信息抽取实验系统

一、综述

我们的信息抽取实验系统，根据收集到的100篇来自北京邮电大学信息门户上发布的通知，实现了采用正则表达式匹配算法进行特定的八个信息点的抽取，同时实现了实体识别及抽取，文章关键词抽取，文章情感分析，事件蒸馏提取，文章句法语法分析，文章智能分段，关系抽取以及关系网络的构建（可视化图网络），最后我们还实现了一个简易的问答机制，系统可以根据使用者所提出的自然语言问题以及对应文章内容返回相应的答案。

除了基本的信息抽取以及我们扩展的更深层的信息抽取功能以外，我们的系统也支持多媒体信息抽取，可将语言信息中的信息点进行抽取。

二、对环境和社会可持续发展影响的思考

传统图书馆中陈列种类繁多的书籍，固然现在已经有了多种可以对书籍进行快速搜索的办法，但是对于一本书籍进行概括，却是十分耗费人力物力的。同时，为满足读者对自己所需求的书籍进行查找或者相关知识的查找，对书籍的知识的抽取又是必不可少的。我们的系统满足对相应书籍的内容进行多元化知识的抽取，可持续发展是一种全新的发展观念和发展战略、是人类实践和科学技术高度发展的产物，是人类以沉重代价换来的重大认识成果。现代科学技术的进步，如我们这个系统的设计与实现，其实是推动了社会的进步的一种体现，是实现社会的可持续发展仍然要依靠科技的发展和作用的体现。

三、系统实现

3.1 工具列表

本次课程实验使用python编写，使用了以下库进行开发：

HarvestText，networkx，matplotlib，argparse，jieba

3.2 运行方法

注：首先需要进行环境配置，相关部分见 README.md。

本系统可以命令行模式运行，主运行逻辑为运行主目录下的 TextInfoExtract.py 文件，可以输入 python TextInfoExtract.py -h 获得命令行参数的帮助信息，结果如下所示：

根据提示的帮助信息可以运行系统，如 python TextInfoExtract.py －all

下面是命令行的帮助信息：

```
usage: TextInfoExtract.py [-h] [--all] [--basic] [--ent] [--k K] [--doc DOC]
[--specificdoc SPECIFICDOC] [--sentence] [--sents] [--event] [--para] [--p P]
                    [--qa] [--network] [--q Q]

optional arguments:
  -h, --help            show this help message and exit
  --all                 Generate all infos
  --basic               Extract basic infos
  --ent                 Recognize entities
  --k K                 topK keywords
  --doc DOC             Num of docs to extract infos
  --specificdoc SPECIFICDOC
                        Specific doc to extract
  --sentence            analyze syntax of sentences
  --sents               analyze sentiment of passage
  --event               Distillation of event
  --para                Cut paragraph
  --p P                 Goal number of paragraphs
  --qa                  Raise questions to system
  --network             Generate a social network
  --q Q                 When enabled qa mode, this is use to record questions
```

特别注：需要进行抽取的信息需要放在数据文件夹下，以数字号命名，从 1 开始，如 1.txt。在使用 specificdoc 参数的时候，需要把文件放到数据文件夹下，在 specificdoc 参数中填入文件的名字，如 1,2 这样，不需要后缀名。

四、系统功能描述与示例

***以下所有的结果均保存在目录下“提取结果”文件夹下。**

4.1 特定信息点抽取

特定信息点抽取根据正则表达式匹配算法实现，设计如下的正则表达式模式：

```
1. pattern_title  =  re.compile(r'[\s]*[\S]+[\s]*\n')
2. pattern_department  =  re.compile(r'发布部门：[\s]*[\S]+')
3. pattern_time  =  re.compile(r'发布时间：[\s]*[\S]+')
4. pattern_count  =  re.compile(r'浏览[\s]*[\d]+[\s]*次')
5. pattern_school  =  re.compile(r'[\S]+学院')
```

实际进行特定信息点抽取时，我们增加了别的方法抽取了更多的特定信息点（因不是所有信息都有固定的模式可用于匹配）。

下面是一个抽取结果的实例：

```
标题:      【获奖通知】关于公布 2020 年全国大学生英语竞赛获奖名单的通知
发布部门:  教务处
发布时间:  2021-03-22
浏览量:    902
学院:      各学院
关键字:    ['获奖', '教务处', '竞赛', '同学', '名单']
篇幅:      339
可能相关:  ['公布', '8 名', '领取', '我校', '获奖名单', '时间', '等奖 1', '竞赛中']
```

