

# 基于深度学习的中文自然语言系统 DeepNLP4.2

## （原 FudanDNN-NLP）使用手册

### 目录

1. 系统介绍.....	3
1.1. 功能简介.....	3
1.2. 版本演进.....	3
2. 部署与开发.....	5
2.1. 系统部署.....	5
3. 使用说明.....	5
3.1. 中文分词.....	5
3.1.1. 使用方法.....	5
3.1.2. 配置文件.....	7
3.1.3. 增加词汇及类型.....	8
3.1.4. 重新或调整训练中文分词网络.....	9
3.1.5. 使用字或词向量初始化网络.....	11
3.2. 中文命名识别.....	11
3.2.1. 使用方法.....	11
3.2.2. 配置文件.....	12
3.2.3. 重新或调整训练中文命名识别网络.....	12
3.3. 中文词性标注.....	12
3.3.1. 使用方法.....	12
3.3.2. 配置文件.....	13
3.3.3. 重新或调整训练中文词性标注网络.....	13
3.4. 句子语义表示模型.....	13
3.4.1. 使用方法.....	14
3.4.2. 训练样本准备.....	14
3.4.3. 句子模型训练.....	14
3.5. 基于 CRF 的中文语义分析.....	16
3.5.1. 使用方法.....	16
3.5.2. 配置文件.....	16
3.5.3. 训练基于 CRF 的中文语义分析模型.....	17
3.5.4. 基于 CRF 的多事件中文语义分析.....	19
3.6. 对话问答.....	20
3.6.1. 系统部署和使用.....	20
3.6.2. 处理流程.....	24
3.6.3. 增量训练.....	25
3.6.4. 自定义问题回答对.....	25
3.6.5. 常见问题检索 (FAQ) 返回 Top-k 的回答.....	27
3.6.6. 配置文件.....	27
3.6.7. 用户上下文信息.....	28

3.7.	依存句法分析.....	28
3.8.	篇章级中文分词和词性标注.....	29
3.9.	领域词典加密.....	29
3.10.	动态增加词汇.....	30
3.10.1.	单个增加词汇.....	30
3.10.2.	批量增加词汇.....	31
3.11.	文本或句子生成特征向量.....	31
3.11.1.	语料集转换成特征向量表示.....	31
3.11.2.	文章或句子转换成特征向量表示.....	32
3.12.	热点新闻发现.....	32
3.13.	文本或句子聚类.....	33
3.13.1.	聚类特征向量生成.....	33
3.13.2.	运行聚类算法.....	33
3.13.3.	聚类结果显示.....	33
3.13.4.	一键式聚类.....	34
3.14.	文本或句子分类.....	34
3.14.1.	分类特征向量生成.....	34
3.14.2.	运行分类算法.....	34
3.14.3.	分类模型使用.....	34
3.15.	文本信息抽取.....	35
3.15.1.	样本准备.....	35
3.15.2.	模型训练.....	35
3.15.3.	信息抽取.....	36
3.16.	情感分析.....	37
3.16.1.	模型训练.....	37
3.16.2.	文本或句子情感分析.....	37
3.16.3.	词汇情感分析.....	38
3.17.	文本摘要和关键词抽取.....	38
3.18.	关键短语和关键词抽取.....	40
3.19.	拼音转换与相似发音分析.....	42
3.19.1.	句子级多音词消歧.....	42
3.19.2.	语句或词汇发音相似性分析.....	42
3.20.	三元组抽取.....	42
3.20.1.	样本准备.....	43
3.20.2.	模型训练.....	43
3.20.3.	模型使用.....	46
4.	注意事项.....	47
5.	联系方式.....	47
6.	参考文献.....	48

# 1. 系统介绍

## 1.1. 功能简介

本使用手册介绍复旦大学计算机学院机器人研究实验室所开发的基于深度学习的中文自然语言处理工具 DeepNLP4.2（原名为 FudanDNN-NLP），该工具目前可用于中文分词、命名识别、词性标注、句法分析、语义分析、知识库访问、对话问答、聚类分类、文本摘要、信息抽取、情感分析。深度学习方法优点在于不需要预先根据任务进行特征选择（特征工程），系统所需参数较少（节省内存开销），并且实际使用远远快于其它相似性能的系统。该工具目前具备以下特点：

- (1) 提供中文分词、命名识别、词性标注、句子分类、语义分析、对话管理等较完整的一套中文处理工具，并且支持由用户所准备的样本进行模型的重新训练和调整训练（即在已经大量样本训练的基础上根据新样本进行精调）。
- (2) 具备动态增加词汇、词典加密、文本聚类、文本分类、文本信息抽取、情感分析、文本摘要、关键短语和关键词抽取、三元组抽取。
- (3) 支持多轮问答和上下文相关问答。上下文相关分为两种情况：一种情况是处理类似上一句问“今天北京天气如何？”，然后下一句追问“上海呢？”（默认开启）；另一种情况是根据对话主题展开、转换和递进给出合适的回答。
- (4) 海量自定义问答对的高效检索。检索匹配时考虑同义词替换，并根据发音相似性纠正可能的错误，并且支持一次提问中包括多个问题的情况。
- (5) 自带包括天气预报、互联网信息检索、出行路线查询、简单算术、诗词朗读、术语解读、菜谱查询、家电控制等场景，并且可以在所提供的框架下方便地新增应用场景。
- (6) 可以与 RDF 表示的知识库（类似知识图谱）对接，用于支持对知识库的访问，以及其上的知识问答。
- (7) 增加预定义问题优先处理机制（通过给出优先处理的问题和回答对文件），以满足一些个性化问答需求。
- (8) 内制包括人名、地名、组织机构、政府权威机构、上市公司、新闻媒体、时间、时期、日期、货币、股票交易代码、度量、邮箱地址、网址、百分比、手机号码、座机号码、成语、俚语、食物、著名景点、文学和影视作品等 30 多类词汇和实体的识别。
- (9) 增加了金融、疾病、药物、动物、植物、化学等专业词汇，并且支持自定义领域词汇。
- (10) 提供客户端用于系统演示和调试，支持本地或服务器快速部署。

## 1.2. 版本演进

4.2较4.1版本，本次主要增加以下功能：

- (1) 增加了用户自定义词类（以词库方式添加），并且可以指定词类之间的优先级。
- (2) 常见问题检索可以返回 Top- $k$  个回答。
- (3) 增加三元组抽取功能。
- (4) 对于序列标注模型（中文分词、词性标注和命名识别等）进行了性能优化。
- (5) 修复了上一版中被发现的错误。

4.1较4.0版本，本次主要增加以下功能：

- (1) 增加了词典动态加载和加密功能。
- (2) 新增了依存句法分析功能。
- (3) 增加了篇章级中文分词和词性标注功能，一定程度上避免在一篇文章中相同实体在不同

出现处识别不一致情况。

- (4) 新增文本和句子转换特征向量的功能，可用于分类和聚类应用。
- (5) 增加了文本聚类和热点新闻发现功能。
- (6) 新增基于随机森林的文本分类功能。
- (7) 增加了文本信息抽取功能。
- (8) 新增文本、语句和词汇情感分析功能。
- (9) 增加了文本摘要、关键短语和关键词抽取功能。
- (10) 优化了对话管理模块，以及修复已知的错误。

4.0较3.0版本，本次主要增加以下功能：

- (1) 上下文相关问答。上下文相关分为两种情况：一种情况是处理类似上一句问“今天北京天气如何？”，然后下一句追问“上海呢？”（默认开启）；另一种情况是根据对话主题展开、转换和递进给出合适的回答，开通此功能需要提供语料和定制。
- (2) 多轮对话。处理类似订购机票的场景。不同场景可以根据情况切换，并且可以插入其他问答。开通此功能需要针对不同场景进行定制。
- (3) 海量自定义问答对的高效检索。检索匹配时考虑同义词替换，根据发音相似性纠正可能的错误，并且支持一次提问中包括多个问题的情况。
- (4) 可以为每一位用户定义各自的上下文信息。
- (5) 客户端用于系统演示和调试，支持本地或服务器快速部署。
- (6) 问答过程中禁用词检测功能。
- (7) 修复了之前版本中已知的一些错误。

3.0较2.2版本，本次主要增加以下功能：

- (1) 融合了任务导向（Task-oriented）的对话问答和聊天机器人（Chatbot）的对话系统。
- (2) 可以与RDF表示的知识库（类似知识图谱）对接，用于支持对知识库的访问，以及其上的知识问答；
- (3) 增加预定义问题优先处理机制（通过给出优先处理的问题和回答对文件），以满足一些个性化问答需求。
- (4) 增加基于发音相似性的匹配功能，部分克服语音识别系统的识别不准确问题，该功能用于知识库检索和预定义问题的匹配。
- (5) 增加了简单算术、诗词朗读、术语解读、菜谱查询、家电控制等场景，并且完善了出行路线外部数据对接。
- (6) 增加了金融、疾病、药物、动物、植物、化学等专业词汇。
- (7) 更新了用于中文分词、词性分析和命名识别的网络参数（原词性分析中“形容词”标签从“JJ”变成“A”）。
- (8) 修复了之前版本中已知的一些错误。

2.2较2.1版本，本次主要增加以下功能：

- (1) 增加了任务导向（Task-oriented）的对话问答功能，可用于智能助理、智能客服、问答机器人等应用。
- (2) 增加了语音接口（语音识别与合成），可以用语音的方式进行对话问答。
- (3) 自带包括天气预报、互联网信息检索、出行路线查询等十个应用场景，并且可在所提供的框架下方便地新增应用场景。

2.1较2.0版本，本次主要增加以下功能：

- (1) 增加了带动态 $k$ -max池化的卷积神经网络的句子语义表示模型，可以根据其产生的句子级语义表示完成各种分类任务。
- (2) 针对同时处理多个事件时，条件随机场CRF（Conditional Random Fields）模型处理速度慢的问题，对CRF模型的训练和解码算法进行优化，使其处理速度与单一事件相同。
- (3) 增加带转移矩阵的双向LSTM模型，能够用于序列标注和语义分析任务（在2.2版中删除）。
- (4) 增加时期类型的识别和支持自定义领域词汇。
- (5) 用于中文分词、命名识别和词性标注的神经网络使用经过预训练的字向量进行重新训练，性能进一步提高。字向量采用最终研究成果的适用于中文的预训练方法。
- (6) 修正了之前版本中的一些Bug。

2.0较1.0版本，本次主要增加以下功能：

- (1) 支持根据用户所准备的样本进行模型的重新训练和调整训练（在已经大量样本训练的基础上根据新样本进行精调）。
- (2) 对网络结构和学习处法进行优化，加速模型训练时间。
- (3) 基于条件随机场CRF（Conditional Random Fields）的语义分析模块。
- (4) 支持自定义词汇和文本规范化（俚语替换）功能。
- (5) 完善各模块之间的整合，提升各种任务的准确性。

## 2. 部署与开发

### 2.1. 系统部署

使用工具进行项目开发的基本流程如下：

- (1) 获取并安装Java JDK1.7版本。
- (2) 获取DeepNLP4.2.zip压缩包，并解压。
- (3) 将解压后的“DeepNLP4.2”目录下的目录及文件全部拷贝到工程。
- (4) 将工程中“lib”目录下所有的包导入工程。
- (5) 根据本手册本节和第3节各种功能的详细使用说明进行应用开发。
- (6) 申请百度的天气预报和路径查询开发者账号，并且更改“config/conf/External.properties”中的API访问密串。请尽快申请和更改，不然会因为每日访问次数限制产生异常。

**注意：**使用本工具时，所有文件都应采用 UTF-8 编码，集成开发环境和编辑工具也应使用 UTF-8 编码，不然可能出现因编码不一致所导致的错误。

为了方法进行系统部署，我们调整了目录结构，所有文件路径都采用相对路径。此外，除了 lib 目录仍在开发项目根目录下，其他目录（随同文件）都转入 config 目录。

## 3. 使用说明

### 3.1. 中文分词

#### 3.1.1. 使用方法

使用中文分词功能的样例代码见“cn.edu.fudan.flow”包下的“WordSegmentorStart.java”，该功能已经集成了自定义词汇、俚语替换和命名识别功能。

自定义词汇和俚语替换通过配置“source”目录中以下18个文件实现：

自定义词汇类别	文件名	备注
人名	person.utf8	每个自定义词条一行

组织机构	organization.utf8	同上
地名	location.utf8	同上
成语	idiom.uft8	同上
食物	food.utf8	同上
著名景点	scene.utf8	同上
设备	device.utf8	同上
动物	animal.utf8	同上
植物	plant.utf8	同上
化学品	chemistry.utf8	同上
常见病症	disease.uft8	同上
金融财务术语	finance.utf8	同上
天气术语	weather.utf8	同上
常见药品	medicine.utf8	同上
常用应用软件	app.utf8	同上
政府权威机构	authorities.utf8	同上
上市公司	listedcompany.utf8	同上
新闻媒体	media.utf8	同上
股票交易代码	stockcode.utf8	同上
作品	title.uft8	同上，可包括著作、影视作品等
领域词汇	domain.utf8	同上，应用领域相关的专用名词
专业术语	terminology.utf8	同上，相关专业的专用名词（较专业或学术性的词汇）
俚语	slang.utf8	分为两列，前一列为俚语、每二列为规范用语，两者以空格分开。一组一行，每一行最前面加“#”号，会忽略该行，删除后恢复。

命名识别模块会对以下类型进行识别：

类别	英文标签	备注
俚语	SLANG	来自slang.utf8文件所定义的词（自动替换，结果中不显示）
邮箱地址	EMAL	
网址	URL	
日期	DATE	
百分比	PERCENT	
度量	MEASURE	细分为：LENGTH（长度）、WEIGHT（重量）、SQUARE（面积）、VOLUMN（体积）、TEMPERATURE（温度）。
时间	TIME	
时期	PERIOD	
货币	CURRENCY	
股票交易代码	STOCKCODE	
手机号码	CELLPHONE	
座机号码	LANDLINE	
领域词汇	DOMAIN	来自domain.uft8文件所定义的词
专业术语	TERMINOLOGY	来自terminology.uft8文件所定义的词
成语	IDIOM	来自idiom.uft8文件所定义的词
常见药品	MEDICINE	来自medicine.uft8文件所定义的词
常见疾病	DISEASE	来自disease.uft8文件所定义的词
化学品	CHEMISTRY	来自chemistry.uft8文件所定义的词
金融财务术语	FINANCE	来自finance.uft8文件所定义的词
动物	ANIMAL	来自animal.uft8文件所定义的词
植物	PLANT	来自plant.uft8文件所定义的词
常用应用软件	APP	来自app.uft8文件所定义的词

天气术语	WEATHER	来自weather.utf8文件所定义的词
食物	FOOD	来自food.utf8文件所定义的词
著名景点	SCENE	来自scene.utf8文件所定义的词
作品	TITLE	来自title.utf8文件所定义的词
设备	DEVICE	来自device.utf8文件所定义的词
政府权威机构	AUTHORITIES	来自authorities.utf8文件所定义的词
上市公司	LISTEDCOMPANY	来自listedcompany.utf8文件所定义的词
新闻媒体	MEDIA	来自media.utf8文件所定义的词
组织机构	ORGANIZATION	合并命名识别模块结果与来自organization.utf8文件所定义的词
地名	LOCATION	合并命名识别模块结果与来自location.utf8文件所定义的词
人名	PERSON	合并命名识别模块结果与来自person.utf8文件所定义的词
外文字符	FOREIGN	
数字	DIGIT	
标点	PUNCUTATION	

