

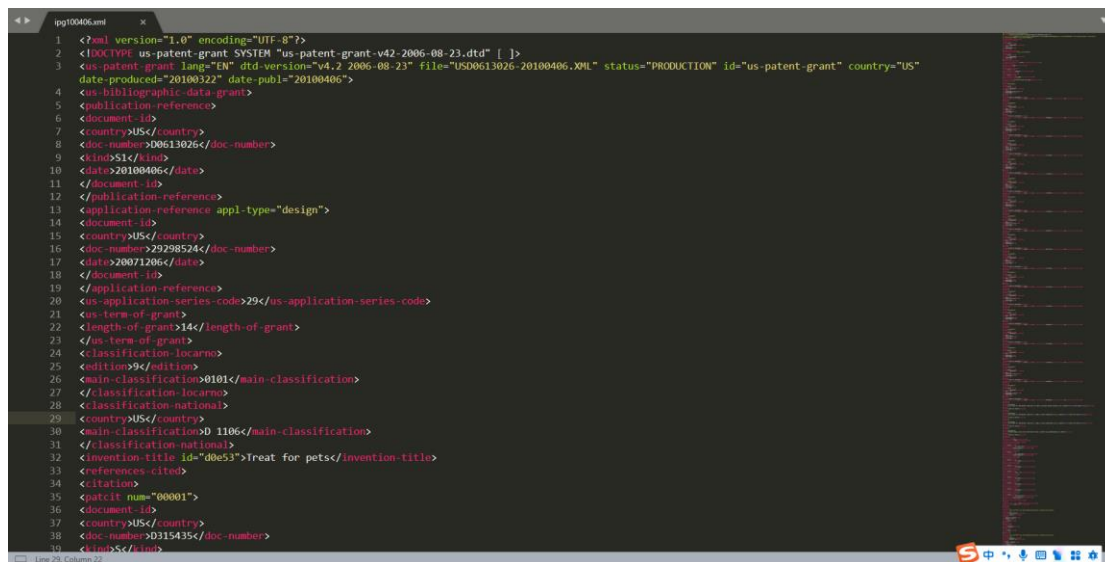
# 论文数据处理主要步骤

## 1. 数据源:

(1) USPTO 下载的 1976-2019 年的专利全文数据; (需修改下面的年份)

下载链接: <https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/2004/>

下图是 2010 年 ipg100406.xml 文件的截图, 需要在类似这种文件中解析出专利引用关系及全文内容。



(2) 从 PatentsView 数据库中筛选 CPC 分类号为 Y02E 10/545 (微晶硅光伏电池)、Y02E 10/546 (多晶硅光伏电池)、Y02E 10/547 (单晶硅光伏电池) 和 Y02E 10/548 (非晶硅光伏电池) 的专利, 专利公开年份限定为 1976-2019 年, 所搜集的目标专利类型均为已授权的发明专利。经过筛选后, 得到 4403 个目标专利。截图如下 (term 列是后续抽取的术语):

对象	candidate_patent @patents...	target_patent @patents_vie...	无标题 - 查询
开始事务	文本	排序	导入
patent_id	pub_date	term	
08143513	2012-03-27	14414,303313,6253,21623,6285,303314	
07842585	2010-11-30	277659,3169,76763,186,269277,58,277660,38,7	
09349890	2016-05-24	395,501,50,8958,338938,856,1694,6811,338939	
08207443	2012-06-26	4392,6563,308072,34758,308073	
07902460	2011-03-08	264302,264303,16487,13958,12368,31269,2643	
08242005	2012-08-14	21160,310785,256,11212,310786,1825	
06707075	2004-03-16	27070,1077,1097,1024,165117	
07812420	2010-10-12	440,9270,19937,64784,1412,395,18748,60536,6	
07029644	2006-04-18	416,198999,305,4793,199000	
04171997	1979-10-23	21,769,582,692,294	
09117948	2015-08-25	1680,304718,373805,4690,373806	
07811900	2010-10-12	39,440,183,69646,101036,46815,1058,186,2749	
08895842	2014-11-25	17373,253080,395,6240,1680,37,628,24060,311	
04775639	1988-10-04	9253,136,1424,3626,2142,32079,32080,38,3208	
06111189	2000-08-29	12422,22309,83653,6253,22309,110344,866,111	
07759572	2010-07-20	13315,29566,4312,186,264	
08993869	2015-03-31	15313,164,365341,305,365342,6778,33589,149	
08426722	2013-04-23	21,156932,21,2495,16887,17765,10154,186,431	
08168882	2012-05-01	4782,129982,303658,123222,285434,1829,1702	
07807544	2010-10-05	148991,440,1851,37870,14371,1375,274297,16	
04415760	1983-11-15	3852,1879,121,899,372	
06518087	2003-02-11	3795,6566,1406,16884,2155,2179,416,146774,1	
09112066	2015-08-18	21,13753,42034,18251,1412,64994	
09478680	2016-10-25	42,42,1010,4392,321,99838,321,74432,20639,4	
06972476	2005-12-06	17379,3371,193601,9270,3371,11733	
04609565	1986-09-02	1516,23042,23041,25338,7872,90,3537	
10504882	2019-12-10	167,167,5970,22639,879,30449,4200,45275,110	
09449818	2016-09-20	58980,394841,81541,95870,394842	
06172408	2001-01-09	352,28,87,38,1181,38932,87,44	
09768334	2017-09-19	30182,54166,30182,412995,412996	

