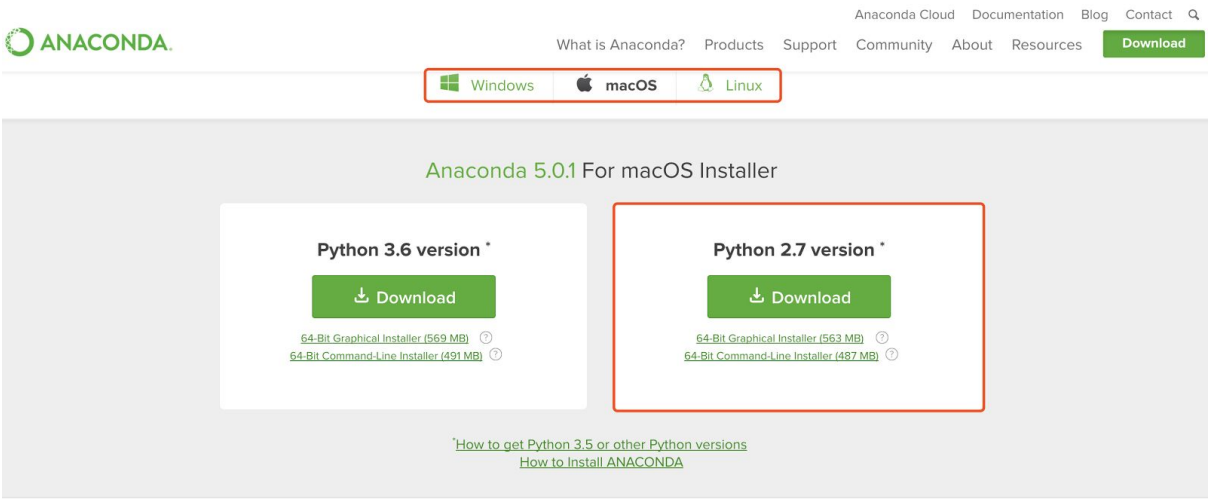
1. 安装

1.1. Python 环境

推荐使用 Python 2.7 的科学计算发行版----Anaconda 2.
请自行下载对应系统的版本, 下载地址为: https://www.anaconda.com/download/



对于不熟悉 Anaconda 的用户, 请允许 Anaconda 为你修改环境变量.

1.2. 配置 Python 第三方库

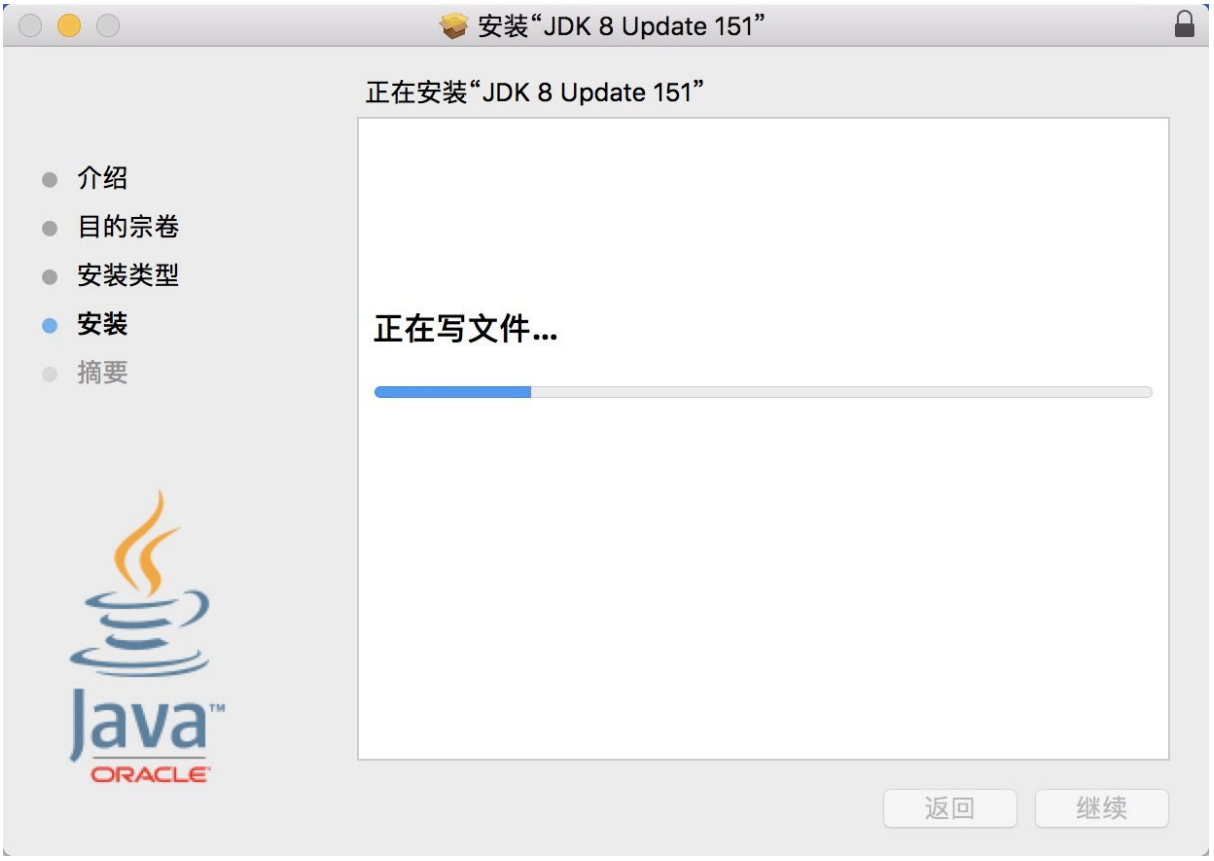
```
# 确认你正在使用正确的 Python 版本
which python
which pip
# 返回的结果中应包含 anaconda 字样

# 开始安装
pip install refo jieba sparqlwrapper
```

