民事司法摘要生成

# #1.摘要模型测评

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Unigram | Bigram | LargestCommon | Weighted |
| Recall | 0.26063 | 0.12974 | 0.25338 |  |
| Precision | 0.58055 | 0.25922 | 0.57089 |  |
| F1 | 0.32617 | 0.15634 | 0.31789 | 0.254926 |

# #2.需求说明

目前司法大数据标注的数据共4047+9484=13531条，包含九种案由：‘劳动合同’, ‘侵权责任’, ‘租赁合同’, ‘借款合同’, ‘继承’, ‘追偿权’, ‘借款’, ‘侵权’, ‘继承关系’；对应法宝案由为：租赁合同、借款合同、劳动合同、继承、侵权责任、追偿权共计六种。

# #3.摘要模型介绍

裁判文书是人民法院公开审判活动、裁判理由、裁判依据和裁判结果的重要载体。司法摘要则是对裁判文书的内容进行压缩、归纳和总结，反映案件审理过程中的裁判过程、事实、理由和判决依据等。裁判文书司法摘要对我国法治建设具有现实意义与必要性。CAIL2020设计了从裁判文书中提炼摘要的任务。

第一阶段从5月15日开始到7月14日，开放第一阶段数据集；从7月15日开始，开放二阶段数据集。第一阶段有4047条标注数据，第二阶段有9484条。

摘要模型分为三个组件，分别为段落提取、关键句提取和关键词提取。段落提取部分利用了文书中标志性词汇来定位不同段落的开始与结尾，例如原告段落常有“原告诉称”、“原告提出诉讼请求”和“提出诉讼”，或者“驳回原告提出诉讼请求”等规则文本；而原告结束段落常跟随者“被告辩称”、“被告答辩”以及“未做答辩”等规则文本。段落共有原告、查明、法院、判决四部分。

第二部分，关键句提取没有现成标注数据可用，因此，利用标注的摘要文本，根据标点符号分割成多个摘要句，利用这些摘要句在裁判文书全文句中查找BM25值最高的语句，标记为相似句，而全文中其他语句则标记为无关句。通过统计相似句与无关句的数量，两者数量比例为1:5。避免不平衡语料对训练模型的影响，因此对量少的语料复制4倍，进行了数据平衡处理。模型结构为首先利用RoBERTa-wwm-ext-Chinese[1]对句子进行编码，将最后两层隐含层合并后取最大值，再经遗忘层连接全连接层和Softmax层，模型的。关键句模型训练集含正负样本782895条，验证集42条，模型训练收敛后在验证集上的准确率为92.06%。

第三部分，关键词提取组件，同样地，需要自己构建标注数据，首先构建了全文关键句与摘要对，共13537条记录。本文对关键词的定义为经过BERT分词后，出现在摘要文本中，未出现在摘要中的词定义为可忽略。在预处理过程中，对关键句按照512为块进行拆分，不足512\**80%的关键句进行舍弃，对于关键句长度小于512\**80%的同样舍弃，对每个关键句块按照512进行填零补齐。关键词提取模型的输出层为512。模型结构为将BERT编码后的句子的后三层和第一层合并，后接入CapsuleNet中，最后接入全连接层。验证集有16条，训练10轮，学习率1E-3，梯度累计步数8，胶囊维度16，压缩胶囊数为128，开启胶囊网络的自适应动态路由，批大小为2。训练模型的验证集精度为60.53%。

模型集成介绍，首先输入的裁判文书经过断句生成数十个句子，经过段落提取后获得关键段落，再经过摘要关键句模型，过滤得到关键段落的关键句，最后经过关键词组件对关键句中的词进行过滤打分，保留高分词得到生成摘要。在利用关键词组件时，由于模型预测字重要性结果很少可以超过0.5，因此采用了TOPK方法对各个词的评分进行了过滤，在本文中选择了87.5%作为K参数，即选择句中评分最高的前87.5%个词作为关键词，组成摘要。