**注意：**当某一词汇同属于两种类型时，按上表所示先后顺序的优先级确定。如：同时属于 **CELLPHONE** 和 **DIGIT**，则识别为 **CELLPHONE** 类型。如果优先级较低类型的词汇完全包含优先级高类型的词汇（即前者较后者长且完全覆盖后者），则分词结果按优先级较低类型的词汇为准，但在词汇类型识别输出列表中，两者都会保留。

识别人名、组织机构名、地名的命名识别模块将先于中文分词模块运行，并将识别结果作用于中文分词模块。

### 3.1.2. 配置文件

中文分词的配置文件见“conf”目录下的“WordSegmentor.properties”，包括以下参数：

参数名称	默认值	备注
inputNetworkSettingFile	model/ WindowConvolutionNetwork _seg_d300_w5_f1_CSV_v2. model	指向用于中文分词的网络参数文件。默认文件的网络参数已经经过大量语料训练，可以实际使用。使用自定义样本重新训练网络后，可以用结果文件进行替换。
isCharacterLevel	true	中文情况设置为true（一般不要修改）
isResultWithTag	false	中文分词设置为false（一般不要修改）
isSegmentation	true	中文分词设置为true（一般不要修改）
isStandard	false	指示是否要对输入语句进行规范化处理（包括全角转半角、繁体转简体、数字和字母统一表示形式）。由于中文分词之前的预处理模块已完成上述工作，所以此处设置为false（一般不要修改）。

中文分词之前会对输入句子进行必要的预处理（包括全角转半角、繁体转简体、数字和字母统一表示形式等），预处理的配置文件见“conf”目录下的“Preprocess.properties”，包括以下参数：

参数名称	默认值	备注
domainFile	source/domain.utf8	指向自定义“领域”词汇的文件路径
terminologyFile	source/terminology.utf8	指向自定义“专业术语”词汇的文件路径
idiomFile	source/idiom.utf8	指向自定义“成语”词汇的文件路径
medicineFile	source/medicine.utf8	指向自定义“常见药品”词汇的文件路径
diseaseFile	source/disease.utf8	指向自定义“常见疾病”词汇的文件路径
chemistryFile	source/chemistry.utf8	指向自定义“化学品”词汇的文件路径
financeFile	source/finance.utf8	指向自定义“金融财务术语”词汇的文件路径

animalFile	source/animal.utf8	指向自定义“动物”词汇的文件路径
plantFile	source/plant.utf8	指向自定义“植物”词汇的文件路径
appFile	source/app.utf8	指向自定义“常用应用软件”词汇的文件路径
foodFile	source/food.utf8	指向自定义“食物名称”词汇的文件路径
weatherFile	source/weather.utf8	指向描述“天气”情况词汇的文件路径
sceneFile	source/scene.utf8	指向自定义“著名景点”词汇的文件路径
titleFile	source/title.utf8	指向自定义“作品”词汇的文件路径
deviceFile	source/device.utf8	指向自定义“设备”词汇的文件路径
slangFile	source/slang.utf8	指向自定义“俚语”词汇的文件路径
authoritiesFile	source/authorities.utf8	指向自定义“政府权威机构”词汇的文件路径
listedcompanyFile	source/listedcompany.utf8	指向自定义“上市公司”词汇的文件路径
mediaFile	source/media.utf8	指向自定义“新闻媒体”词汇的文件路径
organizationFile	source/organization.utf8	指向自定义“组织机构”词汇的文件路径
locationFile	source/location.utf8	指向自定义“地名”词汇的文件路径
personFile	source/person.utf8	指向自定义“人名”词汇的文件路径
nerRecognizerFile	conf/NerRecognizer.properties	指向中文命名识别模块配置文件的文件路径，配置文件相关参数说明详见3.2.2节。
isBIOES	true	指示模型是否采用BIOES标签规范，BIOES规范使用如下的规则：B开头标签表示词的开始字符；I开头表示词的中间字符；E开头表示词的结束字符；S开头表示单独成词的字符；O开头表示与任务无关的字符。如isBIOES设置成“false”，则使用BIO标签规范，即用B替代S，用I替代E。各模块训练样本所采用的标签规范需要与此参数设置一致。默认使用BIOES规范。
isDeleteAllBlank	true	指示是否删除输入句子中的空格。如果设置成“true”，则删除所有中间空格（两个英文单词之间的空格会用特殊字符“□”替换）。如果设置成“false”，所有空格会用特殊字符“□”替换。
isCaseSensitive	true	指示在识别自定义词汇时是否对大小写敏感，如果设置成“true”，输入句子中出现“ibm”，而在source/organization.utf8文件中包含“IBM”词条，则“ibm”不会被识别成公司名，如果设置成“false”，则会被识别成IBM公司的名称。
isDeleteOverlap	true	指示被命名识别模块识别出的词汇列表中是否删除相互重叠的词汇。例如：若设置成“false”，输入语句中包含“猕猴桃”一词，则“猕猴”和“猕猴桃”都会出现的结果列表中；若设置成“true”，则仅出现“猕猴桃”。注意在分词结果按最长匹配原则，即认为“猕猴桃”是词。
isUppercase	true	当isCaseSensitive参数设置被成false时，是否将词表中的词汇和输出中的英文字母全都转化成大写。

### 3.1.3. 增加词汇及类型

通过配置文件增加词汇及其类型。注意：以这种方式增加的自定义词汇都可以加密或非加密的方式增加，因而需要先准备词汇文件，然后使用3.9节中讲述的领域词汇加密方法转换成加密后词汇文件。

通过配置的方式增加词汇及类型的步骤如下：

- (1) 为每一类型的词汇准备词汇文件，该文件每一行为一个领域词汇；
- (2) 使用“corpus”包下的“DictionaryEncryptionStart()”工具将词汇转换成加密词汇文件；

- (3) 配置“conf/CustomerDictionary.utf8”文件，该文件每一行定义一类新增词汇。每一行分为三列，第一列为词汇的类型（如与已有类型名称相同为报错），第二列为该类词汇的词性标签（标签必须是系统所采用的词性标签集合中的一个，否则为报错），第三列为加密词汇文件的路径。注意在文件中较先定义的词汇有较高的识别优先级。

### 3.1.4. 重新或调整训练中文分词网络

重新或调整训练运行“cn.edu.fudan.dnn”包下的“WindowConvolutionNetworkStart.java”，网络训练配置文件见“conf”目录下的“windowConvolutionNetwork.properties”，包括以下参数（注意：尽管以下有些参数的值在特定的情况下会被忽略，但在程序启动时，会被程序读取，所以都需要给出相应的值）：

参数名称	默认值	备注
corpusFile	dataset/segmentation_corpus.utf8	指定训练样本的文件路径。按照默认值所指向文件的格式准备训练样本，每一行包括一个字符与这个字符的标注，每一句之间用一行空格相隔。
tokenFile	conf/characters.utf8	指定中文字符集的文件路径。这个文件一般应包含出现在训练样本中的所有字符。应确保以下字符出现在字符集中：“ <b>⌋</b> ”（句首字符）、“ <b>⌋</b> ”（句尾字符）、“ <b>□</b> ”（未知字符）、英文字母和阿拉伯数字。“conf/characters.utf8”是该文件的一个范本。
labelFile	dataset/segmentation_labels.utf8	指定任务标注集合的文件路径。
isReadEmbedding	false	指定网络是否使用事先准备好的字或词向量进行训练。如果这个值设为“true”时，确保embeddingFile设置成适当的值。
embeddingFile	embedding/Word2Vector.model	指定网络启动时需要读取字或词向量的文件路径（即使用经过训练的字或词向量来进行初始化网络）。如果isReadEmbedding设置成为“false”，则embeddingFile的取值将被忽略。如何正确生成初始化网络字或词向量的“embedding.model”文件见3.1.4节。
isExternalFeature	false	对于字或词向量，是否使用外部特征值。如果这个参数设置为“true”，需要准备外部特征值文件（由externalFeatureFile参数指定储存外部特征值的文件路径），并且tokenFile文件会被忽略，字符集从externalFeatureFile指定的文件中读取。
externalFeatureFile	conf/charactersWithExternalFeature.utf8	指定字或词向量外部特征值文件路径。默认值所指向的文件是该文件的一个范本。
isReadNetworkSetting	false	指定网络参数初值是否从某个经过训练网络的参数中读取。（使用之前训练过的网络参数

		初始化，继续进行训练过程，即调整训练)。如果这个参数设置为“true”，应确保inputNetworkSettingFile设置成适当的值。注意：当这个参数为“true”时，以下参数的值将被忽略：tokenFile、labelFile、isExternalFeature、externalFeatureFile、isReadEmbedding、embeddingFile、featureDimension、internalFeatureDimension、externalFeatureDimension、windowSize和isIgnoreAlphabetNumber。
inputNetworkSettingFile	model/ windowConvolutionNetwork_d100_w3_f1.model	指定网络启动时需要读取网络参数的文件路径。如果isReadNetworkSetting设置成为“false”，则inputNetworkSettingFile的取值将被忽略，网络参数初值将随机产生。
outputNetworkSettingFile	model/ windowConvolutionNetwork_d100_w3_f1.model	指定网络完成训练后，保存网络参数的文件路径。
echoFile	result/ windowConvolutionNetworkEcho.utf8	指定记录训练过程信息的文件路径。
featureDimension	100	指定字或词向量的维度，该值应等于internalFeatureDimension和externalFeatureDimension取值之和，不然程序会报错。
internalFeatureDimension	100	指定字或词向量本身特征维度。
externalFeatureDimension	0	指定字或词向量外部特征维度。
windowSize	3	指定窗口的大小。
featureMap	1	指定卷积层feature map的数量，实验表明，取1时一般就能达到较好的性能。
learningRate	0.05d	指定学习步长。该值越大，学习速度越快。但是取一个较小的值有利于保持网络学习的稳定性。
regularizationRate	0.0001d	指定Regularization参数值。该参数用于防止过拟合。
errorLimit	0.001d	指定期望的误差水平。期望的准确率等于 $(1 - \text{errorLimit})$ 。当网络的标注误差小于设定的值，网络停止训练过程，并且输出相应的网络参数。
learningTimes	100	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，网络停止训练（即使没有达到期望的误差水平）。
isIgnoreAlphabetNumber	true	指定是否忽略不同英文字母和阿拉伯数字的差异。如果这个参数设置为“true”，则所有英文字母的向量表示都相同，所有阿

		拉伯数字的向量表示也都相同。
isAverage	false	指定卷积层产生的多个feature map值之后取平均或求和。当设置为“false”时，进行相加操作，否则求平均值。当featureMap值设置成1时，isAverage取值对网络训练没有影响。

中文分词模块所使用标签及其说明如下表所示：

分词标签	标签说明
B	词的开始字符
I	词的中间字符
E	词的结束字符
S	单独成词的字符

### 3.1.5. 使用字或词向量初始化网络

网络参数的初始值对性能的影响比较大，并且网络参数数量最多的部分是字或词向量。实验结果表明使用经过预训练的字或词向量来初始化网络是提高性能的有效方法。字或词的向量预训练可以使用类似Google的Word2Vector<sup>1</sup>等工具。

中文一般使用和训练字向量，类似英文一般使用和训练词向量。训练好的字向量放置于“embedding”目录下，文件格式见目录下的“Word2Vector.utf8”文件，每一行为一个字符（中文）或词（英文）和其向量表示的各维度值，字或词与每一维度的值之间用空格隔开。运行“cn.edu.fudan.corpus”包下“LookupTableGeneratorStart.java”，会产生符合网络训练初始化要求的“Word2Vector.model”文件。运行“LookupTableGeneratorStart.java”需要配置“conf”目录下的“LookupTableGenerator.properties”文件，该配置文件包括以下参数：

参数名称	默认值	备注
embeddingTextFile	embedding/Word2Vector.utf8	指定保存经过预训练产生字或词向量文件的路径，格式参见默认指向文件。
embeddingFile	embedding/Word2Vector.model	指定转换后满足网络初始化要求的字或词向量文件路径。该文件即为网络训练时embeddingFile参数所指向的文件。
dimension	50	字或词向量的维度，应与预训练所使用的字或词向量维度一致，否则程序会报错。
tokenFile	conf/characters.utf8	指定字符（中文）或词（英文）集合的文件路径。如果该集合含有预训练字或词中没有出现的字或词，则会通过随机初始化补全这些缺失的字或词向量，以确保网络训练不会报错。
isWord2Vector	false	字或词向量文件是否由Word2Vector产生
isChineseCharacter	true	是否是中文字向量

## 3.2. 中文命名识别

### 3.2.1. 使用方法

使用中文命名识别的样例代码见“cn.edu.fudan.flow”包下的“NamedIdentityRecognizerStart.java”，该功能已经集成了自定义词汇和俚语替换功能（相关内容详见3.1.1节）。

中文命名识别模块所使用标签及其说明如下表所示：

<sup>1</sup> <https://code.google.com/archive/p/word2vec/>

命令识别标签	标签说明
PER	人名
ORG	组织机构名
LOC	地名
O	其它

模型采用与中文分词联合标注的方法，即同时识别词及其所属对象。通过采用中文分词标签和命名识别标签相连的方式产生适合于联合标注的标签集合，如：“B\_PER”表示人名的开始字符，其它依此类推。

### 3.2.2. 配置文件

中文命名识别配置文件见“conf”目录下“NerRecognizer.properties”，包括以下参数：

参数名称	默认值	备注
inputNetworkSettingFile	model/ WindowConvolutionNetwork _ner_d300_w5_f1_CSV_v2. model	指向用于中文命名识别的网络参数文件。默认文件的网络参数已经经过大量语料训练，可以实际使用。使用自定义样本重新训练网络后，可用结果文件进行替换。
isCharacterLevel	true	中文情况设置为true（一般不要修改）
isResultWithTag	true	命名识别设置为true（一般不要修改）
isSegmentation	false	命名识别设置为false（一般不要修改）
isStandard	false	指示是否要对输入语句进行规范化处理（包括全角转半角、繁体转简体、数字和字母统一表示形式）。由于中文命名识别之前的预处理模块已完成上述工作，所以此处设置为false（一般不要修改）。

### 3.2.3. 重新或调整训练中文命名识别网络

重新或调整训练运行“cn.edu.fudan.dnn”包下的“WindowConvolutionNetworkStart.java”，网络训练配置文件见“conf”目录下的“windowConvolutionNetwork.properties”，包括类似3.1.3节所介绍的参数。其中：“corpusFile”应指向为中文命名识别任务所准备的训练样本文件，参考格式见“dataset”目录下的“ner\_corpus.utf8”文件，“labelFile”则应指向中文命名识别任务所使用标签集合文件，参考格式见“dataset”目录下的“ner\_labels.utf8”文件。网络训练时使用预训练字或词向量初始化方法见3.1.4节所述。

## 3.3. 中文词性标注

### 3.3.1. 使用方法

使用中文词性标注的样例代码见“cn.edu.fudan.flow”包下的“PosTaggerStart.java”，该功能已经集成了自定义词汇、俚语替换和命名识别功能（相关内容详见3.1.1节）。