可以观察到，我们对于每篇需要进行抽取的文章均进行了八个特定信息点的抽取。

具体抽取结果可以在“./提取结果/信息/”该目录下用文本进行了相关抽取结果的保存。

4.2 实体识别及抽取

系统可以对文章中出现的实体进行识别，并以键值对的数据结构进行存储

下面是一个实体抽取结果的示例：

实体 0: 专题心理培训:机构名
实体 1: 业务水平:其他专名
实体 2: 有效地:其他专名
实体 3: 学生工作部处:机构名
实体 4: 二培训:机构名
实体 5: 西土城:地名
实体 6: 发展中心:机构名
实体 7: 北京邮电大学:机构名
实体 8: 蔺秀云:人名
实体 9: 北京师范大学:机构名
实体 10: 心理学部:机构名
实体 11: 教育部:机构名
实体 12: 长江:地名
实体 13: 家庭治疗:其他专名
实体 14: 中国心理学会:机构名
实体 15: 婚姻家庭:其他专名
实体 16: 咨询专业委员会:机构名
实体 17: 中国教育学会:机构名
实体 18: 学校教育心理学分会:机构名
实体 19: 心理咨询:其他专名
实体 20: 学生促进学生:机构名
实体 21: 七培训:机构名

观察到，实体的类型也被识别出来了。

如成功识别出长江为地名；北京邮电大学，教育部等是机构名；蔺秀云是人名。

4.3 文章关键词抽取

文章的关键词抽取技术依赖于 TFIDF 算法，TF-IDF（term frequency - inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。同时 TF-IDF 还是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。因此，在关键词的抽取我们选择 TFIDF 算法进行。

我们针对 TFIDF 对分词以及去停等词后的文章词汇进行了排序，并通过选择前 topK 的词汇作为文章的关键词。（topK 设置为超参数，默认为 5）。

下面是一个提取到的关键词的实例，用列表存储：

```
关键字:    ['辅导员', '培训', '学生', '心理', '技巧']
```

4.4 文章情感分析

我们开发的文章情感分析功能支持正，反向情感值预测（positive，negative），以及我们自己设计的 7 个情感字的情绪分析结果。

下面是一个情感分析实例：

```
正向情感值: 465.0
负向情感值: 68.0
情绪分析结果:
好: 15
乐: 3
哀: 0
怒: 0
惧: 0
恶: 2
惊: 0
```

4.5 事件蒸馏提取

此功能可以从整篇文章蒸馏抽取出所有事件，并以三元组进行标识。

下面是一个事件提取蒸馏的实例：

['教师', '讲述', '我育人活动通知 发布部门人事处 发布时间 20210507 浏览 279 次关于开展育才']

['我育人活动', '育人', '故事']

['我育人活动通知 发布部门人事处 发布时间 20210507 浏览 279 次关于', '开展', '育才']

['教师', '讲述', '挖掘']

['我育人故事活动通知各学院体育部', '为', '贯彻落实习近平总书记考察']

['生动实践', '践行', '教书育人神圣使命']

['践行教书育人神圣使命生动实践', '强化', '广大教师听党话跟党走自觉']

['广大教师', '听', '党话跟党走']

['北京市委教工委', '发布', '关于']

['第三届北京市大中小幼教师', '讲述', '我我们育人故事活动通知']

['我校', '开', '展教师']

['我育人故事活动通知', '为', '党育人为国育才二活动时间 2021 年 5 月']

['7 月三主要内容我育人故事讲述者', '讲述', '秉承三育人']

['理念', '围绕', '学生成长生动故事']

['广大教师', '坚持', '党教育方针站自觉做中国特色社会主义坚定信仰者做大先生']

['党教育方针', '站在', '一起']

['党教育方针站自觉做中国特色社会主义坚定信仰者', '做', '大先生']

['示范', '为', '人']

['学校党委教师', '工作', '部发布']

['四活动安排工作活动', '工作', '部发布']

['广大教师同心思想共识', '结合', '跟党走']

['活动政治关各院', '推荐', '参加学校']

['活动', '开展', '讲述育人故事活动基础上组织']