模型评价，本任务采用ROUGE(Recall-Oriented Understudy for Gisting Evaluation)评价评价。ROUGE指标将自动生成的摘要与参考摘要进行比较, 其中ROUGE-1衡量unigram匹配情况，ROUGE-2衡量bigram匹配，ROUGE-L记录最长的公共子序列。三者都只采用f-score，且总分的计算方式为：0.2*f-score(R1)+0.4*f-score(R2)+0.4\*f-score(RL)。本文模型加权F1评分为25.49%。详细评分如图1所示：

# #4.接口介绍

## ##4.1请求

### ###4.1.1基本方式（falcon服务框架可能会url长度限制：1000字）

请求方式：HTTP GET方式： 参数说明：1传gid，2传民事裁判文书全文

示例： http://192.168.60.15:58082/z?1=guid&2=徐绍勤、徐艳华等与徐少清继承纠纷一审民事判决书事 判 决 书（2017）辽0782民初2218号：徐绍勤，男，1951年7月12日生，满族，农民，住辽宁省北镇市。：张凤林，系辽宁名崛律师事务所律师。:徐艳华,女,1959年3月8日生,满族，农民，住辽宁省北镇市。:徐艳秋,女,1967年8月8日生,满族，农民，住辽宁省北镇市。：徐少清，男，1956年4月4日生，满族，职工，住辽宁省北镇市。：徐超(被告之子),住辽宁省北镇市北镇街道办事处正兰旗胡同19号。：钱常惠，系北镇市北镇法律服务所法律工作者。、徐艳华、徐艳秋与被告徐少清继承纠纷一案，本院于2017年9月14日立案受理后，依法适用简易程序，公开开庭进行了审理。、被告及委托诉讼代理人均到庭参加了诉讼。。：1.请求继承父、母遗产房屋两间、存款4万元及利息，分割父亲的丧葬费、丧葬补助金137007元；2.要求被告承担诉讼费。：原、被告均系徐世国、朱凌云的子女。（北镇市城建局退休）、母朱凌云在北镇市万紫山社区两间民房居住。2014年11月去世（未对母亲遗产进行继承），父亲由原、被告轮流照顾生活起居，2017年4月去世。、母安葬费用均用父、母积蓄。、母留下的两间平房（现已动迁，被告得一户住宅楼，富源城堡小区），银行存款4万元，均由被告占有。、丧葬补助金137007元也由被告支取。、徐艳秋（依法追加）称，父亲的遗嘱是真实的，听从法院判决。，答辩人的母亲于2013年11月去世。，父亲由答辩人一人照顾生活起居至父亲去世，父亲去世时间是2017年5月。，答辩人的父亲就将母亲的遗产（包括两间房屋的一半）进行了分割，给原、被告四人每人5000元。，现已动迁，答辩人的父亲遗嘱将两间平房给了答辩人。4万元及利息780元和丧抚费137007元确经答辩人手取出，但这些钱答辩人父亲在世的时候，因为答辩人一人独自照顾父亲的生活起居，答辩人的父亲亲口交待，除去用于丧葬之外剩余的钱,均归答辩人所有。：原、被告均系被继承人徐世国、朱凌云的子女。，现已动迁，动迁后置换房屋还没有进行建设。2013年11月去世，当时原、被告父亲(被继承人徐世国)分给原、被告四人每人人民币5000元。，被继承人徐世国于2017年5月去世，被继承人徐世国去世后，由被告进行了发丧。：”房子如有动迁，房子享楼由我二儿子由徐绍青继承。。.24”。，其银行存款4万元及利息780元和徐世国的丧葬费、丧葬补助金137007元由被告支取。，原、被告均系二被继承人的子女，故依法对被继承人的遗产均享有继承权。，原告徐艳华、徐艳秋对其真实性予以确认，原告徐绍勤虽对该遗嘱真实性持否定态度，但未在本院限定时间内预交鉴定费用，故本院对该份遗嘱的真实性予以确认。，被继承人徐世国虽分给原、被告四人每人人民币5000元，但不能说明该笔钱是对被继承人朱凌云遗产（两间房屋的一半）的处分。，但其处分被继承人朱凌云遗产（两间房屋的一半）是无效的。（两间房屋的一半）应按法定继承处理。，置换房屋还未建设，暂无法确认具体继承数额，可待置换房屋建成后另案处理。4万元存款及780元利息，被告没有证据证明被继承人生前答应由其继承该款，故应按法定继承处理。137007元丧抚费，不是被继承人的遗产。，被告对其进行了发丧，故其中的丧葬费28574元，应归被告所有。108433元，系抚恤金。。，故在遗产继承及抚恤金分割时应多分。，依照《中华人民共和国继承法》第三条、第五条、第十条一款、第十三条三款及《中华人民共和国民事诉讼法》第一百四十二条之规定，判决如下：、由三原告各继承被继承人徐世国存款及利息40780元的15%，即6117元，由被告继承55%，即22429元；、徐世国的抚恤金108433元，由三原告各分得15%，即16265元，由被告分得55%，即59638元；、徐世国的丧葬费28574元归被告所有；、二项三原告应得数额，由被告于本判决生效后5日内给付。，应当依照《中华人民共和国民事诉讼法》第二百五十三条的规定，加倍支付迟延履行期间的债务利息。4300元，减半收取2150元，由三原告各负担322元,由被告徐少清负担1184元。，可在判决书送达之日起15日内，向本院递交上诉状，并按对方当事人的人数提出副本，上诉于辽宁省锦州市中级人民法院。　　侯百吉〇一七年十一月九日　　董　健

