基于 StanfordCoreNLP 的生物医学实体间依存分析

2019/3/13 姚昕智

一、实验环境配置（Linux or macOS）

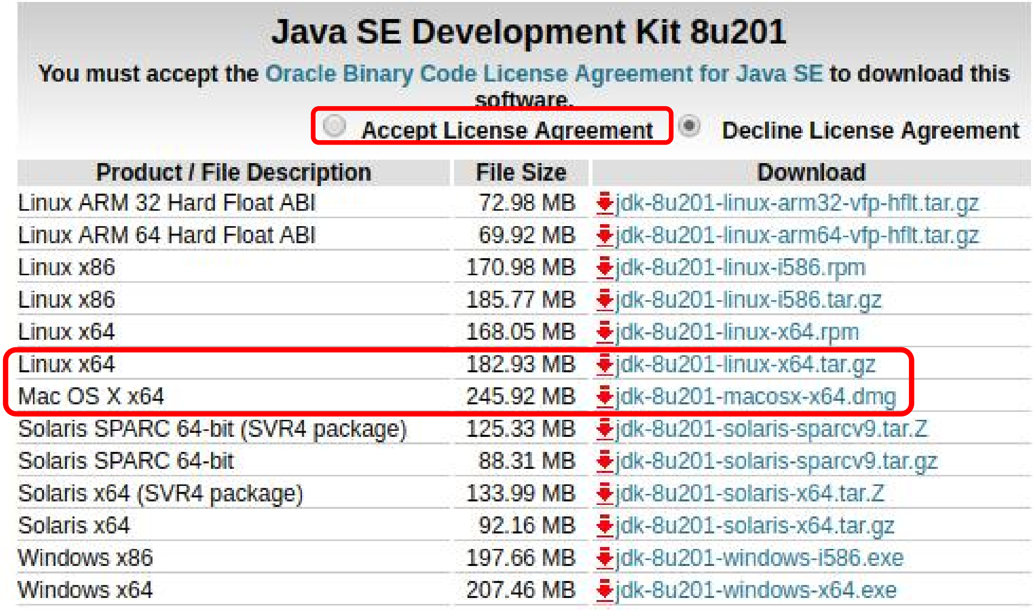
1．配置 JDK 1.8 及以上版本 （建议安装在 home 目录下）

```
java -version # 查看 java 版本， 如果为 1.8.0 及以上版本， 则无需再配置 JDK
```

1. 下载 JDK 安装包

https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

下载系统对应文件：



Java SE Development Kit 8u201

You must accept the Oracle Binary Code License Agreement for Java SE to download this software.

○ Accept License Agreement ● Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.98 MB	jdk-8u201-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.92 MB	jdk-8u201-linux-arm64-vfp-hflt.tar.gz
Linux x86	170.98 MB	jdk-8u201-linux-i586.rpm
Linux x86	185.77 MB	jdk-8u201-linux-i586.tar.gz
Linux x64	168.05 MB	jdk-8u201-linux-x64.rpm
Linux x64	182.93 MB	jdk-8u201-linux-x64.tar.gz
Mac OS X x64	245.92 MB	jdk-8u201-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	125.33 MB	jdk-8u201-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.31 MB	jdk-8u201-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.99 MB	jdk-8u201-solaris-x64.tar.Z
Solaris x64	92.16 MB	jdk-8u201-solaris-x64.tar.gz
Windows x86	197.66 MB	jdk-8u201-windows-i586.exe
Windows x64	207.46 MB	jdk-8u201-windows-x64.exe

2. 安装

(1) Linux 安装

```
tar -zxvf jdk-8u201-linux-x64.tar.gz      # 解压 JDK
cd jdk1.8.0_201
pwd     # 查看当前路径 ‘JDK_path’
```

```
vi ~/.bashrc

# 在 ~/.bashrc 文件最后添加

#####分割线#####

# JDK
JAVA_HOME='JDK_PATH'
export JAVA_HOME
PAHT=$JAVA_HOME/bin:$PATH
export PATH

#####分割线 #####

source ~/.bashrc    # 更新环境变量

java -version    # 检查是否安装成功
```