中文词性标注模块所使用标签及其说明如下表所示：

词性标签	词类说明
D	数词
IJ	语气词或叹词
A	形容词（注意3.0之前版本为JJ）
C	连词
M	量词
NT	时间名词

NR	专有名词
FW	外文词
I	成语或习语
NN	一般名词
U	助词
ON	拟声词
PU	标点
AD	副词
LC	方位词
PN	代词
V	动词
P	介词
X	其它

模型采用与中文分词联合标注的方法，即同时识别词及其所属词性。通过采用中文分词标签和词性标签相连的方式产生适合于联合标注的标签集合，如：“B\_D”表示数词的开始字符，其它依此类推。

### 3.3.2. 配置文件

中文词性标注功能的配置文件见“conf”目录下的“PosTagger.properties”，包括以下参数：

参数名称	默认值	备注
inputNetworkSettingFile	model/ WindowConvolutionNetwork _pos_d300_w5_f1_CSV_v2. model	指向用于中文词性标注的网络参数文件。默认文件的网络参数已经经过大量语料训练，可以实际使用。使用自定义样本重新训练网络后，可用结果文件进行替换。
isCharacterLevel	true	中文情况设置为true（一般不要修改）
isResultWithTag	false	词性标注设置为false（一般不要修改）
isSegmentation	false	词性标注设置为false（一般不要修改）
isStandard	false	指示是否要对输入语句进行规范化处理（包括全角转半角、繁体转简体、数字和字母统一表示形式）。由于中文词性标注之前的预处理模块已完成上述工作，所以此处设置为false（一般不要修改）。

### 3.3.3. 重新或调整训练中文词性标注网络

重新或调整训练运行“cn.edu.fudan.dnn”包下的“WindowConvolutionNetworkStart.java”，网络训练配置文件见“conf”目录下的“windowConvolutionNetwork.properties”，包括类似3.1.3节所介绍的参数。其中：“corpusFile”应指向为中文词性标注任务所准备的训练样本文件，参考格式见“dataset”目录下“postagging\_corpus.utf8”文件，“labelFile”则应指向中文词性标注任务所用标签集合文件，参考格式见“dataset”目录下的“postagging\_labels.utf8”文件。网络训练时使用预训练字或词向量初始化方法见3.1.4节所述。

## 3.4. 句子语义表示模型

本工具采用带动态  $k$ -max 池化的卷积神经网络模型来产生句子的语义表示，然后在此表示的基础上来完成分类等任务。

### 3.4.1. 使用方法

使用句子分类的样例代码见“cn.edu.fudan.sentence”包下的“ConvolutionalSentenceModelDecoderStart.java”程序。对于输入的句子，可以判断句子所属的类别。

### 3.4.2. 训练样本准备

重新训练句子（或短文）分类模型，首先需要准备正负样本，负样本是指应用不感兴趣的句子。正样本的格式见“sentence”目录下的“sample\_sentence.utf8”文件，样本每一行包括句子类型标签和句子内容，标签与句子内容之间用空格隔开。负样本的格式见“sentence”目录下的“sample\_negative.utf8”，负样本不需要标注类型标签（系统默认为“E\_UNKNOWN”）。

运行“cn.edu.fudan.corpus”包下“SentencePrepareStart.java”程序，该程序会产生句子分类模型训练所需要的文件：包括正负样本的文件（顺序随机打乱）、字符表、标签文件。

**注意：**如果提供负样本，负样本数量应与每个类别样本数量大致相当，可以略多一些。高质量的负样本应该是与其他类别语义相近，但也不属于其他类别的句子。

### 3.4.3. 句子模型训练

运行“cn.edu.fudan.sentence”包下“ConvolutionalSentenceModelStart.java”程序，配置文件见“sentence”目录下的“sentence.properties”，包括以下参数：

参数名称	默认值	备注
corpusFile	sentence/sentence_corpus.utf8	指定训练样本的文件路径。该文件由样本准备程序自动生成，见3.4.2节。
tokenFile	sentence/token.utf8	指定训练样本中出现字符的字符表。该文件由样本准备程序自动生成，见3.4.2节。
labelFile	sentence/classificationLabel.utf8	指定任务标签集合的文件路径。该文件由样本准备程序自动生成，见3.4.2节。
isReadEmbedding	false	指定网络是否使用事先准备好的字或词向量进行训练。如果这个值设为“true”时，确保embeddingFile设置成适当的值。
embeddingFile	embedding/Word2Vector.model	指定网络启动时需要读取字或词向量的文件路径（即使用经过训练的字或词向量来进行初始化网络）。如果isReadEmbedding设置成为“false”，则embeddingFile的取值将被忽略。如何正确生成初始化网络字或词向量的“embedding.model”文件见3.1.4节。
isExternalFeature	false	对于字或词向量，是否使用外部特征值。如果这个参数设置为“true”，需要准备外部特征值文件（由externalFeatureFile参数指定储存外部特征值的文件路径），并且tokenFile文件会被忽略，字符集从externalFeatureFile指定的文件中读取。
externalFeatureFile	conf/	指定字或词向量外部特征值文

	charactersWithExternalFeature.utf8	件路径。默认值所指向的文件是该文件的一个范本。
isReadNetworkSetting	false	指定网络参数初值是否从某个经过训练网络的参数中读取。 (使用之前训练过的网络参数初始化, 继续进行训练过程, 即调整训练)。如果这个参数设置为“true”, 应确保inputNetworkSettingFile设置成适当的值。注意: 当这个参数为“true”时, 以下参数的值将被忽略: tokenFile、labelFile、isExternalFeature、externalFeatureFile、isReadEmbedding、embeddingFile、featureDimension、internalFeatureDimension、externalFeatureDimension、windowSize、featureMap、numberOfLayer、numberOfTopK和isIgnoreAlphabetNumber。
inputNetworkSettingFile	model/sentenceModel.model	指定网络启动时需要读取网络参数的文件路径。如果isReadNetworkSetting为“false”, 则inputNetworkSettingFile的取值将被忽略, 网络参数初值将随机产生。
outputNetworkSettingFile	model/sentenceModel.model	指定网络完成训练后, 保存网络参数的文件路径。
echoFile	sentence/sentenceModelEcho.utf8	指定记录训练过程信息的文件路径。
featureDimension	60	指定字或词向量的维度, 该值应等于internalFeatureDimension和externalFeatureDimension取值之和, 不然程序会报错。 这个值最好是2的某个幂的值, 如: 32、64、128等。
internalFeatureDimension	60	指定字或词向量本身特征维度。
externalFeatureDimension	0	指定字或词向量外部特征维度。
numberOfLayer	2	网络层数, 一般取2已经足够。
windowSize	5/3	指定各层网络的窗口大小, 之间用“/”隔开。
featureMap	2/3	指定各层网络卷积层feature map的数量, 之间用“/”隔开。
numberOfTopK	5	最后一层网络的k-max采样的k值大小。
learningRate	0.05	指定学习步长。该值越大, 学习速度越快。但是取一个较小的值有利于保持网络学习的稳定性。
regularizationRate	0.0001	指定Regularization参数值。该参数用于防止过拟合。
errorLimit	0.001	指定期望的误差水平。期望的准确率等于(1 - errorLimit)。当网络的标注误差小于设定的值, 网络停止训练过程, 并且输出相应的网络参数。

learningTimes	1000	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，网络停止训练（即使没有达到期望的误差水平）。
isIgnoreAlphabetNumber	true	指定是否忽略不同英文字母和阿拉伯数字的差异。如果这个参数设置为“true”，则所有英文字母的向量表示都相同，所有阿拉伯数字的向量表示也都相同。

### 3.5. 基于CRF的中文语义分析

基于CRF的中文语义分析分为仅处理单事件和同时处理多事件两个版本，对多事件语义分析需要句子分类作为预处理（即过滤不感兴趣的句子，并且对可能描述某个事件的句子判断其描述事件的类型），句子分类的预处理采用3.4节所述模型。使用单事件语义分析也要确保输入的句子都是描述目标事件的语句。以下先介绍单事件语义分析的训练和使用过程，然后说明进行多事件任务的使用方法。

#### 3.5.1. 使用方法

使用基于CRF模型的中文单事件语义分析的样例代码见“cn.edu.fudan.flow”包下的“CRFSemanticAnalyzerStart.java”程序。语义分析利用中文分词、命名识别和词性标注的结果，分析出输入句子的事件类型和关键的属性值对。

运行“CRFSemanticAnalyzerStart.java”涉及以下参数（在该类中直接设置）：

参数名称	默认值	备注
preprocessFile	conf/Preprocess.properties	指向预处理模块的配置文件路径，该配置文件参数详见3.1.2节
posTaggerFile	conf/PosTagger.properties	指向中文词性标注模块的配置文件路径，该配置文件参数详见3.3.2节
confFile	crf/crfDecoder.properties	指向中文语义分析模块的配置文件路径，该配置文件参数详见3.5.2节
eventKeywordFile	crf/eventKeywords.utf8	指向事件类型描述标签与中文解释之间对应关系的配置文件路径。配置文件中，每一行配置一个事件类型，事件类型描述标签与中文解释之间用空格隔开。
attributeKeywords	crf/attributeKeywords.utf8	指向关键属性描述标签与中文解释之间对应关系的配置文件路径。配置文件中，每一行配置一个关键属性，关键属性描述标签与中文解释之间用空格隔开。

#### 3.5.2. 配置文件

中文语义分析功能的配置文件见“crf/crfDecoder.properties”，包括以下参数：

参数名称	默认值	备注
parameterFile	model/crf_demo.model	指向用于中文语义分析模型的文件。需要根据语义分析需求，使用自定义样本重新训练模型（详见3.5.3节），并将结果文件替换默认文件后才能实际使用。
templateFile	crf/template.utf8	指向条件随机场模型使用的特征模板文件的路径

labelFile	crf/semanticlabels_demo.utf8	指向中文语义分析任务所使用标签集合文件路径
is2gram	true	指示条件随机场模型是否使用前后标签转移特征，应对模型训练时的设置一致。
numberOfColumn	5	表示每一个词汇及其不同模块分析结果标签的总数，默认的取值为5（一般不要修改），包括：中文词汇（数字、网址、时间等用英文标签替换）、语义标签、词性标签、命名识别标签、中文原词形式。表示格式类似“crf”目录下“sample_demo.utf8”文件所示。

### 3.5.3. 训练基于CRF的中文语义分析模型

#### 3.5.3.1. 准备训练样本

运行“cn.edu.fudan.corpus”包下“SemanticCorpusPrepareStart.java”程序，该程序涉及以下参数（在该类中直接设置）：

参数名称	默认值	备注
preprocessFile	conf/Preprocess.properties	指向预处理模块的配置文件路径，该配置文件参数详见3.1.2节
posTaggerFile	conf/PosTagger.properties	指向中文词性标注模块的配置文件路径，该配置文件参数详见3.3.2节
sentenceFile	dataset/single_event_sentence.utf8或dataset/multiple_event_sentence.utf8	指向存储将作为语义分析模型训练样本原始语句的文件路径。对于单事件，该文件包含描述事件的句子，每行一句。对于多事件，每行句子前附加事件类型标签（标签的集合应与句子分类模型使用的一致），事件标签和句子之前用空格隔开。
outputFile	dataset/single_event_sample.utf8或dataset/multiple_event_sample.utf8	指向保存原始语句转化成训练样本格式的文件路径。格式为一个词汇一行，每行共5列，依次分别表示中文词本身（数字、网址、时间等用英文标签替换）、语义标签（此时统一为“O”，表示与任务无关内容，之后需要根据任务进行人工标注）、词性标签、命名识别标签、中文原词形式。每一句样本之间由一空行隔开。
isMultiEvent	true或false	多事件时为true，多事件时为false。
hasHeadingEventType	true或false	多事件时为true，多事件时为false。

对转化成训练样本格式的文件（outputFile所指向的文件）进行语义标注，即使用属性标签修改文件中第二列的语义标签。注意：样本文件最后需空两行。

#### 3.5.3.2. 准备语义标签集合

将语义标注采用的所有标签汇总保存在“crf”目录下“semanticlabels\_demo.utf8”文件中，以每一个语义标签一行的格式存储（包括“O”标签）。