['基础', '讲述', '育人故事活动']

['校级', '讲述', '活动']

['案例五活动要求', '请', '讲述活动']

['主题教育活动重要载体', '宣传', '参与总结发送至联系方式电话 62285129 徐老师 党委教师工作部 2021 年 5 月 7 日']

['组织', '做好', '活动']

['做好活动组织请于 6 月 18 日之前', '组织', '宣传工作']

4.6 文章句法语法分析

本功能可将分句后的文章的句法语法进行分析，对句子的句子结构，词语的语法关系进行抽取。

一个部分的实例如下所示：（因完整分析较长，可以使用系统观察）

```
[0, '关于', 'p', '状中结构', 11]
[1, '开展', 'v', '介宾关系', 0]
[2, '为', 'p', '状中结构', 7]
[3, '党', 'n', '定中关系', 4]
[4, '育人', 'v', '介宾关系', 2]
[5, '为', 'p', '状中结构', 7]
[6, '国', 'n', '介宾关系', 5]
[7, '育才', 'v', '动宾关系', 1]
[10, '教师', 'n', '主谓关系', 11]
[11, '讲述', 'v', '核心关系', -1]
[12, '我', 'r', '定中关系', 16]
[13, '的', 'u', '右附加关系', 12]
[14, '育人', 'vn', '定中关系', 16]
[15, '故事', 'n', '动宾关系', 14]
[16, '活动', 'vn', '定中关系', 19]
[17, '的', 'u', '右附加关系', 16]
```

4.7 文章智能分段

本功能对文章进行只能分段，可以对需要分成的段落的数目进行设置。（通过命令行参数）

下面是一个分段实例：

```
共 10 段, 分段结果:
第 1 段: 关于开展“为党育人，为国育才”——教师讲述我的育人故事活动的通知发布部门：人事处　　发布时间： 2021-05-07　浏览 279 次关于开展“为党育人，为国育才”——教师讲述我的育人故事活动的通知各学院、体育部：为贯彻落实习近平总书记在清华大学考察时的重要讲话精神，挖掘展示教师队伍忠于党的教育事业、践行教书育人神圣使命的生动实
```

践和感人事迹，唱响“为党育人，为国育才”主旋律，进一步强化广大教师“听党话、跟党走”的思想自觉和行动自觉，以实际行动庆祝中国共产党成立100周年，北京市委教工委和北京市教委于4月下旬发布了“关于开展‘为党育人，为国育才’——第三届北京市大中小幼教师讲述我（我们）的育人故事活动的通知”，学校党委教师工作部结合实际和党史学习教育工作安排，现就我校开展教师讲述我的育人故事活动通知如下。一、活动主题为党育人，为国育才

第2段：二、活动时间

第3段：2021年5月—7月三、主要内容

第4段：“我的育人故事”讲述者为一线教师，重点讲述秉承“三全育人”理念，围绕学生成长成才爱岗敬业、无私奉献的生动故事。讲述内容要突出“为党育人，为国育才”的鲜明主题，集中展示广大教师坚持党的教育方针，始终同党和人民站在一起，自觉做中国特色社会主义的坚定信仰者和忠实实践者，做“大先生”，做学生为学、为事、为人的示范，进一步凝聚广大教师同心筑梦的思想共识。

第5段：四、活动安排

第6段：（1）5月上旬，学校党委教师工作部发布活动通知。

第7段：（2）5月中旬—6月中旬，学院（体育部）结合“永远跟党走”主题教育活动“讲起来”的安排，广泛开展讲述育人故事活动，育人故事讲述时间为8分钟左右。学院党委要把好讲述活动的政治关。各院择优推荐1名教师参加学校讲述活动。（3）6月下旬—7月上旬，学校在各学院开展讲述育人故事活动的基础上组织校级讲述活动，并推荐1名教师参加市级层面集中展示的优秀讲述案例。五、活动要求

第8段：请各学院、体育部高度重视，将讲述活动作为党史学习教育和师德教育的重要内容，作为“永远跟党走”主题教育活动的重要载体，广泛宣传，动员全体教师参与，做好活动的组织、宣传工作。请于6月18日之前将活动总结发送至jsgzb@bupt.edu.cn.