2．下载 CoreNLP 3.9.2

```
mkdir yourProject   # 创建工作目录

cd yourProject      # 更改工作目录

# 下载 CoreNLP 压缩包

wget http://nlp.stanford.edu/software/stanford-corenlp-full-2018-10-05.zip


unzip stanford-corenlp-full-2018-10-05.zip   # 解压
```

3．使用 pip 安装 stanfordcorenlp（pip 版本与所用 python 版本相同）：

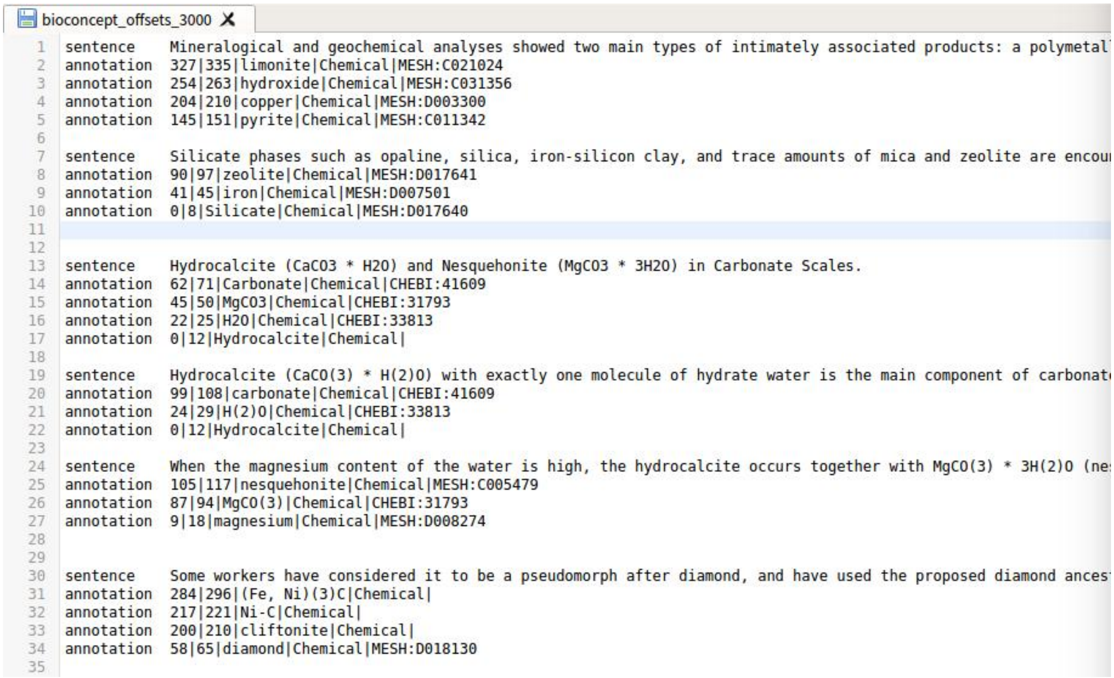
```
pip install stanfordcorenlp   # 安装 python package
```

4. 工作环境配置 （在 /yourProject 目录下）

```
# 下载实验代码
wget https://github.com/YaoXinZhi/CoreNLP_test/archive/master.zip


unzip master.zip     # 解压文件
mv stanford-corenlp-full-2018-10-05 CoreNLP_test-master/CoreNLP_test/

cd CoreNLP_test-master/CoreNLP_test/bin/


# 示例代码 建议大家打开看 StanfordCoreNLP.py 实现
python StanfordCoreNLP.py
python DependencyDistance_StanfordCoreNLP.py -h     # 查看参数


python DependencyDistance_StanfordCoreNLP.py
        -r ../data/bioconcept_offsets_3000 -w ../result     # run
```