在进行语义标注前，请先仔细阅读“corpus/multiple\_event\_robot.utf8”所给出的样例，从中了解标注的基本原则，一般应遵循以下原则：

- (1)样本应覆盖场景中有代表性的表达，但也不要包括一些极少见的句式；
- (2)每个属性的值应是语义相同或相近的，不应有较大差异；

(3)每种语义标签应覆盖连续的一段词汇或短语，中间不能中断；

(4)标注应体现最小原则，提取的属性值只要能够从中了解其表达的意思，其周围附属的词汇就标成“O”；

(5)样本中同一属性的值应出现各种可能的类型和组合。比如地点类型，样本中不要仅包含几个的城市名称，应尽量增加可能出现的描述地点的词汇和短语。

**注意：在样本数量较少的情况下，语义模块训练不收敛大多数是因为错标和漏标等标注不一致造成的。在大样本下，训练结束时有少量错误并不影响使用。训练过程中会显示当前模型仍然标错的词汇，从中可以观察可能的标注错误。**

### 3.5.3.3. 确定模型的特征模板

编辑“crf”目录下的“template.utf8”文件，该文件定义了依据哪些特征进行语义分析。具体内容如下所示（可以使用所提供的默认模板）：

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]
U06:%x[-1,0]/%x[0,0]
U07:%x[-1,0]/%x[1,0]
U08:%x[0,0]/%x[1,0]
U09:%x[1,0]/%x[2,0]
U10:%x[0,0]/%x[0,2]
U11:%x[0,0]/%x[0,3]
U12:%x[0,0]/%x[-1,2]/%x[0,2]
U13:%x[0,0]/%x[0,2]/%x[1,2]
# Bigram
B00:%x[-2,0]
B01:%x[-1,0]
B02:%x[0,0]
B03:%x[1,0]
B04:%x[2,0]
B05:%x[-1,0]/%x[0,0]
B06:%x[-1,0]/%x[1,0]
B07:%x[0,0]/%x[1,0]
B08:%x[0,0]/%x[0,2]
B09:%x[0,0]/%x[0,3]
```

特征模板共有两类，分别以“U”和“B”开头。其中：“U”类特征模板所产生的特征仅考虑当前的语义标签，而“B”类特征模板所产生的特征考虑语义标签之间的转移关系。类似“[-1, 2]”表示利用哪些特征值来对当前的语义标签进行预测，其中“-1”（相对行数）表示利用前一个词的相关特征值，“2”表示（相对列数）哪一列的特征值，“[-1, 2]”则表示前一个词的词性标签。类似“[-1, 2]”可进行自由组合，例如：“%x[0,0]/%x[0,2]/%x[1,2]”，表示联合使用当前词汇、当前词汇词性和后一个词汇词性对当前的语义标签进行预测。

### 3.5.3.4. 模型训练

运行“cn.edu.fudan.crf”包下的“ConditionalRandomFieldStart.java”程序，其配置文件见“crf”目录下的“crf.properties”，包括以下参数：

参数名称	默认值	备注
corpusFile	crf/sample_demo.utf8	指向语义分析训练样本的文件路径。训练样本准备见3.5.3.1节
parameterFile	model/crf_demo.model	指定模型完成训练之后，保存模型参数的文件路径。
templateFile	crf/template.utf8	指向条件随机场模型使用的特征模板文件的路径
labelFile	crf/semanticlabels_demo.utf8	指向中文语义分析任务所使用标签集合文件路径。
numberOfColumn	5	表示每一个词汇及其不同模块分析结果标签的总数，默认的取值为5（一般不要修改），包括：中文词汇（数字、网址、时间等用英文标签替换）、语义标签、词性标签、命名识别标签、中文原词形式。表示格式类似“crf”目录下“sample.utf8”文件所示。
learningTimes	1000	指定最大的迭代次数，即扫描整个训练样本的次数。当迭代次数超过该值，模型停止训练（即使没有达到期望的误差水平）。
errorLimit	0.0001	指定期望的误差水平。期望的准确率等于 $(1 - \text{errorLimit})$ 。当模型的标注误差小于设定的值，模型停止训练过程，并且输出相应的模型参数。
is2gram	true	指示条件随机场模型是否使用前后标签转移特征。

### 3.5.4. 基于CRF的多事件中文语义分析

#### 3.5.4.1. 准备训练样本

见3.5.3.1节，对转化成训练样本格式的文件进行语义标注，与单事件不同的是，每一个样本前行需要增加事件类型标签。样例文件如“crf/sample\_multiple\_demo.utf8”所示。

#### 3.5.4.2. 准备语义标签文件

准备语义标签文件，其样例文件如“crf/semanticlabels\_multiple\_demo.utf8”所示，与单事件语义标签不同在于：不同事件的标签之间用一空行隔开，同一事件的语义标签以事件类型标签开始，然后列出所有该事件所有的属性标签。

#### 3.5.4.3. 准备语义CRF的特征模板

参见3.5.3.3节说明。

#### 3.5.4.4. 模型训练

运行“cn.edu.fudan.crf”包下的“ConditionalRandomFieldMultipleEventStart.java”程序，其配置文件见“crf”目录下的“crf\_multiples.properties”，参数作用基本与单事件相同，其说明见3.5.3.4节。注意：如果模型一般都会正常收敛，如发现不收敛，先检查语义标签文件是否正确完整，然后检查训练样本标注是否一致。

### 3.5.4.5. 训练描述事件的句子分类模型

见3.4节说明。

### 3.5.4.6. 模型使用

使用基于CRF模型的中文单事件语义分析的样例代码见“cn.edu.fudan.flow”包下的“CRFSemanticAnalyzerMultipleEventStart.java”程序。

## 3.6. 对话问答

本工具融合任务导向（Task-oriented）的对话问答和聊天机器人（Chatbot），即能支持包括简单算术、诗词朗读、术语解读、菜谱查询、家电控制、机器人运动控制、天气预报、互联网查询、出行路线查询、备忘提醒、时间日期询问等场景，还可以支持开放式问题回答。4.0版本开始支持多轮对话和上下文相关问答（需要订制）。工具可以与RDF表示的知识库（类似知识图谱）对接，用于支持对知识库的访问，以及其上的知识问答，增加预定义问题优先处理机制（通过给出优先处理的问题和回答对文件），以满足个性化问答需求。内制了基于发音相似性的匹配功能，部分克服语音识别系统的识别不准确问题，该功能用于知识库检索和预定义问题的匹配。工具可用于智能助理、智能客服、问答机器人等应用开发。带有语音接口（语音识别与合成），可以用语音方式进行对话问答。

### 3.6.1. 系统部署和使用

使用客户端进行系统演示和调试运行cn.edu.fudan.gui.ClientStart.java，支持本地（默认）或服务器两种部署方式。客户端启动之后如图4所示。

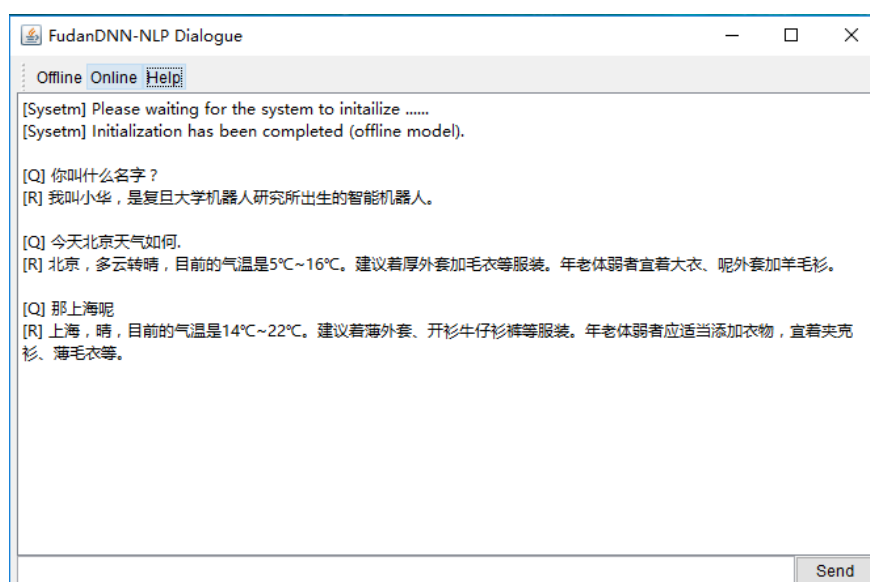


图 4. 对话系统调试界面

如需在服务器端部署对话系统，然后采用客户端访问。首先在服务器部署DeepNLP然后运行“cn.edu.fudan.server”包下“ServerStart.java”程序。然后在客户端的上方点击“Online”按钮，并且根据提示输入服务器的IP地址。可点击“Offline”按钮回到本地系统。

语音对话的情况下，运行“cn.edu.fudan.dialog”包下“SpeechDialogStart.java”程序。该程序调用了科大讯飞的语音识别和语音合成功能。

使用语音对话功能之前，请先申请科大讯飞的开发者帐号（否则可以无法正常使用），并且更改“com.iflytek.util”包下“Version.java”中“版本号”和“开发者 id 号”，同时替换工程根目录下 libmsc32.so、libmsc64.so、msc32.dll、msc64.dll 文件。目前 Apple 计算机不能使用语音功能。

不通过客户端进行手工测试的情况下，运行“cn.edu.fudan.dialog”包下“DialogStart.java”程序。程序运行时会先输出若干个例子，然后出前“问题：”提示符，在此之后输入问句回车后，输出相应的回答。如果输入“exit”，则程序停止。

目前工具自带包括基于自定义问答、简单算术、诗词朗读、术语解读、菜谱查询、家电控制、机器人运动控制、天气预报、互联网查询、出行路线查询、备忘提醒、时间日期询问、寒暄、安防功能（温度、湿度、有害气体、巡逻、监控）、机器人移动控制指令、控制应用软件操作指令、介绍功能、应对未知事件十余个应用场景。

对于“天气预报”和“路线查询”目前是使用百度的 API，在 conf 目录的 External.properties 文件保存访问调用这两个 API 的开发者 key 值，可以根据需要替换成自己申请的 key 值串，以缩短访问的延时。

### 上下文相关问答

上下文相关分为两种情况：一种情况是处理类似上一句问“今天北京天气如何？”，然后下一句追问“上海呢？”（默认开启）；另一种情况是根据对话主题展开、转换和递进给出合适的回答，开通此功能需要提供语料和定制。

第一种情况的问答效果可使用以下句子作为测试：

#### (1) 天气预报

用例：

“北京今天天气如何。”

“上海呢。”

#### (2) 出行路线

用例：

“从上海到杭州怎么走。”

“到北京呢。”

“开车呢。”

#### (3) 菜谱查询

用例：

“糖醋排骨怎么做。”

“那鱼茄子呢。”

第二种情况的问答效果可使用以下句子作为测试：

#### (1) 询问贷款相关问题（递进关系）

“我想贷款。”

“保单如何处理。”

“都有哪些渠道。”

“APP 上怎么操作。”

“如何查看是否需要确认。”

“如何操作。”

(2) 询问理赔相关问题（多个问题）

“我想要理赔。”

“寿险境外可以么。在国外怎么申请。”

“医院有什么要求。”

**多轮对话**

多轮对话。处理类似订购机票场景（开通此功能需要针对不同场景进行定制）。可使用以下两个场景进行测试（不同场景可以根据情况切换，并且可以插入其他问答）。

(1) 激活信用卡

用例：

“我想激活银行卡。”

(2) 取消出险报案

“我要注销报案。”

(3) 结束当前会话

“不需要了。”、“不用了。”、“不必了。”等等

**其他例句**

查看工具的问答效果可使用以下句子作为测试（对于软硬件控制命令，会输出内部指令）：

(1) 自定义问题

用例：

“太阳是什么颜色。”

“鳄鱼为什么吞石块。”

“眼睛为何不怕冷。”

(2) 自我介绍

用例：

“你能做什么。”

“你是谁。”

“你会哪些事。”

(3) 打招呼

用例：

“您好。”

“下午好。”

“早安。”

(4) 在线搜索

用例：

“网上查询自由女神像。”

“查询茉莉花革命。”

“搜索习近平。”

(5) 天气预报

用例：

“今天天气如何。”

“厦门今天会下雨吗。”

“杭州明天天气预报。”

(6) 地图应用

用例：

“到北京大学怎么走。”

“从武汉到昆明坐高铁怎么去。”

“开车从陆家嘴到世贸大厦怎么走。”

(7) 提醒功能

用例：

“每天的上午 7 点半提醒奶奶吃药。”

“每年的 8 月 15 号提醒我外婆的生日。”

“每个月的 3 号提醒儿子回家。”

(8) 时间日期

用例：

“现在几点了。”

“今天是星期一吗。”

“今天是几号。”

(9) 简单算术

用例：

“3 除以 0 是多少。”

“221 乘以-3.14 的结果。”

“19 加上 0 等于多少。”

(10) 术语解释

用例：

“什么是光合作用。”

“互联网是什么意思。”

“人工神经网络是什么。”

注意：目前仅少量术语有相关的知识（大熊猫、人工神经网络、光合作用、互联网、云计算）

(11) 菜谱查询

用例：

“糖醋排骨怎么做。”

“怎么做鱼香茄子。”

“怎么煮东坡肉。”

注意：目前仅少量菜谱有相关的内容（东坡肉、鱼香茄子、糖醋排骨、龙井虾仁、西湖醋鱼）

(12) 诗词朗诵

用例：

“来念首诗听听。”

“来读一首李清照的词吧。”

“给我读一首辛弃疾的破阵子。”

注意：目前仅少量作家和作品有（作家有：李白、杜甫、辛弃疾、李清照、苏轼；作品有：静夜思、将进酒、月下独酌、春望、行路难、念奴娇、江城子、破阵子、声声慢、水调歌头）

(13) 讲故事

用例：

“讲个故事吧。”

“给我讲个白雪公主的故事。”

“给我讲个故事。”

注意：目前仅少量儿童故事（三只小猪、白雪公主、小红帽、灰姑娘、三个和尚）

#### (14) 聊天机器人（由于训练语料质量不佳，目前已经关闭）

用例：

“今天是我生日。”

“我家宝贝真可爱。”

“欢迎使用小米手机。”

“月下花前。”

“江苏舜天战胜了上海申花。”

“中华人民共和国成立之前的中国。”

“看看你的名字在古代是什么职业。”

“有多少人，还记得当年学校的校训的，请举手。”

“你家宝贝真可爱。”

“青椒肉丝味道鲜美。”

“一个人的命运要考虑历史的行程。”

“喵星人个个都是武林高手。”

“慢镜头回放狗狗看到食物的可爱反应。”

“成功就在坚持的下一秒。”

“今天婚礼我们一定会幸福的。”

“老爵爷，向您致敬，没有您，就没有曼联。”

“独在异乡为异客，每逢佳节胖三斤。”

“在最好的角落，坐拥书城。”

“渔舟唱晚，不知乘月几人归。”

### 3.6.2. 处理流程

对话问答功能的处理流程如下：

- (1) 在语音进行识别，转换成文字表示的句子；
- (2) 对句子进行分类，确定其事件类型（详见 3.4 节的句子语义表示模型）；
- (3) 根据句子分类的结果进行语义分析（详见 3.5.4 节的基于 CRF 的多事件中文语义分析）；
- (4) 根据语义分析的结果封装相应的 Frame（本工具采用 Frame 的知识表示模型，新增应用场景时需将“cn.edu.fudan.frame”包下增加相应 Frame 的 JavaBean 定义。Frame 的构造与属性填充由“cn.edu.fudan.dialog”包下的“EventFrameGenerator.java”完成，在新增应用场景该类也需要相应修改）；
- (5) 对于需要调用外部资源的应用（如：天气预报和互联网查询），使用封装好的 Frame 构造相应的查询（查询构造和执行需要实现“cn.edu.fudan.resource”包下相应的类）；
- (6) 通过调用外部资源执行查询，并且返回查询结果；
- (7) 根据查询结果，生成回答句子（由“cn.edu.fudan.generator”包下的“LanguageGenerator.java”完成答句生成，新增应用场景时也要做相应修改）。
- (8) 根据生成的答句，调用语音合成功能，产生语音回答。

注意：存放在 corpus 目录下的 multiple.event.robot.utf8、sentence\_positive.utf8 以及 sentence\_negative.utf8 中给出了目前所有的样本及标注供参考，可以根据应用准备样本重新训练场景分类和语义解析模型只需启动 cn.edu.fudan.dialog 包下的 DialogLearningStart.java。

### 3.6.3. 增量训练

此节说明如果通过增加样本来提高问答的准确性和增加新的应用场景。

- (1) 对于目前不能正确解析的句子，在“corpus/sentence\_positive\_addition.utf8”；  
E\_MAP 到附近的超市怎么走。  
E\_SERACH 帮我查复旦大学的信息。  
E\_WEATHER 明天杭州温度多少度。  
注：在“corpus/sentence\_negative.utf8”中增加负样本（这些句子被分类到 E\_UNKNOWN 类别，由聊天机器人负责产生回答）
- (2) 准备事件分析模型样本。运行“cn.edu.fudan.corpus”包下的“SentencePrepareStart.java”程序（以“corpus/sentence\_robot\_sample.utf8”作为输入）。
- (3) 训练事件分类模型。运行“cn.edu.fudan.sentence”包下“ConvolutionalSentenceModelStart.java”程序（以上一步产生的文件作为输入，详见 3.4 节的句子语义表示模型）。
- (4) 准备新增语义解析样本。运行“cn.edu.fudan.corpus”包下的“SemanticCorpusPrepareStart.java”程序，该程序在“corpus/multiple\_event\_robot\_addition.utf8”中生成增加句子的语义解析样本（以“corpus/sentence\_robot\_addition.utf8”作为输入），对于该样本第二列标注正确的语义标签（详见 3.5.3.1 节的准备训练样本）。
- (5) 如果在生成的语义样本中有影响语义解析准确性的重要分词错误，可在“source”目录下的相应文件中增加自定义词汇（详见 3.1.2 节的配置文件）。
- (6) 将经过语义标注的“corpus/multiple\_event\_robot\_addition.utf8”文件中的句子增加到“corpus/multiple\_event\_robot.utf8”之中。注意“corpus/multiple\_event\_robot.utf8”的文件最后应留两个空行（否则最后一个样本将不会被用于训练）。
- (7) 重新对语义解析模型进行训练，以体现新增的样本。运行“cn.edu.fudan.crf”包下的“ConditionalRandomFieldMultipleEventStart.java”程序。
- (8) 修改“robot\Dialog.properties”配置，使其指向重新训练好的模型参数文件。若新增应用场景，需对“robot”目录下“eventkeywords\_robot.utf8”和“attributekeywords\_robot.utf8”做相应修改。
- (9) 运行“cn.edu.fudan.dialog”包下的“DialogStart.java”程序对新增加的句式进行测试。

### 3.6.4. 自定义问题回答对

自定义问题回答对分成三组，分别在 corpus 目录下 question\_daily.utf8、question\_faq.utf8 和 question\_bank.utf8 配置。这三个文件中问题和回答通过相邻的两个语句来定义，前一句为问题，后一句为回答。每一对问题和回答由一空行隔开（注意：文件末尾需要留出两个空行）。一个问题有多种提问方式或多种可选回答由“||”隔开。如以下格式所示：

你喜欢什么颜色。

蓝色，其实白色或红色也不错。

你是男的还是女的。|| 你是男孩还是女孩。

如果你是男生，我就是女生；如果你是女生，我就男生。|| 人家是女生啊。

你平时喜欢做什么事。

运动、旅游、看书、美食、美人，所有高大上的我都喜欢。

注意：在准备语料时，相同的问题的不同表述，以及同一问题存在多个可选回答时用“||”

隔开。另外，相同问题不要写成多个问答对的方式，要用以“||”隔开方式写在一起。同一问题的多个可选回答也要写在一起。**问答对文件最后要空两行。**

文件 question\_daily.utf8 应主要配置与机器人性格和本身有关的一些寒暄问题，不需要进行训练。寒暄问题的配置文件为 Robot\Question\_daily.properties，各项配置说明如下：

dailyQuestionCorpusFile = corpus/question\_daily.utf8 // 寒暄问题回答对，最后须空两行。

isChooseOneAnswerAtRandom = true // 多个可选问答时是否随机选择一个作为回复。

isNormalizeAnswer = false // 是否对回答进行语言规范化处理，这时逗号和冒号都会以中文“，”和“：”出现。

isReplacePunctuation = false // 是否替换不规范的标号。

文件 question\_faq.utf8 是与应用最相关的经常被问到的问题（Frequency ask question），这些问题将被优先进行匹配。文件 question\_bank.utf8 一般更大的问题集合，但重要性较低，因而优先级较 question\_faq.utf8 文件所定义的问题低。为了加快检索速度和语义匹配的性能，这两个文件所定义的问答对需要进行训练。训练方法如下：

- (1) 从每一个问题中选择一至两个领域关键词（这些词应与问题密切相关，并且一般不会引起分词错误，即如果出现，就单独成词），关键词一般按最细粒度原则入选，比如：“风险投资”中“风险”和“投资”都是词，仅放入“风险”和“投资”，且不必放入“风险投资”。注意：一个字的词原则上下入选。选好的领域关键词可放到任何文件中，并且在配置文件中指明文件位置，但是 question\_faq.utf8 中的关键词应放到 source\domain.utf8 中，以避免可能的分词错误。
- (2) 将 cn.edu.fudan.retrieval.QuestionAnswerPairsPrepareStart.java 中的 corpusFile 变量设置为问答语料文件（即 question\_faq.utf8 或 question\_bank.utf8），同时修改模型参数文件 statisticsFile 的值，并且运行该程序（对于 question\_faq.utf8 和 question\_bank.utf8 各自运行一次，并且生成不同的模型参数文件）。
- (3) 将 Robot 目录下 Question\_faq.properties 或 Question\_bank.properties 文件中的 statisticsFile 属性设置成以上 QuestionAnswerPairsPrepareStart.java 中的 statisticsFile 的值。

以 Robot\Question\_faq.properties 文件的配置为例，各项配置说明如下：

stopwordFile = conf/stopword.utf8 // 停止词列表，停止词是指汉语中大量出现，起辅助作用的词，如：的、地、得。

statisticsFile = model/question\_faq\_statistics.model // FAQ 问答对训练后的模型文件。

domainFile = source/domain.utf8 // FAQ 问答所使用的领域词表。

speechSimilarityThreshold = 0.70 // 默认设置，一般无需修改。

matchThreshold = 0.7 // 值越接近于 1.0，问题匹配的语义相似性要求越高。

possibleThreshold = 0.8 // 值越接近于 1.0，问题候选集产生的语义相似度要求越高。

gapThreshold = 0.7 // 默认设置，一般无需修改。

queryWeight = 0.6 // 默认设置，一般无需修改。

domainWordPlusStd = 1 // 默认设置，一般无需修改。

minQueryLength = 3 // 问题的最小字符长度，低于此长度将不进行检索。

candidateSelectedByCommonWord = false // 在挑选候选回答时，是否仅使用领域词汇。如何问答集合数量较小时可以设置成 true，以提高匹配的准确性。

recoverBySpeechSimilarity = true // 设置是否通过发音相似来匹配领域词汇。

allWordRecoverBySpeechSimilarity = true // 默认设置，一般无需修改。

isIntersection = false // 默认设置，一般无需修改。

domainDescription = domain // 默认设置，一般无需修改。

overlapLimit = 0 // 默认设置，一般无需修改。