对象	candidate_patent @patents...	target_patent @patents_vie...	无标题 - 查询
保存	查询创建工具	美化 SQL	代码片段
Mysql	patents_view	运行	停止
1 SELECT * FROM `target_patent` where pub_date < "2020-01-01"			
信息	Result 1	剖析	状态
patent_id	pub_date	term	
08143513	2012-03-27	14414,303313,6253,21623,6285,303314	
07842585	2010-11-30	277659,3169,76763,186,269277,58,277660,38,7229,6215,2037,168	
09349890	2016-05-24	395,501,50,8958,338938,856,1694,6811,338939,338940,36888,58	
08207443	2012-06-26	4392,6563,308072,34758,308073	
07902460	2011-03-08	264302,264303,16487,13958,12368,31269,264304,264305,155930,15	
08242005	2012-08-14	21160,310785,256,11212,310786,1825	
06707075	2004-03-16	27070,1077,1097,1024,165117	
07812420	2010-10-12	440,9270,19937,64784,1412,395,18748,60536,658,497,18748,476,41	
07029644	2006-04-18	416,198999,305,4793,199000	
04171997	1979-10-23	21,769,582,692,294	
09117948	2015-08-25	1680,304718,373805,4690,373806	
07811900	2010-10-12	39,440,183,69646,101036,46815,1058,186,274911,352,16061,1148	
08895842	2014-11-25	17373,253080,395,6240,1680,37,628,24060,311551,253081,190207,2	
04775639	1988-10-04	9253,136,1424,3626,2142,32079,32080,38,32081,2272	
06111189	2000-08-29	12422,22309,83653,6253,22309,110344,866,110345,13704,110346,1	
07759572	2010-07-20	13315,29566,4312,186,264	
08993869	2015-03-31	15313,164,365341,305,365342,6778,33589,1497,1948,54396,25576,4	
08426722	2013-04-23	21,156932,21,2495,16887,17765,10154,186,4312,7729,156932,30446	
08168882	2012-05-01	4782,129982,303658,123222,285434,1829,17020,303659,303660,554	
07807544	2010-10-05	148991,440,1851,37870,14371,1375,274297,1678,3309,22927	
04415760	1983-11-15	3852,1879,711,899,372	
06518087	2003-02-11	3795,6566,1406,16884,2155,2179,416,146774,145077,11677,14823,1	
09112066	2015-08-18	21,13753,42034,18251,1412,64994	
09478680	2016-10-25	42,42,1010,4392,321,99838,321,74432,20639,4463,370651,440,6285,	