第9段：联系方式：电话：62285129，徐老师。

第10段：党委教师工作部2021年5月7日

4.8 关系抽取以及关系网络的构建

本功能以关键词为核心，对文章中的实体关系进行建模。

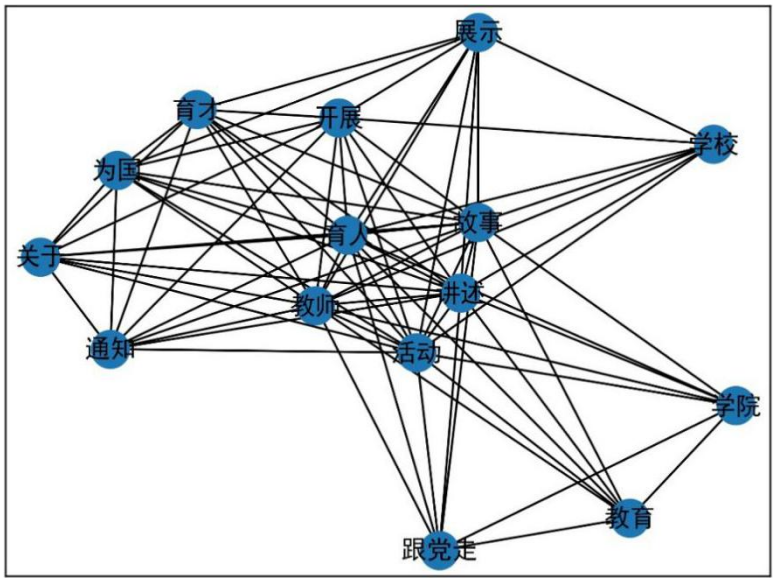
对文章中实体关系进行抽取，构造成一个带权无向图，无向图可以以一个字典的形式展现出所有的边，我们同时也给出了可视化的关系网络图。

例如下的关系抽取示例：

关系网络字典表示:

{('关于', '通知'): {'weight': 3}, ('关于', '开展'): {'weight': 3}, ('关于', '育人'): {'weight': 3}, ('关于', '教师'): {'weight': 3}, ('关于', '为国'): {'weight': 3}, ('关于', '讲述'): {'weight': 3}, ('关于', '育才'): {'weight': 3}, ('关于', '活动'): {'weight': 3}, ('关于', '故事'): {'weight': 3}, ('通知', '开展'): {'weight': 3}, ('通知', '育人'): {'weight': 3}, ('通知', '教师'): {'weight': 3}, ('通知', '为国'): {'weight': 3}, ('通知', '讲述'): {'weight': 3}, ('通知', '育才'): {'weight': 3}, ('通知', '活动'): {'weight': 3}, ('通知', '故事'): {'weight': 3}, ('开展', '育人'): {'weight': 4}, ('开展', '教师'): {'weight': 4}, ('开展', '为国'): {'weight': 3}, ('开展', '讲述'): {'weight': 4}, ('开展', '育才'): {'weight': 3}, ('开展', '活动'): {'weight': 4}, ('开展', '故事'): {'weight': 4}, ('开展', '展示'): {'weight': 2}, ('开展', '学校'): {'weight': 2}, ('育人', '教育'): {'weight': 2}, ('育人', '学校'): {'weight': 2}, ('育人', '跟党走'): {'weight': 2}, ('育人', '教师'): {'weight': 6}, ('育人', '为国'): {'weight': 4}, ('育人', '讲述'): {'weight': 7}, ('育人', '育才'): {'weight': 4}, ('育人', '活动'): {'weight': 5}, ('育人', '故事'): {'weight': 6}, ('育人', '展示'): {'weight': 3}, ('育人', '学院'): {'weight': 2}, ('教师', '教育'): {'weight': 2}, ('教师', '学校'): {'weight': 3}, ('教师', '跟党走'): {'weight': 2}, ('教师', '为国'): {'weight': 4}, ('教师', '讲述'): {'weight': 8}, ('教师', '育才'): {'weight': 4}, ('教师', '活动'): {'weight': 6}, ('教师', '故事'): {'weight': 5}, ('教师', '展示'): {'weight': 3}, ('教师', '学院'): {'weight': 2}, ('为国', '讲述'): {'weight': 4}, ('为国', '育才'): {'weight': 4}, ('为国', '活动'): {'weight': 3}, ('为国', '故事'): {'weight': 3}, ('为国', '展示'): {'weight': 2}, ('讲述', '教育'): {'weight': 3}, ('讲述', '学校'): {'weight': 3}, ('讲述', '跟党走'): {'weight': 3}, ('讲述', '育才'): {'weight': 4}, ('讲述', '活动'): {'weight': 8}, ('讲述', '故事'): {'weight': 6}, ('讲述', '展示'): {'weight': 3}, ('讲述', '学院'): {'weight': 4}, ('育才', '展示'): {'weight': 2}, ('育才', '活动'): {'weight': 3}, ('育才', '故事'): {'weight': 3}, ('活动', '教育'): {'weight': 3}, ('活动', '学校'): {'weight': 3}, ('活动', '跟党走'): {'weight': 3}, ('活动', '展示'): {'weight': 2}, ('活动', '故事'): {'weight': 5}, ('活动', '学院'): {'weight': 4}, ('故事', '教育'): {'weight': 2}, ('故事', '学校'): {'weight': 2}, ('故事', '跟党走'): {'weight': 2}, ('故事', '展示'): {'weight': 2}, ('故事', '学院'): {'weight': 2}, ('教育', '跟党走'): {'weight': 3}, ('教育', '学院'): {'weight': 2}, ('跟党走', '学院'): {'weight': 2}, ('展示', '学校'): {'weight': 2}}