```
isChooseOneAnswerAtRandom = true // 多种可选问答时，是否随机选择一个作为回复。  
isOnewayMatch = false // 是否只考虑所列标准提问的匹配度
```

### 3.6.5. 常见问题检索（FAQ）返回Top- $k$ 的回答

对于自定义问答对的检索，可以返回排名前  $k$ （Top- $k$ ）个回答，使用的方法如以下例子所示（详见 cn.edu.fudan.retrieval 包下 AnswerRetrievalStart()程序），其中参数 topK 指定返回可能回答的数量， $k$  个回答按其语义相似性降序排列。

```
query = "市政债券是什么。";  
topKresponse = retrieval.getAnswer(query, topK);  
  
System.out.println("Q: " + query);  
its = topKresponse.iterator();  
nubmer = 0;  
while (its.hasNext()) {  
    nubmer++;  
    System.out.println "[" + nubmer + "]: " + its.next());  
}  
System.out.println("");
```

### 3.6.6. 配置文件

在“robot/Generator.properties”文件设置在寒暄、功能介绍、要求复述问题、未知事件时应答句子的文件。每一个文件可以输入多条答句（每行一条句子），在回答时会随时选择一条作为应答，以模拟人的行为。

在“robot/System.utf8”文件设置问答程序（机器人）相关的参数，配置文件格式如下：

robot 小华（机器人名字）

hasChatbot false // 是否挂接聊天机器人，由于需要根据领域准备合适语料，该功能暂停  
hasDailyQuestion true // 是否有自定义寒暄问答对，通过修改 robot\Question\_daily.properties 文件配置寒暄问答。

hasContextQuestion true // 是否开启上下文相关对话功能

hasMultislots true // 是否开启多轮对话功能

hasFaqQuestion true // 是否应用相关的经常被问到的问答集合，需要提供语料和训练

hasQuestionBank true // 是否有额外的问答集合，优先级低于经常被问到的问答集合

hasInstruction true // 是否处理机器人本体控制指令，包括打开软件、运动控制等

hasForbiddenWordCheck true // 是否进行禁用词检测。通过 source 目录下 complaint.utf8 配置“抱怨或辱骂”的词汇和 forbidden.utf8 配置敏感词汇，定义文件第一句是在问题中出现禁用词后如何回复的句子

shareInformation false // 多轮对话中，不同话题中的信息是否进行共享，如在一个话题中获取了用户的身份证号，再另一话题中是否不必再寻问

chatbot http://116.62.236.184/ // 聊天机器人访问地址

knowledgebaseFile knowledgebase/knowledgebase.owl // 知识库文件存储地址

knowledgebaseID 1495373640 // 知识库文件序列号

matchThreshold 0.85 // 知识库检索时使用的阈值，当超过该阈值，则作为可选结果  
questionThreshold 0.8 // 预定义问题匹配时的阈值，若问题识别率低，可调低此阈值  
maxGapTimes 3 // 在上下文相关问答，连续几个问题为非该上下文时，使该上下文失效

注意：知识库文件序列号为知识库内容文件 knowledgebase/Knowledgebase.owl 中以下一行 `xml:base="http://www.owl-ontologies.com/Ontology1495373640.owl">` 中的数字，如果采用 Protégé 编辑工具添加知识，每一次保存会生成新的数字串，所以需要相应修改。目前知识库数据较少，可以自行根据需要按 OWL/RDF 标准添加相关内容。

在“robot/Dialog.utf8”对话系统相关资源相关的配置，该配置文件格式如下：

```
preprocessFile = conf/Preprocess.properties // 默认设置，一般无需修改
posTaggerFile = conf/PosTagger.properties // 默认设置，一般无需修改
wordSegmentorFile = conf/WordSegmentor.properties // 默认设置，一般无需修改
confFile = robot/crfDecoder_multiple_robot.properties // 默认设置，一般无需修改
eventKeywordFile = robot/eventKeywords_robot.utf8 // 默认设置，一般无需修改
attributeKeywordFile = robot/attributeKeywords_robot.utf8 // 默认设置，一般无需修改
sentenceModelSettingFile = model/sentenceModel_robot.model // 默认设置，一般无需修改
languageGeneratorFile = robot/Generator.properties // 默认设置，一般无需修改。
speechSimilarityEstimatorFile = conf/SpeechSimilarityEstimator.properties // 默认设置，一般无需修改。
dailyQuestionFile = corpus/question_daily.properties // 定义机器人性格和本身有关寒暄问题的配置文件
contextQuestionFile = robot/Question_context.properties // 上下文相关问答配置文件
faqFile = robot/Question_faq.properties // 应用相关的经常被问到的问答的配置文件
questionBankFile = robot/Question_bank.properties // 额外的问答的配置文件，这些问题的优先级低于经常被问到的问答集合
complaintWordFile = source/complaint.utf8 // 抱怨或辱骂词列表
forbiddenWordFile = source/forbidden.utf8 // 敏感词列表
```

目前系统的 question\_faq.utf8 是一些金融方面的常见问题，而 question\_bank.utf8 是类似十万个为什么的问题。

### 3.6.7. 用户上下文信息

可为每一位用户维护各自的上下文信息，通过构建 ClientContext 类实例来实现，用于支持 Web 等应用。用户上下文信息包括：姓名、所在城市、所在位置、搜索引擎（因其他搜索引擎访问受限，目前只能是百度）、上下文主题和多轮对话的信息等。其中上下文主题和多轮对话的信息由系统自动维护。使用详见“cn.edu.fudan.dialog”包下的“DialogStart.java”中的实例。除了存储姓名、所在城市、所在位置、默认搜索引擎等信息外，还有一个类型为 HashMap<String, String>的变量 attributes 可用于存储每一个用户的信息，它是一个 Hash 表，可以根据自定义的 key 值存储相应的 value，其中 key 和 value 都是 String 类型。

## 3.7. 依存句法分析

依存句法是指通过分析将输入的句子表示成一棵依存句法树，树中的结点之间的弧表示句子中各个词语之间的依存关系（修饰关系），也即指出了词语之间在句法上的搭配关系。

**注意：此功能需要付费开通。**

如果输入“上海浦东开发与法制建设同步”，输出以下形式的分析结果。其中：第一列表示词汇在句子中的编号、第二列是句中的词汇、第三列是词性分析结果，第四列就是依存关系，表示该词修饰句中哪一个词、第五列表示依存关系的类型标签。

```

1  上海  NR  2  NMOD
2  浦东  NR  6  NMOD
3  开发  NN  6  NMOD
4  与    C   6  NMOD
5  法制  NN  6  NMOD
6  建设  NN  7  SUB
7  同步  V   0  ROOT

```

依存关系的类型标签说明如下：

依存关系的类型标签	标签说明
P	介词
PMOD	被修饰词为介词
ROOT	虚拟根结点
AMOD	被修饰词为形容词
SUB	主谓依存关系
SBAR	引导各类从句，如定语从句，时间状语从句等
DEP	依赖关系
VC	系动词。“是”和“为”被标记为系动词
PRD	谓语动词
NMOD	被修饰词为名词
VMOD	被修饰词为动词
OBJ	宾语，中心词为谓词或介词

### 3.8. 篇章级中文分词和词性标注

篇章级中文分词和词性标注会对输入篇章中的实体进行统一识别，尽可能避免某些句子中的实体被识别，而另一些句子中相同的实体没有被识别的情况。

**注意：此功能需要付费开通。**

篇章级中文分词运行“cn.edu.fudan.flow”包下“WordSegmentorForParagraphStart.java”。

篇章级词性标注运行“cn.edu.fudan.flow”包下“PreprocessForParagraphStart.java”。

其中：上述两个 java 程序参数 occurrenceLimit 设置同一实体被确认最少的捕获次数。

### 3.9. 领域词典加密

对于需要将所收集整理的领域词汇进行加密的情况，可以使用“cn.edu.fudan.corpus”包下的“DictionaryEncryptionStart.java”对词典进行加密处理。

**注意：此功能需要付费开通。**

相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
dictionaryFile	dictionary/domain.utf8	明文词典的文件路径。文件格式为每一行一个词汇
encryptionFile	source/domain.dic	加密词典的文件的的路径
isCaseSensitive	false	外文字符是否大小写敏感。敏感设置为

		“true”，不敏感则设置为“false”
isUppercase	false	若设置外文字符大小写不敏感，设置统一转换成大写还是小写。如果设置外文字符敏感，此参数的设置不发生作用
isIncrementalWay	false	指定是否以增量方式进行添加加密词典。若为“true”，dictionaryFile所指向文件中的词将与之前的加密词典合并；若为“false”，则替换原来的加密词典文件

### 3.10. 动态增加词汇

为了在不重新启动系统的情况下增加词汇，系统提供了以下两种动态增加词汇的方式。

#### 3.10.1. 单个增加词汇

获取 Preprocess 类的实例，WordSegmentor、PosTagger、NamedIdentityRecognizer、WordSegmentorForParagraph、PosTaggerForParagraph 和 Dialog 类的实例都能通过以下方式获得 Preprocess 类的实例：

```
Preprocess preprocess = wordSegmentor.getPreprocess();
```

然后可以通过 preprocess 调用 addWord(String word, int type)来逐一添加词汇。例如：

```
preprocess.addWord("艾尔伯特", Item.PERSON);
```

```
preprocess.addWord("英国皇家海军研究院应用物理所", Item.ORGANIZATION);
```

```
preprocess.addWord("量子力学", Item.TERMINOLOGY);
```

其中 addWord(String word, int type)方法将返回 boolean 型值，返回值为 true 时表示添加成功，为 false 时表示添加失败，该方法的第二个参数只可取下表所列值之一：

参数值	词汇类别
Item.PERSON	人名
Item.ORGANIZATION	组织机构
Item.LOCATION	地名
Item.IDIOM	成语
Item.FOOD	食物
Item.SCENE	著名景点
Item.DEVICE	设备
Item.ANIMAL	动物
Item.PLANT	植物
Item.CHEMISTRY	化学品
Item.DISEASE	常见病症
Item.FINANCE	金融财务术语
Item.WEATHER	天气术语
Item.MEDICINE	常见药品
Item.APP	常用应用软件
Item.AUTHORITIES	政府权威机构
Item.LISTEDCOMPANY	上市公司
Item.MEDIA	新闻媒体
Item.STOCKCODE	股票交易代码
Item.TITLE	作品

Item.DOMAIN	领域词汇
Item.TERMINOLOGY	专业术语

注意：按上述方法动态添加的词，在系统运行结束后会失效。

### 3.10.2. 批量增加词汇

获取 Preprocess 类的实例，WordSegmentor、PosTagger、NamedIdentityRecognizer、WordSegmentorForParagraph、PosTaggerForParagraph 和 Dialog 类的实例都能通过以下方式获得 Preprocess 类的实例：