候选专利集既包含目标专利的施引专利信息，也包含被引专利以及被引专利的施引专利信息，截图如下（term 列是后续抽取的术语）：

对象	candidate_patent @patents...	target_patent @patents_vie...
开始事务	文本	筛选
patent_id	pub_date	term
10000336	2018-06-19	2376,424961,11262,42178,424961
10000411	2018-06-19	27932,351,2322,70069,4021,4430,424962,22151,22151,417986,62
10000472	2018-06-19	222355,6993,14470,1426,106122,703,1595,63832,512,16527,2223
10000609	2018-06-19	389034,391393,391394,58958,541,42794,126948,389032,10000,28
10000630	2018-06-19	389328,397934,397933,397933,397935,397943,9801,220604,3979
10000645	2018-06-19	167381,54663,1268,411257,323698,11808,1268,411258,90,58,233
10000665	2018-06-19	110816,63471,424976,904,424977
10000670	2018-06-19	28837,28837,158610,169524,247984
10000680	2018-06-19	904,256924,703,943,7117
10000690	2018-06-19	305831,904,5649,628,17521,9161,4030,3579,2806,764,346519,381
10000695	2018-06-19	128,259219,424978,9341,19820,904,146964,352,35234,378607,13
10000834	2018-06-19	8600,10311,422980,55973,55974,400912,210000,424981
10000840	2018-06-19	17561,17561,7500,13144,82823
10000845	2018-06-19	14562,424982,241118,424983,352,131607,424984,64090,67808,86
10000847	2018-06-19	1143,6404,424985,3419,49755
10000852	2018-06-19	2348,25574,7986,122884,424986,39090,424987,256218,15739,440
10000864	2018-06-19	424988,808,32,295,10480,351409,6663,352,425,424989
10000934	2018-06-19	156987,382598,18727,382599,211153,424990,18727
10000965	2018-06-19	27932,351,2322,70069,4021,4430,424962,22151,22151,6215,4179
10001087	2018-06-19	424991,376421,10457,110347,5921,255996,110347,190388
10001297	2018-06-19	4468,4468,29731,259080,424992,52993,43925,424993,4211
10001327	2018-06-19	424994,424995,424996,424997,924,61754,928,372598,352499,129
10001406	2018-06-19	4262,23774,5118,24459,23774,33753,3306,167789,92578,424998,
10001407	2018-06-19	467,5660,5661,285365,425009
10001414	2018-06-19	77823,167,14802,129315,16319,333109,671,1118,21624,11976,44,
10001442	2018-06-19	197962,609,45453,1079,214485
10001444	2018-06-19	12774,32301,270655,477,425010,856,15291,25052,425011,99838,
10001450	2018-06-19	128817,425013,31829,79785,48634,77790,425014,425015,7489,42
10001473	2018-06-19	211130,425018,203263,15479,354680,703,425019,211131,354679,
10001475	2018-06-19	211130,425018,203263,15479,425020,354680,703,425019,211131,

## 2. 术语抽取

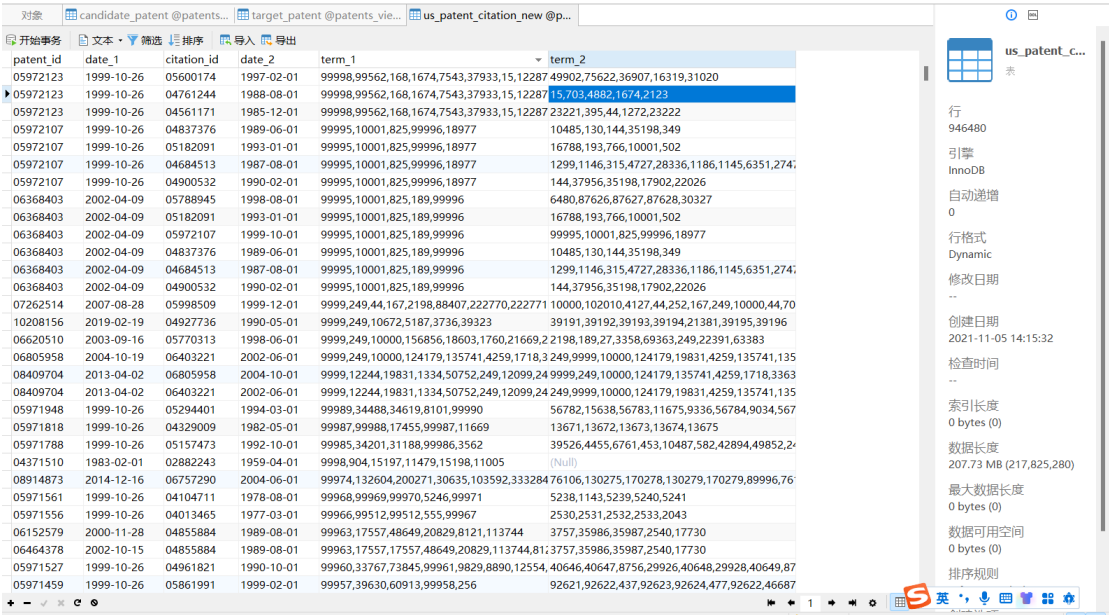
C-Value 抽取术语过程详见代码 C\_Values.py 和 extract\_term\_corenlp.py 文件，主要分为以下几步：读取全文本、切分段落、文本词性标注、抽取候选术语集、计算 C-value 值、术语筛选等。术语抽取后，得到 230696 个专利原始术语文件（一个专利对应一个文件），详见 ./term/extracted\_term/ 文件夹。