### ###4.1.2建议方式

请求方式：HTTP POST方式：

参数说明： POST /z HTTP/1.1 Content-Length: 6178 Host: 192.168.4.250:58082 Content-Type: application/json

数据： { “1”:“guid” , “2”:“徐绍勤、徐艳华等与徐少清继承纠纷一审民事判决书事 判 决 书（2017）辽0782民初2218号：徐绍勤，男，1951年7月12日生，满族，农民，住辽宁省北镇市。：张凤林，系辽宁名崛律师事务所律师。:徐艳华,女,1959年3月8日生,满族，农民，住辽宁省北镇市。:徐艳秋,女,1967年8月8日生,满族，农民，住辽宁省北镇市。：徐少清，男，1956年4月4日生，满族，职工，住辽宁省北镇市。：徐超(被告之子),住辽宁省北镇市北镇街道办事处正兰旗胡同19号。：钱常惠，系北镇市北镇法律服务所法律工作者。、徐艳华、徐艳秋与被告徐少清继承纠纷一案，本院于2017年9月14日立案受理后，依法适用简易程序，公开开庭进行了审理。、被告及委托诉讼代理人均到庭参加了诉讼。。：1.请求继承父、母遗产房屋两间、存款4万元及利息，分割父亲的丧葬费、丧葬补助金137007元；2.要求被告承担诉讼费。：原、被告均系徐世国、朱凌云的子女。（北镇市城建局退休）、母朱凌云在北镇市万紫山社区两间民房居住。2014年11月去世（未对母亲遗产进行继承），父亲由原、被告轮流照顾生活起居，2017年4月去世。、母安葬费用均用父、母积蓄。、母留下的两间平房（现已动迁，被告得一户住宅楼，富源城堡小区），银行存款4万元，均由被告占有。、丧葬补助金137007元也由被告支取。、徐艳秋（依法追加）称，父亲的遗嘱是真实的，听从法院判决。，答辩人的母亲于2013年11月去世。，父亲由答辩人一人照顾生活起居至父亲去世，父亲去世时间是2017年5月。，答辩人的父亲就将母亲的遗产（包括两间房屋的一半）进行了分割，给原、被告四人每人5000元。，现已动迁，答辩人的父亲遗嘱将两间平房给了答辩人。4万元及利息780元和丧抚费137007元确经答辩人手取出，但这些钱答辩人父亲在世的时候，因为答辩人一人独自照顾父亲的生活起居，答辩人的父亲亲口交待，除去用于丧葬之外剩余的钱,均归答辩人所有。：原、被告均系被继承人徐世国、朱凌云的子女。，现已动迁，动迁后置换房屋还没有进行建设。2013年11月去世，当时原、被告父亲(被继承人徐世国)分给原、被告四人每人人民币5000元。，被继承人徐世国于2017年5月去世，被继承人徐世国去世后，由被告进行了发丧。：”房子如有动迁，房子享楼由我二儿子由徐绍青继承。。.24”。，其银行存款4万元及利息780元和徐世国的丧葬费、丧葬补助金137007元由被告支取。，原、被告均系二被继承人的子女，故依法对被继承人的遗产均享有继承权。，原告徐艳华、徐艳秋对其真实性予以确认，原告徐绍勤虽对该遗嘱真实性持否定态度，但未在本院限定时间内预交鉴定费用，故本院对该份遗嘱的真实性予以确认。，被继承人徐世国虽分给原、被告四人每人人民币5000元，但不能说明该笔钱是对被继承人朱凌云遗产（两间房屋的一半）的处分。，但其处分被继承人朱凌云遗产（两间房屋的一半）是无效的。（两间房屋的一半）应按法定继承处理。，置换房屋还未建设，暂无法确认具体继承数额，可待置换房屋建成后另案处理。4万元存款及780元利息，被告没有证据证明被继承人生前答应由其继承该款，故应按法定继承处理。137007元丧抚费，不是被继承人的遗产。，被告对其进行了发丧，故其中的丧葬费28574元，应归被告所有。108433元，系抚恤金。。，故在遗产继承及抚恤金分割时应多分。，依照《中华人民共和国继承法》第三条、第五条、第十条一款、第十三条三款及《中华人民共和国民事诉讼法》第一百四十二条之规定，判决如下：、由三原告各继承被继承人徐世国存款及利息40780元的15%，即6117元，由被告继承55%，即22429元；、徐世国的抚恤金108433元，由三原告各分得15%，即16265元，由被告分得55%，即59638元；、徐世国的丧葬费28574元归被告所有；、二项三原告应得数额，由被告于本判决生效后5日内给付。，应当依照《中华人民共和国民事诉讼法》第二百五十三条的规定，加倍支付迟延履行期间的债务利息。4300元，减半收取2150元，由三原告各负担322元,由被告徐少清负担1184元。，可在判决书送达之日起15日内，向本院递交上诉状，并按对方当事人的人数提出副本，上诉于辽宁省锦州市中级人民法院。　　侯百吉〇一七年十一月九日　　董　健" }