```
Preprocess preprocess = wordSegmentor.getPreprocess();
```

然后在“source”目录下批量添加相关文件的词汇，最后通过 preprocess 调用 reload(int type)方法重载所指定类型的词表。例如：

```
preprocess.reload(Item.PERSON);
preprocess.reload(Item.AUTHORITIES);
preprocess.reload(Item.TITLE);
```

reload(int type)方法的参数取值同 3.10.1 节表格所列出的参数值。

## 3.11. 文本或句子生成特征向量

将文本或句子转换成特征向量表示，用于之后的分类或聚类任务。**注意：此功能需要付费开通。**

### 3.11.1. 语料集转换成特征向量表示

给定文本的集合，根据后续任务（文本聚类或分类等）将文本转换成相应的特征向量。运行“cn.edu.fudan.representation”包下的“ArticleToVectorStart.java”程序。参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
corpusFile	N/A	原始样本语料，每篇文章一行。如为分类任务，每篇文章前需有一标签表示其所属的类别，然后用一特殊符号（可自定义）与文章正文分开
domainFile	source/domain_clustering.utf8	领域内特别重要的词，目前为空。可根据需要添加。出现在这个文件中的词汇将比一般词汇在后续的分类和聚类任务起更重要的作用
stopwordFile	conf/stopword.utf8	停止词表。停止词表中的词汇将不会影响所生成的文本特征向量表示
preprocessFile	conf/Preprocess.properties	预处理配置文件
wordSegmentorFile	conf/WordSegmentor.properties	分词配置文件
labelFile	N/A	如果是为分类任务准备特征向量，将在该文件中输出所有的标签集合
segmentedFile	N/A	输出的分词后的样本语料
vectorFile	N/A	输出换成特征向量表示的文件
articleToVectorSettingFile	*.model	输出之后用于其他文章或句子转成相同空间特征向量的参数文件
delimiter		如果为分类任务准备特征向量，样本语料将用该特殊符号分隔标签和正文。如果为

		非分类任务，该特殊符号可用于分隔文章标题与正文
minFrequency	5	构成特征向量的词汇在样本语料中应出现的最少次数
minWordLength	2	构成特征向量的词汇应包含的最少字数
domainWordPlusStd	0	上述“domainFile”中列出领域词汇增加影响的标准差阈值，一般应大于或等于零
truncateMinusStd	-2	构成特征向量的词汇的tf-idf最小标准差阈值，一般应小于或等于零
truncateMinusChi2Std	1	如果为分类任务准备特征向量，构成特征向量的词汇的卡方最小标准差阈值，一般应大于或等于零
hotMinusStd	-1	如果需同时发现热点文章，构成热点文章特征向量词汇的入选减项标准差的阈值，一般应小于或等于零
hotPlusStd	1	如果需同时发现热点文章，构成热点文章特征向量词汇的入选增项标准差的阈值，一般应大于或等于零
similarLimit	0.85	如果需同时发现热点文章，删除重复文章的相似度阈值，系统使用Cosine距离来计算文章的相似度
isFindHotArticle	false;	设置是否同时发现热点文章
isNorm	true	生成的文章特征向量是否除去模长
isClassification	true	设置是否为分类任务准备特征向量。此时每篇文章前需有一标签表示其所属类别，然后使用一特殊符号（可自定义）与文章正文分开
hotArticleFile	N/A	如需同时发现热点文章，输入所发现热点文章的文件

注意：“articleToVectorSettingFile”参数所设置的文件将保存之后用于其他文章或句子转成相同空间特征向量的参数文件。

### 3.11.2. 文章或句子转换成特征向量表示

见“cn.edu.fudan.representation”包下的“ArticleToVectorRepresentationStart.java”的例子。相关的参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
preprocessFile	conf/Preprocess.properties	预处理配置文件
wordSegmentorFile	conf/WordSegmentor.properties	分词配置文件
articleToVectorSettingFile	*.model	上一节语料集转换成特征向量表示步骤所生成的之后用于其他文章或句子转成相同空间特征向量的参数文件

输入文章或句子，程序会返回其特征向量。

### 3.12. 热点新闻发现

详见 3.11 节，并且将“isFindHotArticle”参数设置为“true”，并且正确设置“hotMinusStd、hotPlusStd、similarLimit”等参数。**注意：此功能需要付费开通。**

### 3.13. 文本或句子聚类

文本或句子的聚类过程包括三个主要过程：特征向量生成（包括热点文本抽取）、运行聚类算法和结果显示。聚类算法中包括初始细粒度聚类后的优化，即合并相似度较高的类别。  
**注意：此功能需要付费开通。**

#### 3.13.1. 聚类特征向量生成

详见 3.11 节，并且将“isClassification”参数设置为“false”。

#### 3.13.2. 运行聚类算法

运行“cn.edu.fudan.clustering”包下“SphericalKmeansStart.java”程序。参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
vectorFile	N/A	第一步转换成特征向量表示的文件
prototypeFile	N/A	存储聚类后每一个类的代表特征向量的文件
numberOfCluster	N/A	聚类的类别数量，根据样本语料规模选择合适值。该值应小于聚类的文章数量
clusteringAccuracy	0.9995	聚类算法的停止精度。即当所有样本中该比例的样本所在类别没有发生变动之时，聚类算法停止
isNorm	true	特征向量表示是否已经除去了向量模长
prototypeSelectedBySample	true	最初的代表向量是否从样本中采出。设置为“false”时，初始代表向量将采用随机方法产生
isRefine	true	聚类结束时是否根据聚类结果进行优化，即合并相似度较高的类别
mergeSimilarityLimit	0.75	对聚类结果进行优化时两个类别合并的相似度阈值，系统使用Cosine距离来计算类间相似度

#### 3.13.3. 聚类结果显示

运行“cn.edu.fudan.clustering”包下“InterpretClusterResultStart.java”程序，将输出显示聚类结果的文件，文件每一类之间用一空行隔开，类中文章或句子在一个段落中显示。相关的参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
vectorFile	N/A	第一步转换成特征向量表示的文件
prototypeFile	N/A	存储聚类后每一个类的代表特征向量的文件
corpusFile	N/A	聚类的原始样本语料，每篇文章一行
resultFile	N/A	输出的聚类结果文件
isArticle	true	是否是对文章或句子进行聚类
delimiter		如果是对文章进行聚类，该特殊符号用于分隔文章标题与正文

### 3.13.4. 一键式聚类

运行“cn.edu.fudan.clustering”包下“ArticleSphericalKmeansClusteringStart.java”程序，相关参数见上述三节内容。

## 3.14. 文本或句子分类

文本或句子分类过程包括三个主要过程：特征向量生成、运行分类算法和分类模型使用。系统采用目前准确度较高的随机森林算法作为分类算法。**注意：此功能需要付费开通。**

### 3.14.1. 分类特征向量生成

详见 3.11 节，并且将“isClassification”参数设置为“true”。

### 3.14.2. 运行分类算法

系统采用目前准确度较高的随机森林算法。运行“cn.edu.fudan.classification”包下“RandomForestStart.java”程序，相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
corpusFile	N/A	第一步转换成特征向量表示的文件
randomForestFile	*.model	输出的分类模型参数文件
sampleProportion	-1.0	构建随机森林中每棵树的样本采样比例。若为负数或大于1，则使用算法默认采样比例
featureProportion	-1.0	构建随机森林中每棵树的特征采样比例。若为负数或大于1，则使用算法默认采样比例
forestSize	500	随机森林中树的数量
partitionNumber	5	特征离散化所使用的分段数量
accuracyLimit	0.9	每棵树结点分裂的精度要求。当归为某一结点的样本属于一种类别的比例超过此设定的值，则该结点停止分裂
isSamplingWithReplacement	true	样本采样时是否采用放回采样的方法
isReadIntoMemory	true	如果训练样本可以全部放入内存设置为“true”以加速训练过程，否则需为“false”

### 3.14.3. 分类模型使用

运行“cn.edu.fudan.classification”包下“RandomForestClassifierStart.java”例子，相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
preprocessFile	conf/Preprocess.properties	预处理配置文件
wordSegmentorFile	conf/WordSegmentor.properties	分词配置文件
articleToVectorSettingFile	*.model	第一步所输出用于文章或句子转成相同空间特征向量的参数文件

randomForestFile	*.model	第二步所输出的分类模型参数文件
------------------	---------	-----------------

输入文章或句子, 输出 ClassificationResult 的实例, 通过该实例可得到类别标签及概率。

### 3.15. 文本信息抽取

文本信息抽取过程包括三个主要过程：样本准备、模型训练和信息抽取。**注意：此功能需要付费开通。**

#### 3.15.1. 样本准备

运行“cn.edu.fudan.extraction”包下“InformationExtractionCorpusPrepareStart.java”的程序，相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
inputFile	N/A	样本语料集。每一个样本由若干行组成，前几行表示需要抽取的内容（内容为多值时，采用自定义的特殊符号隔开），最后一行为原文，不同样本之间留一空行
numberOfRow	N/A	一个样本的行数。包括抽取内容和原文
corpusFile	N/A	输出的信息抽取训练样本
preprocessFile	conf/Preprocess.properties	预处理配置文件
nerRecognizerFile	conf/NerRecognizer.properties	命名识别配置文件
posTaggerFile	conf/PosTagger.properties	词性标注配置文件
isRemoveWithKeyword	true	文本中的语句是否用关键词表进行筛选。如果文本较长，并且句子是否包含有需要抽取内容可通过关键词进行定位则设成“true”以加速处理和抽取过程；否则设成“false”
keywordFile	N/A	若句子是否包含有需要抽取内容可通过关键词进行定位，该参数指向保存关键词的文件。该文件每个关键词一行
occurrenceLimit	2	篇章中同一实体被确认的最少捕获次数
itemDelimiter		抽取内容为多值时，采用的特殊分隔符
isNormalized	false	样本语料集是否经过了文本规范化处理
isCaseSensitive	false	英文字符是否大小写敏感
isUppercase	false	如果大小写不敏感，是否统一转换成大写字母

#### 3.15.2. 模型训练

运行“cn.edu.fudan.extraction”包下“InformationExtractionLearningStart.java”的程序，需要配置相应的 properties 参数文件，该文件中参数说明如下：

参数名称	默认值	备注
corpusFile	N/A	第一步所输出的信息抽取训练样本
parameterFile	N/A	输出的信息抽取模型参数文件
templateFile	extraction/template.utf8	特征模板文件
labelFile	N/A	信息抽取任务所使用标签集合文件
is2gram	true	是否使用前后标签转移特征，应对模型训

		训练时的设置一致。
numberOfColumn	5	表示每一个词汇及其不同模块分析结果标签的总数，默认的取值为5（一般不要修改），包括：中文词汇（数字、网址、时间等用英文标签替换）、语义标签、词性标签、命名识别标签、中文原词形式。
learningTimes	50	训练迭代次数
errorLimit	0.0001	误差精度
isLook	false	训练过程中是否显示当前仍错误的标签。从中可以发现可能的标注不一致或错误的情况

### 3.15.3. 信息抽取

运行“cn.edu.fudan.extraction”包下“InformationExtractorStart.java”的程序，相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
preprocessFile	conf/Preprocess.properties	预处理配置文件
nerRecognizerFile	conf/NerRecognizer.properties	命名识别配置文件
posTaggerFile	conf/PosTagger.properties	词性标注配置文件
isRemoveWithKeyword	true	文本中的语句是否用关键词表进行筛选。如果文本较长，并且句子是否包含有需要抽取内容可通过关键词进行定位则设成“true”以加速处理和抽取过程；否则设成“false”
keywordFile	N/A	若句子是否包含有需要抽取内容可通过关键词进行定位，该参数指向保存关键词的文件。该文件每个关键词一行
occurrenceLimit	2	篇章中同一实体被确认的最少捕获次数
occurrenceMin	2	抽取内容被确认所需最少抽取到的次数。如果只会出现一次，则设置为1即可
confFile	extraction/crfDecoder_lawsuit.properties	信息抽取模型配置文件，见此表后说明
eventKeywordFile	extraction/eventKeywords.utf8	事件类型关键词文件
attributeKeywordFile	extraction/attributeKeywords.utf8	英文字符是否大小写敏感
eventType	N/A	自定义抽取文本类型名称

其中上表中的“confFile”文件参数说明如下：

参数名称	默认值	备注
parameterFile	N/A	输出的信息抽取模型参数文件
templateFile	extraction/template.utf8	特征模板文件
labelFile	N/A	信息抽取任务所使用标签集合文件
is2gram	true	是否使用前后标签转移特征，应对模型训练时的设置一致。
numberOfColumn	5	表示每一个词汇及其不同模块分析结果标签的总数，默认的取值为5（一般不要修改），包括：中文词汇（数字、网址、时间等用英文标签替换）、语义标签、词性标签、命名识别标签、中文原词形式。

### 3.16. 情感分析

系统的情感分析模型可以从语料中同时学习到篇章、句子和词汇的情感倾向及其概率或分值。**注意：此功能需要付费开通。**

#### 3.16.1. 模型训练

运行“cn.edu.fudan.sentiment”包下“SentimentLexiconGeneratorStart.java”程序，相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
preprocessFile	conf/Preprocess.properties	预处理配置文件
nerRecognizerFile	conf/NerRecognizer.properties	命名识别配置文件
wordSegmentorFile	conf/WordSegmentor.properties	中文分词配置文件
occurrenceLimit	2	篇章中同一实体被确认的最少捕获次数
stopwordFile	sentiment/stopword_sentiment. utf8	情感分析专用停止词表
adverbFile	sentiment/adverb.utf8	情感分析专用程度副词表
positiveCorpusFile	N/A	正面倾向情感语料文件。一篇文章或一句 句子一行
negativeCorpusFile	N/A	负面倾向情感语料文件。一篇文章或一句 句子一行
corpusFile	N/A	经分词处理后的情感语料文件（合并正负 面倾向情感语料）
sentimentLexiconFile	*.model	训练后的情感词表参数文件
sentimentAnalysisNetwork SettingFile	*.model	训练后的情感分析模型参数文件
isSegmented	false	之前是否对原始情感分析语料经过分词 处理，如果设置为“true”，则训练无需 对原始情感语料进行分词处理，直接从 “corpusFile”指向的文件读出训练样本
superlativeMultiplier	2.0	最高级词的情感倍数，一般大于1.0
strengthenMultiplier	1.5	加强级词的情感倍数，一般大于1.0
