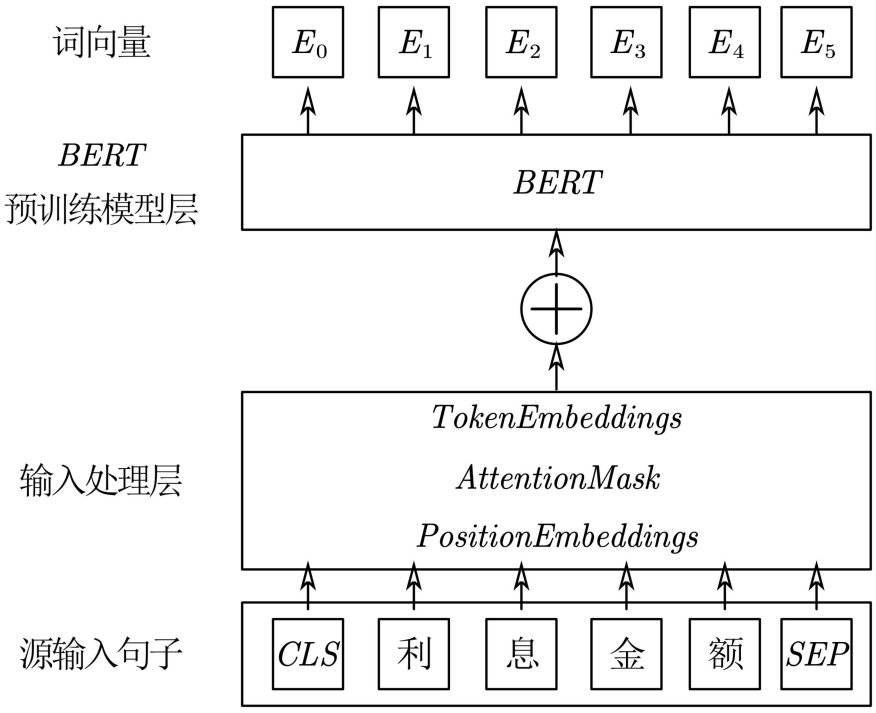

图 2-7 BERT 模型输入输出说明示意

对于段落中根据正则提取器得到的金额实体$s_j, j \in \{1,2,\cdots,|s|\}$，我们需要将他们分为了实体类别$y_i, i \in \{1,2,\cdots,K\}$，也就是拟合出一个映射$f$，满足映射关系$y \leftarrow f(s) \in R^{|s|\times K}$。

对于一个金额实体s_j，其周围的词对分类其实体类别是非常关键的。记$E(\cdot)$为某词经过$BERT$模型提取后的词向量，我们可以抽取该金额实体s_j所在位置左右α个词的词向量特征表示作为分类的特征：

$$X = [E(x_1), E(x_2), \cdots, E(x_{2\alpha})]$$

在得到了实体s_j 的特征表示 后，我们可以基于逻辑回归（Softmax Regression）来学习其类别归属概率 $P(y = i|s_j; W)$，相关参数W的拟合采用梯度下降法进行求解。

$$P(y = i|s_j; W) = \frac{e^{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}}}{\sum_{l=0}^{K-1} e^{\mathbf{w}_l^{\mathrm{T}}\mathbf{x}}}$$

基于 BERT 和双指针的抵押物字段提取模型

抵押物的提取部分，我们首先通过关键词规则定位到抵押物所处的句子，然后将这些句子输入基于 BERT+双指针的模型中，模型由 BERT+LAYERNORM+FC 构成，输出是句子每个位置的分类类别：是否为抵押物开头、是否为抵押物结尾。根据头尾连接关系，得到最终预测的抵押物字段。训练数据通过标注的 3k 份抵押物和十套抵押物句子模板构造而成。

2.3 接口说明

软件采用 API 调用（POST）的方式与外界进行交互，采用 FLASK 框架搭建了后端接口。接口的输入和输出说明如表 2-3、表 2-4 所示。

1）输入字段

表 2-3 输入字段说明

序号	字段名	字段类型	字段描述
1	Text	字符串	裁判文书全文，纯文本

2）输出字段

表 2-4 输出字段说明

序号	字段名	字段类型	字段描述
1	MSG	Int	调用状态码：0 表示调用失败；1 表示调用成功。
2	Content	Dict	提取的结果
3	amount	Dict	提取的十类金额
4	compound_interest_balance	List[str]	复利余额
5	debt_amount	List[str]	贷款本金
6	interest	List[str]	利息余额
7	maximum_guaranteed_amount	List[str]	最大保证金额
8	maximum_mortgage_amount	List[str]	抵押物最高抵押金额
9	other_amount	List[str]	其他金额
10	payment_balance	List[str]	罚息余额
11	principal_balance	List[str]	本金余额
12	sum_insterest_debt	List[str]	本息合计
13	guarantor_maximum_guaranteed_amount	Dict	保证人-保证金额的关系对
14	guarantor	List[str]	保证人
15	maximum_guaranteed_amount	str	保证金额
16	mortgage_maximum_mortgage_amount	Dict	抵押物-抵押物最高抵押金额的关系对
17	maximum_mortgage_amount	str	抵押物最高抵押金额
18	mortgage	List[str]	抵押物