## ##4.2响应

id为gid，summay为摘要

示例： {“answer”: [{“id”: “gid”, “summary”: “原被告系继承关系。1.请求继承父、母遗产房屋两间、存款4万元及利息，分割父亲的丧葬费、丧葬补助金137007元；、朱凌云的子女。人的子女，故依法对被继承人的遗产均享有继承权。人徐世国虽分给原、被告四人每人人民币5000元，但不能说明该笔钱是对被继承人朱凌云遗产（两间房屋的一半）的处分。虽是真实的，但其处分被继承人朱凌云遗产（两间房屋的一半）是无效的。一半）应按法定继承处理。设，暂无法确认具体继承数额，可待置换房屋建成后另案处理。没有证据证明被继承人生前答应由其继承该款，故应按法定继承处理。进行了发丧，故其中的丧葬费28574元，应归被告所有。要的赡养义务，故在遗产继承及抚恤金分割时应多分。法》第三条、第五条、第十条一款、第十三条三款及《中华人民共和国民事诉讼法》第一百四十二条之规定，判决如下：国存款及利息40780元的15%，即6117元，由被告继承55%，即22429元；三原告各分得15%，即16265元，由被告分得55%，即59638元；所有；被告于本判决生效后5日内给付。钱给付义务的，应当依照《中华人民共和国民事诉讼法》第二百五十三条的规定，加倍支付迟延履行期间的债务利息。”}]}

# #5.存储需求

存储gid、案由（名称）、摘要（长度约500-1500，个别小于500字）三个字段，存至mysql数据库中一张表；也可以存储成csv文件。

# #6.时间需求

摘要接口服务器的GPU显存为11G，每个请求消耗2G显存，因此理论上并发支持5个，平均1秒返回一个摘要结果。由于本工作是为了下一步的向量检索的准备工作，因此可准备1000万条以内的结果。按照一天40万篇的计算量，整个计算过程需要25天。