关系网络的可视化表示如下图所示：



4.9 问答机制（实验性）

本功能是一个实验性的功能，可以针对用户所提问题，我们所开发的系统会根据问题以及文章内容给予使用者相应的回复。

下面是两个问答机制的实例，可以看出此功能还只是个实验性的功能，值得一提的是，现在 NLP 领域发表的 paper 中针对这个问题也并没有太多的进展，中文文本难度更高，期待之后的进展。

```
python TextInfoExtract.py --specificdoc=34 --all --q='党支部活动'
```

上面语句的意思是对名称为 34 的文档进行信息抽取，问题是党支部活动。

下面是问答机制的输出：

```
问题: '党支部活动'
回答: 共建
```

```
python TextInfoExtract.py --specificdoc=1 --all --q='支部开展了什么'
```

上面语句的意思是对名称为 1 的文档进行信息抽取，问题是支部开展了什么。

下面是问答机制的输出：

```
问题: '支部开展了什么'
回答: 组织党建工作
```

五、信息抽取结果准确率人工评价

5.1 信息抽取系统准确率人工评价

我们对特定信息点的匹配结果通过 Accuracy，Precision，Recall，F1-score 四个指标进行了信息抽取系统准确率的评价。

结果如下：（保存在“结果指标”文本下）

	标题	部门	时间	浏览量	学院	关键字	相关词
Accuracy :	100%	100%	100%	100%	100%	100%	100%
Precision :	100%	100%	100%	100%	100%	100%	100%
Recall :	100%	100%	100%	100%	100%	100%	100%
F1-score :	100%	100%	100%	100%	100%	100%	100%

可以看到，因为正则表达式匹配算法的严谨性，以及信息门户的通知格式具有一定的标准格式，以及其他因素的影响（如样本数过少等），所有指标均达到了 100%，证明了我们信息抽取系统的有效性。

六、多媒体信息抽取

多媒体信息抽取基于实现的基本文本信息抽取系统，输入为 wav 语音文件，通过对语音信息进行文字转换，之后以转换后的文本信息为输入，运行信息抽取系统，相当于将语言信息中的信息点进行抽取。同时由文字抽取系统识别出的内容具有一定的语义信息和纠错机制，对结果有一定的优化效果。在命令行和网页模式下会显示对应的抽取结果。

（注：语音信息的文字转换采用 google 开发的在线 api，其他的免费工具效果很不理想。因此需要在可以连接境外网络的环境下执行程序，否则会出现网络请求错误。本工具由谷歌团队进行维护和优化）