comparativeMultiplier	1.2	比较级词的情感倍数，一般大于1.0
epochTimes	500	最大的迭代次数
learningRate	0.001	指定学习步长。该值越大，学习速度越快。 但是取一个较小的值有利于保持网络学 习的稳定性
errorLimit	0.01	指定期望的误差水平
margin	1.0	训练时区分情感倾向所要求的最小分差

#### 3.16.2. 文本或句子情感分析

运行“cn.edu.fudan.sentiment”包下“SentimentAnalyserStart.java”的例子，需要使用在第一步中训练完成的“sentimentAnalysisNetworkSettingFile”所指向的参数文件。输入文章或句子，输出 SentimentResult 的实例，通过该实例可以得到情感标签以及相应的概率。

### 3.16.3. 词汇情感分析

运行“cn.edu.fudan.sentiment”包下“SentimentEmbeddingAnalyzerStart.java”的例子，需要使用在第一步中训练完成的“sentimentLexiconFile”所指向的参数文件。输入某一词汇，分别输出正面和负面情感倾向的得分。

### 3.17. 文本摘要和关键词抽取

此功能一体完成文本摘要和关键词抽取。

**注意：此功能需要付费开通。**

使用该功能需运行“cn.edu.fudan.Summary”包下“SummaryKeywordJointStart.java”例子，相关参数说明如下：

参数名称	默认值	备注
currentDirectory	config/	当前目录
preprocessFile	conf/Preprocess.properties	预处理配置文件
nerRecognizerFile	conf/NerRecognizer.properties	命名识别配置文件
posTaggerFile	conf/PosTagger.properties	词性标注配置文件
occurrenceLimit	2	篇章中同一实体被确认的最少捕获次数
isFineSegmentation	true	是否进行精细分词。如果为“true”，会将较长的地址、机构或作品名进行细分词
stopwordFile	sentiment/stopword.utf8	停止词表
embeddingFile	embedding/skipgram_cn_d100_w5_n5_i5_m10.model	中文词向量参数文件
summaryMaxLength	100	摘要所含的最长词数

输入文章，会输入该文章的摘要及关键词。如输入以下文章：

“导语：《要点》提出,强化保险公司维护消费者权益的主体责任;督促保险公司切实提高和改进保险服务质量;规范保险消费投诉处理工作;加大对损害保险消费者合法权益行为的检查与曝光力度;加大对损害保险消费者合法权益行为的检查与曝光力度。以下为保监会公告原文:保监消保(2018)64号机关各部门,各保监局,培训中心,中国保险保障基金有限责任公司、中国保险信息技术管理有限责任公司、中国保险报业股份有限公司,中国保险行业协会、中国保险学会:现将《2018年保险消费者权益保护工作要点》印发给你们,请结合本单位实际认真贯彻执行。2018年保险消费者权益保护工作要点 2018年保险消费者权益保护工作要以习近平新时代中国特色社会主义思想为指导,全面贯彻党的十九大、中央经济工作会议和全国金融工作会议精神,按照保监会“1+4”系列文件要求和2018年保险监管工作会议部署,坚持稳中求进工作总基调,坚持“以人民为中心”的发展思想,坚持“保险业姓保、监管姓监”,以强化保险公司维护消费者权益主体责任为工作主线,以督促保险公司切实提高和改进保险服务水平为切入点,以强化投诉处理和矛盾化解、加大查处力度、加强消费者教育和风险提示、突出透明度监管、推进保险业信用体系建设为关键环节,以稳步推进保险消费者权益保护制度机制为保障,打好治理损害保险消费者合法权益行为、防范声誉风险攻坚战,为人民群众的美好生活保驾护航。一、强化保险公司维护消费者权益的主体责任(一)落实保险公司各级机构一把手负责制。督促保险公司各级机构落实以总经理为第一责任人的维护消费者合法权益责任制;督促保险公司将维护消费者合法权益工作纳入各级机构管理层的重要议事日程,定期研究和及时解决问题;督促保险公司做实总经理定期接待日制度。(二)落实保险公司消费者事务委员会职责。发布保险公司消费者事务委员会工作指引,指导保险公司各级机构健全消费者事务委员会组织机构、职责权限和工作规则;督促保险公司消费者事务委员会切实发挥职能,推

动保险公司构建内部“大消保”工作机制。(三)落实保险公司消费者权益保护约束考核要求。督促保险公司将维护消费者合法权益情况、服务质量情况、投诉处理工作情况、涉及消费者权益有关信息的披露情况等纳入各级机构经营考核指标体系,并与各级机构主要负责人、引发问题责任人的薪酬分配、职务晋升挂钩。建立健全监管机构对保险公司消费者权益保护工作的考核指标体系,定期组织考核,督促保险公司切实维护消费者合法权益。二、督促保险公司切实提高和改进保险服务质量(四)完善保险公司服务评价。调整完善服务评价指标体系,规范重要服务创新和重大负面事件评价流程,使评价结果更能反映保险公司服务水平;优化保险公司服务评价系统,改进系统功能,强化信息安全;提高保险公司服务数据质量,完善对保险服务评价指标和数据的核查机制,强化对有关数据和材料的审核;加强服务评价结果运用,加大服务评价对保险公司各级机构经营管理的约束力度,鼓励保险公司积极创新服务方式,推动保险公司增强以服务创造价值的内生动力,进而实现行业高质量发展;探索建立区域性保险服务评价机制。(五)加强保险小额理赔监测。丰富和完善保险小额理赔服务监测指标,提升监测工作的时效性、针对性和约束力,指导保险公司进一步简化理赔索赔流程,减少理赔单证种类,推进单证无纸化,并适时开展数据真实性现场检查。(六)推进保险服务标准化建设。在2017年初步搭建框架的基础上,完成保险服务标准体系整体框架构建,同时逐步推进服务标准化各个子项目落地。(七)监控保险公司销售行为。抓好保险销售行为可回溯制度落实,研究制定互联网保险销售行为可回溯管理细则,督促保险公司记录和保存保险销售过程关键环节,加强对制度落实情况的督导检查,根据实施情况逐步探索扩大可回溯管理暂行办法所涵盖的险种(人群)范围。三、规范保险消费投诉处理工作(八)完善制度流程。总结梳理投诉处理实务中存在的问题,完善保险消费投诉处理制度流程。加大对保险公司督导检查力度,选择若干保险公司及分支机构进行督导检查。推动保险公司各级机构高度重视并积极参与行业调解和诉调对接工作,完善调解程序,强化调解协议约束力,并将保险公司各级机构参与调处工作情况纳入投诉考评、服务评价指标体系。(九)升级管理系统。理顺12378热线管理体制,升级改造保险投诉管理系统,提升投诉管理系统的智能化和精细化水平。对保险消费投诉数据资源深度挖掘、综合分析,为改进保险监管、完善政策提供参考。(十)加强投诉通报。定期发布保险消费投诉通报,督促保险公司各级机构依规处理投诉事项,及时妥善解决消费者诉求,将矛盾化解在公司内部,避免矛盾激化升级。对投诉案件较多、处理不到位的保险公司采取有力监管措施,倒逼保险公司改进保险消费投诉处理。(十一)强化投诉考评。加大对保险公司投诉事项办理回复报告的审查力度,督促保险公司深入分析引发投诉的深层次原因,溯源整改。以投诉考评为抓手,落实保险公司投诉处理主体责任,提高保险公司解决矛盾纠纷的积极性、主动性。(十二)防范群体性事件风险。加强对消费者反映突出问题的分析研判,密切关注舆论动态,抓早抓小,及时排查和化解投诉纠纷重点难点问题,防范群体性事件风险。四、加大对损害保险消费者合法权益行为的检查与曝光力度(十三)加大违法违规查处力度。针对损害保险消费者合法权益的典型问题和突出公司,持续“亮剑”,组织开展“精准打击行动”,揪住一点、查深查透。坚持“严字当头”,从严整治、从快处理、从重问责,发挥警示和震慑作用。组织检查可回溯办法落实情况,重点查处欺骗保险消费者、隐瞒与保险合同相关的重要事项、强制搭售保险等严重损害消费者知情权、公平交易权、选择权的问题。(十四)突出透明度监管。继续督促保险公司加强对涉及保险消费者权益有关信息的披露,及时、真实、准确、全面地向保险消费者披露可能影响决策的信息。定期公布保险公司被投诉情况、损害消费者合法权益的典型案例、保险公司服务评价结果和保险小额理赔服务监测情况。探索开展“以案说险”等教育引导。适时曝光保险公司、保险从业人员失信行为。五、加强保险消费者权益保护基础建设(十五)完善保险消费者权益保护制度机制。出台《保险实名登记管理规定》,完成保险实名查验登记平台建设,认真做好规定发布后的试点实施、健全标准、宣传培训、督导检查等后续工作。研究建立保险广告行为监管规范,建立健全保险广告正、负面清单,推进保险违法违规广告行为治理。修订《保险消费投诉管理办法》,进一步提升保险消费投诉处理机制的针对性和有效性。制定指导性文件规范保险纠纷调解组织建设,推动保险公司积极参与保险纠纷调处及诉调

对接工作,完善保险纠纷调解处理机制。(十六)加强消费者教育和风险提示。多层次、多渠道、多样式、多频次开展保险消费者教育引导活动,强化“保监微课堂”普及保险知识作用。督促行业贯彻落实《中国保监会关于加强保险消费风险提示工作的意见》,加强对行业保险消费者教育及风险提示工作的指导、监督,落实保监局的属地监管责任和保险机构的主体责任。(十七)推进保险业信用体系建设。编制保险业信用信息目录,研究制定保险业信用信息系统管理使用办法,逐步建立健全保险业信用体系建设制度规范。积极参与新制定的守信联合激励和失信联合惩戒备忘录签署并抓好已签署相关备忘录落实。加快推进保险业信用信息系统建设,确保如期完成建设目标。推动将违法违规失信行为作为各级监管部门查处的重要内容,依法严肃追究相关单位和责任人的责任,积极稳妥防范和处置各种信用风险。”

输出的关键词和摘要内容如下:

**关键词:** 金融、导语、合同、动态、要点

**摘要:**《要点》提出,强化保险公司维护消费者权益的主体责任;督促保险公司切实提高和改进保险服务质量;规范保险消费投诉处理工作;加大对损害保险消费者合法权益行为的检查与曝光力度;加大对损害保险消费者合法权益行为的检查与曝光力度。以下为保监会公告:保监消保(2018)64号机关各部门,各保监局,培训中心,中国保险保障基金有限责任公司、中国保险信息技术管理有限责任公司、中国保险报业股份有限公司,中国保险行业协会、中国保险学会:现将《2018年保险消费者权益保护工作要点》印发给你们,请结合本单位实际认真贯彻执行。

### 3.18. 关键短语和关键词抽取

此功能从文本中一体抽取关键短语和关键词。

**注意: 此功能需要付费开通。**

使用该功能需运行“cn.edu.fudan.Summary”包下“KeyphraseExtractionStart.java”例子,相关参数说明如下:

参数名称	默认值	备注
currentDirectory	config/	当前目录
preprocessFile	conf/Preprocess.properties	预处理配置文件
nerRecognizeFile	conf/NerRecognize.properties	命名识别配置文件
posTaggerFile	conf/PosTagger.properties	词性标注配置文件
occurrenceLimit	2	篇章中同一实体被确认的最少捕获次数
stopwordFile	sentiment/stopword.utf8	停止词表
gapLimit	5	相邻关键短语合并时所允许的最长间隔字数
lengthMaxLimit	15	最长的关键短语的字数
lengthMinLimit	5	最短的关键短语的字数
countMinLimit	3	所需抽取的最少关键短语数量
rank	0.05	所有候选关键短语中抽取最重要的关键短语比例

输入文章,会输入该文章的关键短语及关键词。如输入以下文章:

“贾跃亭的妻子甘薇也被列入失信被执行人(即俗称的“老赖”)名单。4月14日,澎湃新闻从中国执行信息公开网上看到,因与中泰创展的纠纷,案号(2017)京03执646号,贾跃亭、甘薇、

乐视控股被列为失信被执行人。根据公告，执行法院为北京市第三中级人民法院，立案时间为2017年8月8日，生效法律文书确定的义务为：一，向浙江中泰创展企业管理有限公司支付人民币14.03亿元；二，向浙江中泰创展企业管理有限公司支付违约金（以人民币14亿元为基数，按照年利率24%计算，自2017年7月11日至实际清偿之日）；三，向浙江中泰创展企业管理有限公司支付迟延履行期间的债务利息。被执行人的履行情况为全部未履行，失信被执行人行为具体情形：违反财产报告制度，发布时间为2018年3月7日。中国执行信息公开网显示，贾跃亭已经7次被列为失信被执行人，乐视控股为6次，甘薇则是首次被列入。此次纠纷的涉及主体为浙江中泰创展企业管理有限公司，中泰创展则与中植系关系紧密。工商资料显示，浙江中泰创展企业管理有限公司成立于2015年，注册资本35亿元，法定代表人为邱晓健。浙江中泰创展企业管理有限公司的股东为北京中泰创展企业管理有限公司，北京中泰创展的法定代表人为周律，而周律也是中泰创展控股有限公司的法定代表人。官网资料显示，中泰创展控股有限公司创建于2008年，是中植企业集团旗下大型新金融控股公司，致力于为优质企业量身打造定制化综合金融解决方案，提供全方位的财务支持和战略资源整合支持。截至2016年末，资产管理规模345亿元，营业收入25亿元，净利润8亿元，资本市场浮盈约15亿元。中植企业集团创建于1995年，总部位于北京，是国内最大的民营资产管理公司之一，目前管理资产超过1万亿元人民币。中植企业集团的实际控制人是资本大佬解直锟。中泰创展和乐视体系间外界熟知的一次交集，发生在2016年11月，牵扯到的核心公司为易到，当时乐视方面持有易到70%的股权，为控股股东。据《证券日报》2017年12月的一则报道，2016年11月份，鉴于易到贷款困难，乐视控股以名下乐视大厦作为抵押物，以乐视汽车生态内的易到为主体取得的一笔14亿元联合贷款，这笔资金，仅有1亿元用于易到，其余13亿元都流入了乐视汽车生态之中。值得一提的是，这14亿元贷款只有1亿元留给易到，还引发了2017年4月，易到创始人周航公开斥责贾跃亭挪用资金的事件，当时乐视方面反击称其为“农夫与蛇”。如今的易到早已被乐视用于断臂求生，新的金主为韬蕴资本和中信银行。报道称，乐视以南京银行为通道将乐视大厦进行了抵押，这笔14亿元贷款资金来自于中泰创展控股有限公司，期限为两年，年利率为8%，总利息为2.24亿元。这14亿元的贷款数额，恰好与上述案件所涉及的金额相符合。另从工商资料看看到，2016年11月9日，乐视控股法定代表人吴孟将持有55万元易到股权（指北京东方车云信息技术有限公司）质押给了浙江中泰创展企业管理有限公司。至于甘薇为何会被卷入其中，目前还不甚清晰。不过，自贾跃亭于2017年7月出走美国后，甘薇更多在国内帮助贾跃亭处理债务问题。2017年12月31日从美国返回北京，甘薇更是正式宣布，与贾跃亭的哥哥贾跃民一同行使上市公司股东权利和履行股东责任，进行资产处置工作。乐视系资金链危机自2016年下半年以来持续发酵。2017年6月，招商银行向法院申请财产保全冻结贾跃亭、甘薇、乐视三家公司12亿资产开始，整个乐视非上市体系陷入溃败停滞，贾跃亭也因多次被债权人申请财产保全，名下资产多处于冻结状态。资产冻结、运营停滞从而债务无法有效偿还，贾跃亭以及乐视多家公司被列入失信被执行人名单。失信被执行人即俗称的老赖。2016年，中共中央办公厅、国务院办公厅印发了《关于加快推进失信被执行人信用监督、警示和惩戒机制建设的意见》，加大了对失信被执行人的联合惩戒力度。《意见》指出，失信被执行人，将遭受从事特定行业或项目限制、政府支持或补贴限制、任职资格限制、准入资格限制、荣誉和授信限制、特殊市场交易限制、限制高消费及有关消费等多项限制。其中，在限制消费方面，失信被执行人及失信被执行人的法定代表人、主要负责人、实际控制人、影响债务履行的直接责任人员将在乘坐火车、飞机、住宿宾馆饭店、高消费旅游、子女就读高收费学校等方面受到限制。有报道据此分析，鉴于甘薇也被列入“老赖”名单，贾氏夫妇下一次团聚将因此多有不便。”

输出的关键短语和关键词如下（中括号数值为重要性得分）：

**关键短语：**

[18.5] 上市公司股东权利和履行股东责任  
[14.8] 债权人申请财产保全  
[14.3] 北京东方车云信息技术有限公司  
[14.2] 中植企业集团  
[14.0] 乐视系资金链危机  
[13.1] 民营资产管理公司  
[12.0] 金融解决方案  
[10.6] 中泰创展控股有限公司  
[10.0] 财务支持和战略资源整合

**关键词：**

[9.0] 大型  
[9.0] 旗下  
[6.0] 金融  
[5.0] 集团  
[5.0] 中植  
[4.3] 控股  
[4.2] 企业

### **3.19. 拼音转换与相似发音分析**

#### **3.19.1. 句子级多音词消歧**

汉语有较多且复杂的一字多音、多字同音现象，此功能将输入句子或文本转换成相应的读音，并且能够根据上下文对多音字进行消歧。

**注意：此功能需要付费开通。**

运行文本或句子转拼音工具（声调用最后一位数字表示，对应一至四声，轻声为 0）。

如输入：“大人，我今天去看大夫。”

输出：“da4 ren2 , wo3 jin1 tian1 qv4 kan4 dai4 fu1 。”

#### **3.19.2. 语句或词汇发音相似性分析**

输入一对短语或词汇，输出其发音相似性分值（分值为 0 至 1 之间）。

如输入：“李清照”和“李青照”

输出：0.83

如输入：“恒生电子”和“恒升电子”

输出：0.875

### **3.20. 三元组抽取**

**注意：此功能需要付费开通。**

### 3.20.1. 样本准备

样本文件格式如：triple/tripleCorpus.utf8 格式所示，三个样本的例子如下：

湖南大宗商品交易中心有限公司目前的董事长兼总经理叫做邓宇，公司历任高管中也没有过一个叫刘俊余的人。

总经理、湖南大宗商品交易中心有限公司、邓宇

董事长、湖南大宗商品交易中心有限公司、邓宇

6月15日上午，兴港投资集团第二届第一次监事会会议在监事会办公室召开，主席汪洋主持会议。

监事会主席、兴港投资集团、汪洋

两年前，当桑贾伊·沙阿成为亚马逊副总裁的时候，谁也没想到，他会以这种方式离开。

RELATION、ARG0、ARG1

其中：每个样本第一行为文本(程序会自动拆解成句子，并且利用整篇文本来识别实体)，之后的几行为需要抽取的三元组，三元组以“关系、实体 0、实体 1”表示。对于整个句子无任何需抽取的三元组，但为比较重要的负样本，则在文本后添加一行“RELATION、ARG0、ARG1”即可。**注意：**文件最后一个样本后需要空两行。

### 3.20.2. 模型训练

运行 cn.edu.fudan.triple 包下的 TripleExtractionModelStart.java，主要参数如下：