为了剔除一些噪声术语，本文进行了一些操作，详见 ./term/write\_term.py 程序。主要操作包括：进行词形还原、剔除包含数字等其他非英语字符的术语、剔除所有单词不超过三个字符的术语、剔除存在单词字符长度超过 45 的术语、剔

除全部单词均为一样的术语、剔除单词存在两个以上横杠的术语等。处理后，得到术语总计 451100 个，详见 ./term/term\_dict.txt 文件，并同时专利对应的术语写入数据库中。

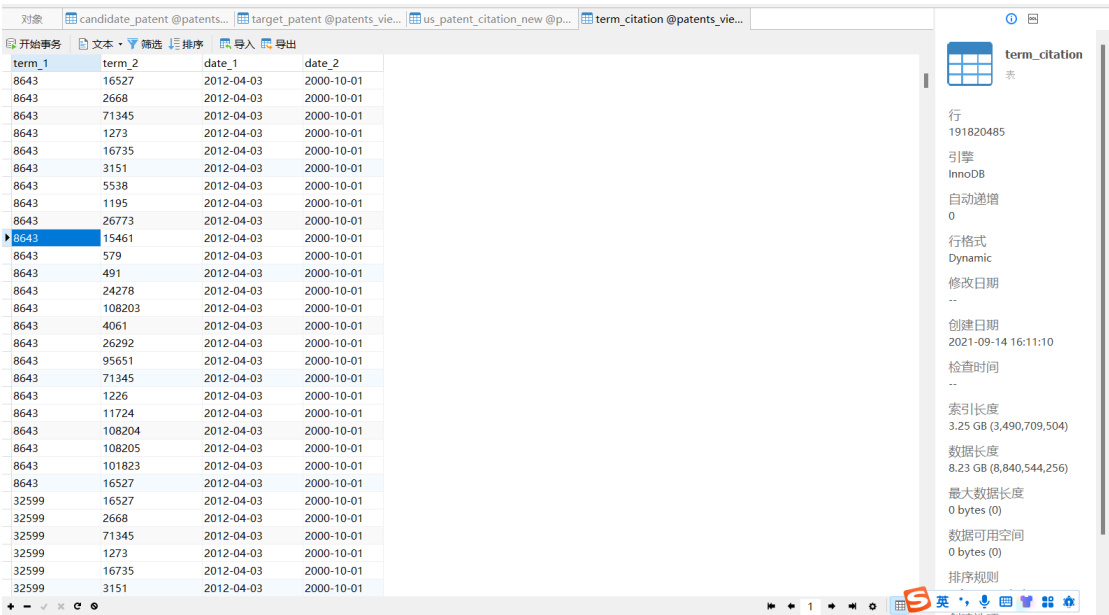
### 3. 指标计算

(1) 根据专利引用关系，构建术语引用关系，详见 term\_citation.py 程序。  
如：PatA 术语有 T1, T2; PatB 有术语 T3, T4; PatB 引用了 PatA，那么术语引用就有 T3 引用 T1 和 T2、T4 引用 T1 和 T2 这四种引用关系。（即采用笛卡尔积形式）  
专利引用数据截图：