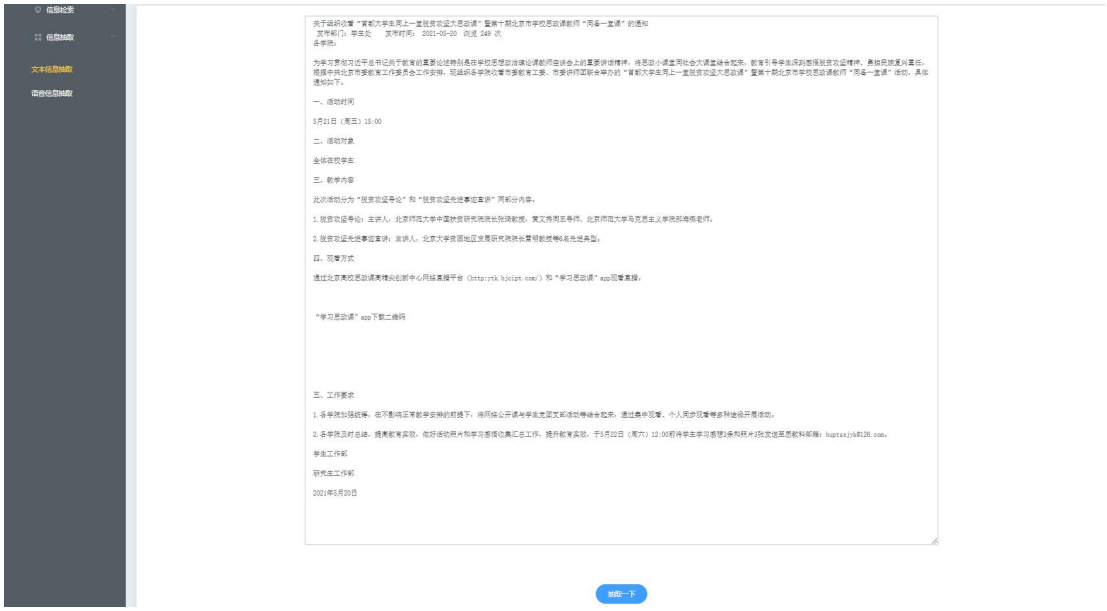
使用方法：

在 Extract 文件夹下，执行 python VoiceTextInfoExtract.py --source=[] --target=[] 来执行语音抽取，其中 source 是源语音地址，只能是 wav 格式，target 是抽取输出，为一个 txt 和一个 jpg 文件，分别为抽取的内容和网络关系图。其中抽取内容会命名为'{target}'，网络关系图会命名为'{target}.jpg'。