```
String currentDirectory = "config/"; // 当前目录
String parameterFile = "triple/parameter_position.utf8"; // 参数文件
String preprocessFile = "conf/Preprocess.properties"; // 预处理配置文件
String nerRecognizerFile = "conf/NerRecognizer.properties"; // 命名识别配置文件
String posTaggerFile = "conf/PosTagger.properties"; // 词性标注配置文件
String tripleCorpusFile = "triple/tripleCorpus.utf8"; // 样本文件
String templateFile = "triple/template_position.utf8"; // 三元组抽取模板定义
String stopwordFile = "conf/stopword.utf8"; // 停止词文件
int occurrenceLimit = 2; // 命名识别最少确认次数
int maxGapSize = 10; // 三元组中两个实体间最大词间隔
int minSentenceLength = 10; // 三元组抽取的最短句子长度
boolean isAddPunctuation = true; // 是否在句末补充标点
int learningTimes = 30; // 学习迭代次数
double errorLimit = 0.00001; // 学习所要求的准确率
double errorDisplayLimit = 0.99; // 错误显示的准确率阈值
double margin = 1.0d; // 模型进行判断所需的最小Margin差异
double delta = 0.2d; // 参数调整的步长
```

比较重要的是配置抽取目标关系的特征模板，比如按 triple/template\_position.utf8 所示：

POSITION:职位

NER:ORGANIZATION NER:LISTEDCOMPANY

NER:PERSON

董事长  
法定代表人  
董事会主席  
董事会提名委员会主席  
副董事长  
执行董事  
董事总经理  
董事、委派董事  
独立董事、独董  
非独立董事  
非执行董事、独立非执行董事  
董事会秘书、董秘  
董事长助理  
监事会主席  
监事会副主席  
监事  
监事代表  
首席执行官、CEO  
总裁  
副总裁、资深副总裁、高级副总裁、商业化副总裁  
总经理、品牌总经理  
副总经理、常务副总经理、生产副总经理  
经理  
财务负责人  
财务总监  
党委书记  
党组副书记  
研究官  
行长

M%w[1]  
M%w[2]  
M%w[3]  
M%w[1]/M%w[2]  
M%w[1]/M%w[2]/M%w[3]  
M%w[-1]  
M%w[-2]  
M%w[-3]  
M%w[-1]/M%w[-2]  
M%w[-1]/M%w[-2]/M%w[-3]  
L%w[-1]  
L%w[-2]  
L%w[-3]  
L%w[-1]/L%w[-2]

L%w[-1]/L%w[-2]/L%w[-3]  
R%w[1]  
R%w[2]  
R%w[3]  
R%w[4]  
R%w[5]  
R%w[1]/R%w[2]  
R%w[1]/R%w[2]/R%w[3]  
R%w[3]/R%w[4]/R%w[5]

特征模板分为三个部分：元信息、目标关系、特征模板定义。

元信息部分共三行。第一行是关系的英文和中文表示，用“:”分隔。第二行为关系中 AEG0 的识别类型定义。第三行为关系中 AEG1 的识别类型定义。如有多种类型，可以使用空格隔开，比如“NER:ORGANIZATION NER:LISTEDCOMPANY”。每一识别类型定义分两个部分，并用“:”分隔。一是规定是以词性（POS）还是命名识别（NER）为识别对象，二是规定其子类。NER 可以是 PERSON、ORGANIZATION、CELLPHONE 等；POS 可以是 NR、NN、NT 等。

第二部分按行列出所需要抽取的关系，同一关系不同表述在同一行中列出，并且以“、”分隔。默认以第一个列出的表述作为标准表述。

第三部分为特征模板定义，每一特征模板占一行。每个特征模板可以由多个子项组成，子项用“/”分隔。每一子项由三个部分表示。以“M%w[1]”为例，开始的“M”表示特征提取的位置，可以取 M、L、R 三个不同的值。M 表示在关系的两个实体间的内容，L 表示第一个实体（可能是 ARG0，也可以是 ARG1）出现的左侧，R 表示第二个实体（可能是 ARG0，也可以是 ARG1）出现的右侧。之后跟“%”符号分隔。后一字符可以取 w、p、n 三种不同的值。w 表示单词，p 表示词性，n 则表示命名识别类型。最后一项在中括号内的为特征相对于实体的出现位置，正号数字表示在实体出现之后的位置，而负号数字则表示在实体出现之前的位置。如位置设为 L，则一般取负值，如为 R，则一般取正值。

模型训练过程中，会出现如下提示：

The training time: 1 with the accuracy 0.5339805825242718 (220/412).  
The training time: 2 with the accuracy 0.912621359223301 (376/412).  
The training time: 3 with the accuracy 0.9660194174757282 (398/412).

民航资源网2017年12月19日消息：12月18日新浪发布公告表示，任命携程董事长梁建章为公司的独立董事。

L 携程 梁建章 [UNKNOWN/独立董事]

选举华融贵州公司副总经理张旭东及该公司股权经营部负责人朱江伟出任董事会董事，华融贵州分公司股权部副经理余佳出任监事会监事。

L 华融贵州 张旭东 [UNKNOWN/董事]

选举华融贵州公司副总经理张旭东及该公司股权经营部负责人朱江伟出任董事会董事，华融贵州分公司股权部副经理余佳出任监事会监事。

L 华融贵州 朱江伟 [董事/UNKNOWN]

徐鸣辞职后，猎豹移动总裁职责将由公司ceo傅盛和高级副总裁孙明焱共同分担。

L 猎豹移动 盛和 [UNKNOWN/副总裁]

The training time: 4 with the accuracy 0.9902912621359223 (408/412).

The training time: 5 with the accuracy 1.0 (412/412).

The parameters of the model have been written into the disk.

The learning process has been completed.

**注意：**在模型准确率超过一定程度时，会显示当前模型仍然错判的例子，从中可以分析问题并调整特征模板，或者发现样本中可能存在的冲突。当前抽取错误三元组后的中括号里，第一项为标准关系，后一项为当前模型判断的关系。

### 3.20.3. 模型使用

运行 `cn.edu.fudan.triple` 包下的 `TripleExtractorStart.java`。主要参数如下：

```
String currentDirectory = "config/";  
String parameterFile = "triple/parameter_position.utf8"; // 参数文件  
String preprocessFile = "conf/Preprocess.properties"; // 预处理配置文件  
String nerRecognizerFile = "conf/NerRecognizer.properties"; // 命名识别配置文件  
String posTaggerFile = "conf/PosTagger.properties"; // 词性标注配置文件  
String templateFile = "triple/template_position.utf8"; // 三元组抽取模板定义（需与训练时的一致）  
String stopwordFile = "conf/stopword.utf8"; // 停止词文件
```

输入文本（程序会自动拆解成句子，并且利用整篇文本来识别实体）或句子，程序输出抽取到的三元组，如下所示：

入职长城后，文飞将负责品牌战略、市场营销及产品传播工作，并统筹管理销售计划、商务政策以及蓝标哈弗独立经销商网络的能力提升，直接向长城汽车股份公司副总裁、长城汽车销售公司总经理李瑞峰汇报。

副总裁、长城汽车股份公司、李瑞峰

总经理、长城汽车销售公司、李瑞峰

新京报快讯（记者马婧）6月22日晚间，中国联通发布公告称，公司董事李彦宏申请辞去公司董事及董事会发展战略委员会委员职务。同时，增补百度公司副总裁王路为公司董事。

董事、百度、王路

副总裁、百度、王路

王建伟主持召开了2018年股东会首次会议。会上，与会股东代表审议并表决通过了《关于变更水钢公司董事、监事的议案》。选举华融贵州公司副总经理张旭东及该公司股权经营部负责人朱江伟出任董事会董事，华融贵州分公司股权部副经理余佳出任监事会监事。会议宣读了《关于申燕出任职工董事、吴学龙出任职工监事的通报》。

副总经理、华融贵州、张旭东

董事、华融贵州、朱江伟

王路先生，52岁，百度公司副总裁，负责公司级的品牌市场、公关、校园品牌市场、政府关系、行政和工程建设、采购及职业道德建设等业务。王路先生毕业于北京联合大学并在北京大学获得EMBA学位。王先生是中国数字信息领域里的老兵，拥有丰富的传媒管理、商业拓展经验，对用户在互联网及移动世界里的行为深蕴其道。加入百度之前，王先生曾任沃尔玛电商亚洲区总裁兼CEO，成功协助沃尔玛全资收购1号店，并领导1号店在华业务，后又成功协助沃尔玛达成与京东集团的深度全球战略合作。王路先生现任北京市人大代表。王先生目前同时还担任爱奇艺、中信百信银行股份有限公司

公司董事。

副总裁、百度、王路

21世纪经济报道记者从接近苏州银行业人士处获悉，近日，苏州资产管理有限公司党委书记、董事长、总裁赵琨接任苏州银行行长一职。

行长、苏州银行、赵琨

## 4. 注意事项

使用本系统 DeepNLP4.2(原 FudanDNN-NLP)受到多项发明专利权和软件著作权保护，用于开发商业目的应用需要购买许可证。

## 5. 联系方式

工具使用过程中，如有任何问题或建议，请通过邮件 [zhengxq@fudan.edu.cn](mailto:zhengxq@fudan.edu.cn)（郑骁庆）联系我们。

## 6. 参考文献

- [1] Haoyuan Peng, Lu Liu, Yi Zhou, Junying Zhou, **Xiaoqing Zheng\***. Attention-based belief or disbelief feature extraction for dependency parsing. *In Proc. AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.
- [2] Jiangtao Feng, **Xiaoqing Zheng\***. Geometric relationship between word and context representations. *In Proc. AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.
- [3] Yi Zhou, Junying Zhou, Lu Liu, Jiangtao Feng, **Xiaoqing Zheng\***. RNN-based sequence-preserved attention for dependency parsing. *In Proc. AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.
- [4] **Xiaoqing Zheng**. Incremental graph-based neural dependency parsing. *In Proc. Conference on Empirical Methods on Natural Language Processing (EMNLP'17)*, 2017.
- [5] **Xiaoqing Zheng**, Jiangtao Feng, Yi Chen, Haoyuan Peng, Wenqiang Zhang. Learning context-specific word/character embeddings. *In Proc. AAAI Conference on Artificial Intelligence (AAAI'17)*, 2017.
- [6] **Xiaoqing Zheng**, Jiangtao Feng, Mengxiao Lin, Wenqiang Zhang. Context-specific and multi-prototype character representations. *In Proc. The Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*, 2016.
- [7] **Xiaoqing Zheng**, Hanyang Chen, Tianyu Xu. Deep learning for Chinese word segmentation and POS tagging. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2013)*, Seattle, Washington, USA, 18-21 October, 2013, pp. 647–657.
- [8] **Xiaoqing Zheng**, Haoyuan Peng, Yi Chen, Pengjing Zhang, Wenqiang Zhang. Character-based parsing with convolutional neural network. *In Proc. The Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI'15)*, 2015.
- [9] Nal Kalchbrenner, Edward Grefenstette, Phil Blunsom. A convolutional neural network for modelling sentences. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*.
- [10] Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. *In Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*.
- [11] Ronan Collobert, Jason Weston, L  on Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12: 2493–2537.
- [12] John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the International Conference on Machine learning (ICML'01)*.
- [13] Hwee T. Ng and Jin K. Lou. 2004. Chinese part-of- speech tagging: one-at-a time or all-at-once? word-based or character-based? *In Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*.
- [14] Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1): 29–48.
- [15] Yue Zhang and Stephen Clark. 2008. Joint word segmentation and pos tagging using a single perceptron. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*.