二、测试数据

data/bioconcept_offssets_3000



```
sentence    Mineralogical and geochemical analyses showed two main types of intimately associated products: a polymeta
annotation  327|335|limonite|Chemical|MESH:C021024
annotation  254|263|hydroxide|Chemical|MESH:C031356
annotation  204|210|copper|Chemical|MESH:D003300
annotation  145|151|pyrite|Chemical|MESH:C011342

sentence    Silicate phases such as opaline, silica, iron-silicon clay, and trace amounts of mica and zeolite are encou
annotation  90|97|zeolite|Chemical|MESH:D017641
annotation  41|45|iron|Chemical|MESH:D007501
annotation  0|8|Silicate|Chemical|MESH:D017640


sentence    Hydrocalcite (CaCO3 * H2O) and Nesquehonite (MgCO3 * 3H2O) in Carbonate Scales.
annotation  62|71|Carbonate|Chemical|CHEBI:41609
annotation  45|50|MgCO3|Chemical|CHEBI:31793
annotation  22|25|H2O|Chemical|CHEBI:33813
annotation  0|12|Hydrocalcite|Chemical|

sentence    Hydrocalcite (CaCO(3) * H(2)O) with exactly one molecule of hydrate water is the main component of carbonat
annotation  99|108|carbonate|Chemical|CHEBI:41609
annotation  24|29|H(2)O|Chemical|CHEBI:33813
annotation  0|12|Hydrocalcite|Chemical|

sentence    When the magnesium content of the water is high, the hydrocalcite occurs together with MgCO(3) * 3H(2)O (ne
annotation  105|117|nesquehonite|Chemical|MESH:C005479
annotation  87|94|MgCO(3)|Chemical|CHEBI:31793
annotation  9|18|magnesium|Chemical|MESH:D008274


sentence    Some workers have considered it to be a pseudomorph after diamond, and have used the proposed diamond ances
annotation  284|296|(Fe, Ni)(3)C|Chemical|
annotation  217|221|Ni-C|Chemical|
annotation  200|210|cliftonite|Chemical|
annotation  58|65|diamond|Chemical|MESH:D018130

```

注： sentence 行 为提供给 StanfordCoreNLP 依存分析的句子

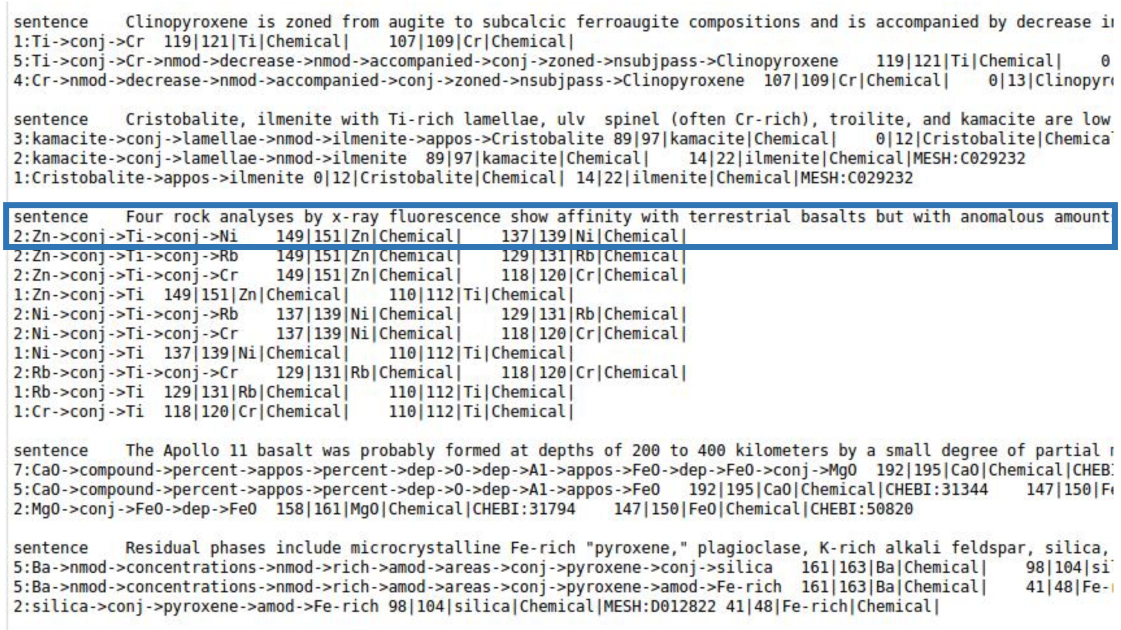
annotation 行为 Pubtator 对句子中出现生物医学实体的注释 其中每一行为一条注释