patent_id	date_1	citation_id	date_2	term_1	term_2
05972123	1999-10-26	05600174	1997-02-01	99998,99562,168,1674,7543,37933,15,12287,49902,75622,36907,16319,31020	
05972123	1999-10-26	04761244	1988-08-01	99998,99562,168,1674,7543,37933,15,12287	15,703,4882,1674,2123
05972123	1999-10-26	04561171	1985-12-01	99998,99562,168,1674,7543,37933,15,12287	23221,395,44,1272,23222
05972107	1999-10-26	04837376	1989-06-01	99995,10001,825,99996,18977	10485,130,144,35198,349
05972107	1999-10-26	05182091	1993-01-01	99995,10001,825,99996,18977	16788,193,766,10001,502
05972107	1999-10-26	04684513	1987-08-01	99995,10001,825,99996,18977	1299,1146,315,4727,28336,1186,1145,6351,274
05972107	1999-10-26	04900532	1990-02-01	99995,10001,825,99996,18977	144,37956,35198,17902,22026
06368403	2002-04-09	05788945	1998-08-01	99995,10001,825,189,99996	6480,87626,87627,87628,30327
06368403	2002-04-09	05182091	1993-01-01	99995,10001,825,189,99996	16788,193,766,10001,502
06368403	2002-04-09	05972107	1999-10-01	99995,10001,825,189,99996	99995,10001,825,99996,18977
06368403	2002-04-09	04837376	1989-06-01	99995,10001,825,189,99996	10485,130,144,35198,349
06368403	2002-04-09	04684513	1987-08-01	99995,10001,825,189,99996	1299,1146,315,4727,28336,1186,1145,6351,274
06368403	2002-04-09	04900532	1990-02-01	99995,10001,825,189,99996	144,37956,35198,17902,22026
07262514	2007-08-28	05998509	1999-12-01	9999,249,44,167,2198,88407,222770,222771	10000,102010,4127,44,252,167,249,10000,44,70
10208156	2019-02-19	04927736	1990-05-01	9999,249,10672,5187,3736,39323	39191,39192,39193,39194,21381,39195,39196
06620510	2003-09-16	05770313	1998-06-01	9999,249,10000,156856,18603,1760,21669,2	2198,189,27,3358,69363,249,22391,63383
06805958	2004-10-19	06403221	2002-06-01	9999,249,10000,124179,135741,4259,1718,3	249,9999,10000,124179,19831,4259,135741,135
08409704	2013-04-02	06805958	2004-10-01	9999,12244,19831,1334,50752,249,12099,24	9999,249,10000,124179,135741,4259,1718,3363
08409704	2013-04-02	06403221	2002-06-01	9999,12244,19831,1334,50752,249,12099,24	249,9999,10000,124179,19831,4259,135741,135
05971948	1999-10-26	05294401	1994-03-01	99989,34488,34619,8101,99990	56782,15638,56783,11675,9336,56784,9034,567
05971818	1999-10-26	04329009	1982-05-01	99987,99988,17455,99987,11669	13671,13672,13673,13674,13675
05971788	1999-10-26	05157473	1992-10-01	99985,34201,31188,99986,3562	39526,4455,6761,453,10487,582,42894,49852,2
04371510	1983-02-01	02882243	1959-04-01	9998,904,15197,11479,15198,11005	(Null)
08914873	2014-12-16	06757290	2004-06-01	99974,132604,200271,30635,103592,333284	76106,130275,170278,130279,170279,89996,76
05971561	1999-10-26	04104711	1978-08-01	99968,99969,99970,5246,99971	5238,1143,5239,5240,5241
05971556	1999-10-26	04013465	1977-03-01	99966,99512,99512,555,99967	2530,2531,2532,2533,2043
06152579	2000-11-28	04855884	1989-08-01	99963,17557,48649,20829,8121,113744	3757,35986,35987,2540,17730
06464378	2002-10-15	04855884	1989-08-01	99963,17557,17557,48649,20829,113744,81	3757,35986,35987,2540,17730
05971527	1999-10-26	04961821	1990-10-01	99960,33767,73845,99961,9829,8890,12554	40646,40647,8756,29926,40648,29928,40649,87
05971459	1999-10-26	05861991	1999-02-01	99957,39630,60913,99958,256	92621,92622,437,92623,92624,477,92622,46687

术语引用数据截图：



term_1	term_2	date_1	date_2
8643	16527	2012-04-03	2000-10-01
8643	2668	2012-04-03	2000-10-01
8643	71345	2012-04-03	2000-10-01
8643	1273	2012-04-03	2000-10-01
8643	16735	2012-04-03	2000-10-01
8643	3151	2012-04-03	2000-10-01
8643	5538	2012-04-03	2000-10-01
8643	1195	2012-04-03	2000-10-01
8643	26773	2012-04-03	2000-10-01
8643	15461	2012-04-03	2000-10-01
8643	579	2012-04-03	2000-10-01
8643	491	2012-04-03	2000-10-01
8643	24278	2012-04-03	2000-10-01
8643	108203	2012-04-03	2000-10-01
8643	4061	2012-04-03	2000-10-01
8643	26292	2012-04-03	2000-10-01
8643	95651	2012-04-03	2000-10-01
8643	71345	2012-04-03	2000-10-01
8643	1226	2012-04-03	2000-10-01
8643	11724	2012-04-03	2000-10-01
8643	108204	2012-04-03	2000-10-01
8643	108205	2012-04-03	2000-10-01
8643	101823	2012-04-03	2000-10-01
8643	16527	2012-04-03	2000-10-01
32599	16527	2012-04-03	2000-10-01
32599	2668	2012-04-03	2000-10-01
32599	71345	2012-04-03	2000-10-01
32599	1273	2012-04-03	2000-10-01
32599	16735	2012-04-03	2000-10-01
32599	3151	2012-04-03	2000-10-01