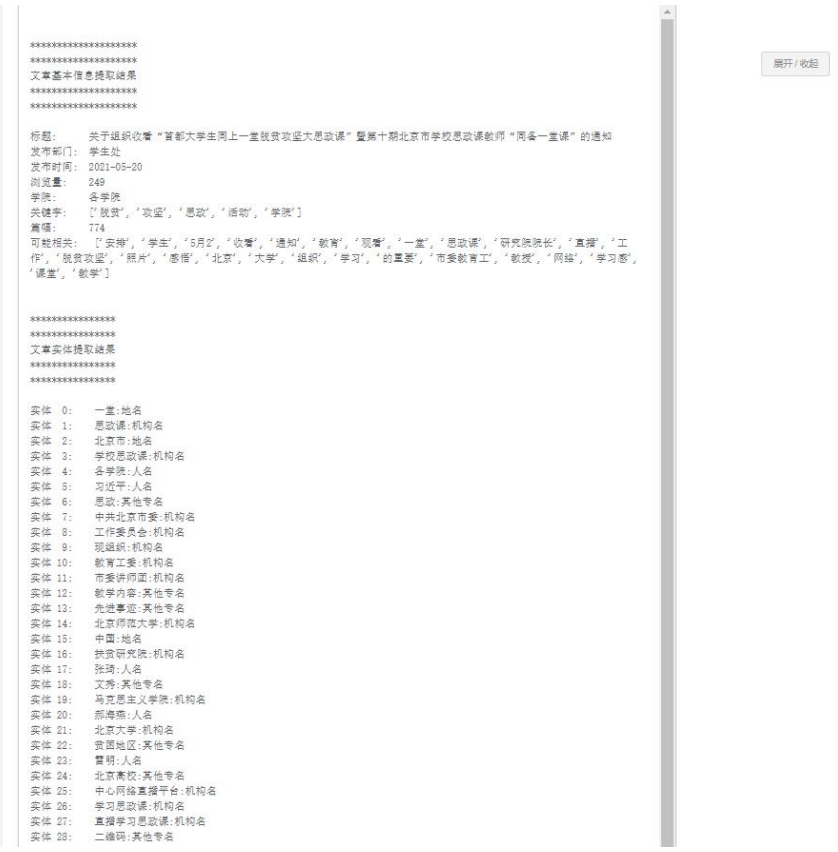
七，运行实例

本系统同时支持图形化界面（前端）以及命令行使用，在此仅展现图形化界面的运行实例

运行界面如下，以及输入了要进行信息抽取的文本。



抽取结果的文本显示如下：



****
****
文章情感分析结果
****
****

正向情感值：645.0
负向情感值：1.0

情绪分析结果：
好：19
乐：7
哀：0
怒：0
惧：0
恶：0
惊：0

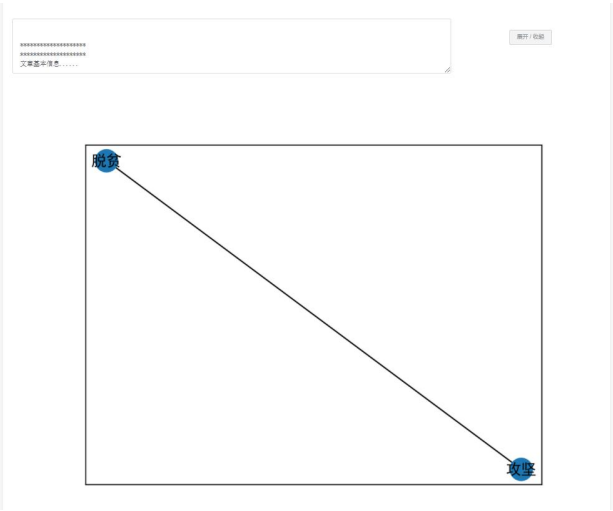
****
****
文章事件提取结果（三元组表示）
****
****

['北京市学校', '思', '政课']
['关于组织一堂脱贫攻坚大思政课', '各', '一堂课']
['精神', '担', '复兴']
['中共北京市委', '教育', '工作委员会']
['教育工作', '教育', '工作委员会']
['关于组织一堂脱贫攻坚大思政课各一堂课通知 发布部门学生处 发布时间 20210520 浏览 249 次各学院', '组织', '收看市委教育工委市委讲师团举办首都大学生同上一堂脱贫攻坚大思政课']
['一堂脱贫攻坚大思政课', '收看', '市委教育工委市委讲师团举办首都大学生同上']
['北京市学校', '思', '政课']
['第十期北京市学校思政课教师', '各', '一堂课']
['如下一活动时间5月21日周五15:00二活动对象', '分为', '通过']
['北京高校', '思政课', '创']
['新中心网络直播平台http', '观看', '直接学习思政课']
['app下载二维码五工作', '要求', '加强统筹']
['前提下', '影响', '正常教学安排']
['影响正常教学安排前提下', '结', '合起']

****
****
根据文章内容智能分段提取结果
****
****

共 10 段，分段结果：
第1段：关于组织收看“首都大学生同上一堂脱贫攻坚大思政课”暨第十期北京市学校思政课教师“同备一堂课”的通知发布部门：学生处　　发布时间：2021-05-20　浏览 249 次各学院：为学习贯彻习近平总书记关于教育的重要论述特别是在学校思想政治理论课教师座谈会上的重要讲话精神，将思政小课堂同社会大课堂结合起来，教育引导学生深刻感悟脱贫攻坚精神、勇担民族复兴重任，根据中共北京市委教育工作委员会工作安排，现组织各学院收看市委教育工委、市委讲师团联合举办的“首都大学生同上一堂脱贫攻坚大思政课”暨第十期北京市学校思政课教师“同备一堂课”活动，具体通知如下。
第2段：一、活动时间
第3段：5月21日（周五）15:00二、活动对象全体在校学生
第4段：三、教学内容
第5段：此次活动分为“脱贫攻坚导论”和“脱贫攻坚先进事迹宣讲”两部分内容。1.脱贫攻坚导论：主讲人，北京师范大学中国扶

对关系抽取到的关系网络图如下所示：



八、引用参考

https://blog.csdn.net/weixin_43758551/article/details/108482905

https://harvesttext.readthedocs.io/en/latest/

https://blog.csdn.net/hfutdog/article/details/88085878?utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromMachineLearnPai2%7Edefault-1.control&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromMachineLearnPai2%7Edefault-1.control