1.3. 安装 JDK 8.0

我们的知识库以服务的形式挂载在后台, 接受 HTTP 请求并返回结果. 运行之前需要安装 JDK 8.0 环境. 请自行下载对应系统的版本, 下载地址为:
http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html


JDK 8 Update 151
Java Development Kit
Double-click on icon to install
JDK 8 Update 151.pkg


安装“JDK 8 Update 151”
正在安装“JDK 8 Update 151”
介绍
目的宗卷
安装类型
安装
摘要
正在写文件...
返回
继续

2. 使用

cnquepy 项目的主要文件树如下:

```
.
└── test.py
└── backend
    └── apache-jena-fuseki-3.5.0
    └── DB
    └── KBdemo.xml
└── README.pdf
```

2.1. 文件树简介

./test.py 是运行测试例子的脚本. 内部包含了从自然语言问题到 SPARQL 查询语句的转换逻辑, 也包含了利用 SPARQL 查询知识库的逻辑, 最终得到对应问题的答案, 并输出.

./backend 文件夹下包含三部分:

1. **./backend/apache-jena-fuseki-3.5.0** 是 Jena 的 Fuseki 模块(基于 Java 构建). 运行后在本地监听(http://localhost:3030/), 接收 JSON 格式的 SPARQL 语句, 通过查询后台的 TDB数据库, 返回查询结果..
2. **./backend/DB** 是 Fuseki 的后台数据库文件, 包含结构化的知识库数据文件, 由文件 KBdemo.xml转化得到.
3. **./backend/KBdemo.xml** 是 Fuseki 的后台数据文件. 由原始的三元组数据集转化成 XML 格式的 RDF 文件.

2.2. 代码逻辑说明

1. 预定义 3 类共 5 个示例问题. 包括:
    - "谁是苑茵?",
    - "丁洪奎是谁?",
    - "苏进木来自哪里?",
    - "苑茵哪个族的?",
    - "苑茵是什么民族的人?".

2. 利用结巴分词对中文句子进行分词, 同时进行词性标注. (词性标注中使用的词性兼容了 ICTCLAS 汉语词性标准, 详情可参考: https://gist.github.com/luw2007/6016931
3. 将词的文本和词性打包, 视为"词对象". 对应 **:class:Word(token, pos)**.

4. 利用 REfO 模块对词进行对象级别的 (object-level) 的正则匹配. 判断问题属于 3 种类型中的哪一种, 并产生对应的 SPARQL. 对应 **:class:Rule(condition, action)**.
5. 如果成功匹配并成功产生 SPARQL 查询语句, 立刻请求 Fuseki 服务并返回结果. 打印相关内容.

2.3. 数据集补充说明

1. 采用部分人物数据集作为知识库的内容
2. 原始数据采用三元组形式**<subject attribute object>**, 即**<人物名称, 属性名, 对应属性值>**.
3. 示例如下:

```
丁洪奎 birthDate 1944年11月
丁洪奎 birthPlace 涟水县
丁洪奎 description 现任天明化工厂总工程师、淮阳分厂厂长、盱眙县人大常委会副主任。
苏赫拉布·莫拉迪 award 2012年亚洲举重锦标赛85公斤级挺举金牌
苏赫拉布·莫拉迪 weight 80kg
苏进木 birthDate 1910
苑茵 birthDate 1919年
苑茵 description 苑茵，辽宁本溪人。1942年毕业于复旦大学经济系。历任重庆妇女辅导院人事干事，重庆、天津、沈阳中央信托局业务科主任，天津人民银行科级行员，地质部干部英语班、中国音乐学院英语教师，北京市文史馆馆员。1995年加入中国作家协会。
苑茵 gender 女
苑茵 nationality 中国
姚明 gender 男
姚明 height 226厘米
```

2.4. 运行测试代码

```
# 开启服务(进入此项目的根文件夹 cnquepy)
cd backend/apache-jena-fuseki-3.5.0
nohup ./fuseki-server --loc=../DB /demo > log.txt 2>&1 &

# 运行测试代码, 尝试对比不同句式的结果
cd ../..
python test.py
```

3. 致谢

1. 本文代码中对自然语言问句进行正则匹配的逻辑兼容 REfO. 主要参考代码为: https://github.com/machinalis/refo/blob/master/examples/words.py
2. 本文代码的后续改进可参考: https://github.com/machinalis/quepy
    - 使用邻接链表表示自然语言问句
    - 通过遍历有向图或子图匹配方法构造 SPAPQL 查询语句