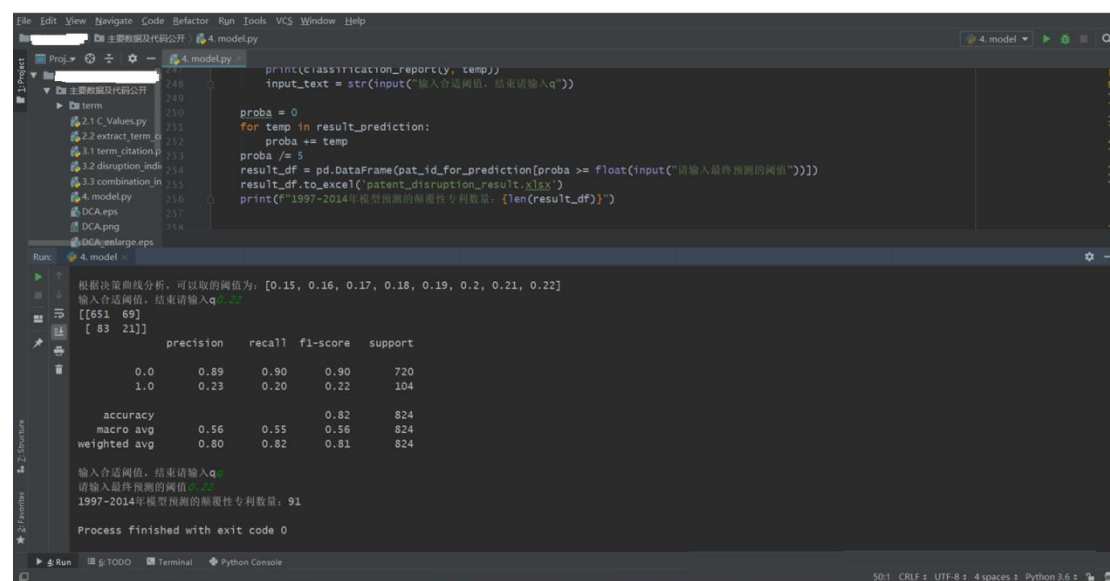
(2) 计算专利公开 5 年后的影响力指标，详见 disruption\_indicator.py 文件。

(3) 计算计算基于术语创新和术语复用的四个指标，详见 combination\_indicator.py 文件。

(4) 将上述数据按照 patent\_id 整合起来得到最终的特征数据 model\_data.xlsx。根据 G. F. Netmet 等和 B. Sun 等的研究，将 1977-1996 年期间公开的 824 个专利作为参加训练模型的数据，这些专利中有 104 个专利被上述两个研究确定为具有重大颠覆的专利，剩余的 720 个专利被确定为非颠覆性专利。

## 4. 模型训练及预测

详见 model.py 文件。主要步骤：读取 model\_data 数据，1977-1996 年期间 824 个专利进行模型训练，绘制 DCA 曲线以确定最优分类阈值。确定分类阈值后，对 1997-2014 年专利进行预测，最后的预测结果 patent\_disruption\_result.xlsx。程序运行截图如下：



```
print(classification_report(y, temp))
input_text = str(input("输入合适阈值，结束请输入q"))

proba = 0
for temp in result_prediction:
    proba += temp
proba /= 5
result_df = pd.DataFrame(pat_id_for_prediction[proba >= float(input("请输入最终预测的阈值"))])
result_df.to_excel('patent_disruption_result.xlsx')
print(f"1997-2014年模型预测的颠覆性专利数量: {len(result_df)}")
```

根据决策曲线分析，可以取的阈值为: [0.15, 0.16, 0.17, 0.18, 0.19, 0.2, 0.21, 0.22]  
输入合适阈值，结束请输入q 0.22  
[[651 69]  
[ 83 21]]

	precision	recall	f1-score	support
0.0	0.89	0.90	0.90	720
1.0	0.23	0.20	0.22	104
accuracy			0.82	824
macro avg	0.56	0.55	0.56	824
weighted avg	0.80	0.82	0.81	824

输入合适阈值，结束请输入q  
请输入最终预测的阈值 0.22  
1997-2014年模型预测的颠覆性专利数量: 91  
Process finished with exit code 0