通过‘|’分隔 分别为 起始位置|终止位置|注释词语|实体类型|对应 ID（如果有）

三、结果解释

1. 通过上面的代码 结果存放在

/result/bioconcept_offsets_3000_DepencyParsing



```
sentence    Clinopyroxene is zoned from augite to subcalcic ferroaugite compositions and is accompanied by decrease i
1:Ti->conj->Cr  119|121|Ti|Chemical|    107|109|Cr|Chemical|
5:Ti->conj->Cr->nmod->decrease->nmod->accompanied->conj->zoned->nsubjpass->Clinopyroxene    119|121|Ti|Chemical|    0
4:Cr->nmod->decrease->nmod->accompanied->conj->zoned->nsubjpass->Clinopyroxene  107|109|Cr|Chemical|    0|13|Clinopyr

sentence    Cristobalite, ilmenite with Ti-rich lamellae, ulv  spinel (often Cr-rich), troilite, and kamacite are low
3:kamacite->conj->lamellae->nmod->ilmenite->appos->Cristobalite 89|97|kamacite|Chemical|    0|12|Cristobalite|Chemica
2:kamacite->conj->lamellae->nmod->ilmenite  89|97|kamacite|Chemical|    14|22|ilmenite|Chemical|MESH:C029232
1:Cristobalite->appos->ilmenite 0|12|Cristobalite|Chemical| 14|22|ilmenite|Chemical|MESH:C029232

sentence    Four rock analyses by x-ray fluorescence show affinity with terrestrial basalts but with anomalous amount
2:Zn->conj->Ti->conj->Ni    149|151|Zn|Chemical|    137|139|Ni|Chemical|
2:Zn->conj->Ti->conj->Rb    149|151|Zn|Chemical|    129|131|Rb|Chemical|
2:Zn->conj->Ti->conj->Cr    149|151|Zn|Chemical|    118|120|Cr|Chemical|
1:Zn->conj->Ti  149|151|Zn|Chemical|    110|112|Ti|Chemical|
2:Ni->conj->Ti->conj->Rb    137|139|Ni|Chemical|    129|131|Rb|Chemical|
2:Ni->conj->Ti->conj->Cr    137|139|Ni|Chemical|    118|120|Cr|Chemical|
1:Ni->conj->Ti  137|139|Ni|Chemical|    110|112|Ti|Chemical|
2:Rb->conj->Ti->conj->Cr    129|131|Rb|Chemical|    118|120|Cr|Chemical|
1:Rb->conj->Ti  129|131|Rb|Chemical|    110|112|Ti|Chemical|
1:Cr->conj->Ti  118|120|Cr|Chemical|    110|112|Ti|Chemical|

sentence    The Apollo 11 basalt was probably formed at depths of 200 to 400 kilometers by a small degree of partial
7:CaO->compound->percent->appos->percent->dep->O->dep->Al->appos->FeO->dep->FeO->conj->MgO  192|195|CaO|Chemical|CHEB
5:CaO->compound->percent->appos->percent->dep->O->dep->Al->appos->FeO   192|195|CaO|Chemical|CHEBI:31344    147|150|F
2:MgO->conj->FeO->dep->FeO  158|161|MgO|Chemical|CHEBI:31794    147|150|FeO|Chemical|CHEBI:50820

sentence    Residual phases include microcrystalline Fe-rich "pyroxene," plagioclase, K-rich alkali feldspar, silica,
5:Ba->nmod->concentrations->nmod->rich->amod->areas->conj->pyroxene->conj->silica   161|163|Ba|Chemical|    98|104|si
5:Ba->nmod->concentrations->nmod->rich->amod->areas->conj->pyroxene->amod->Fe-rich  161|163|Ba|Chemical|    41|48|Fe-
2:silica->conj->pyroxene->amod->Fe-rich 98|104|silica|Chemical|MESH:D012822 41|48|Fe-rich|Chemical|
```

注： sentence 行为解析的句子

下面 行 为 Pubtator 在该句子中注释出的实体两两之间的依存树分析结果

冒号 前的数字为 两实体间最短依存路径距离

冒号 后为两实体间的依存路径

最后为 Pubtator 所注释出的该句子中所含有的两实体自由组合

例如蓝框中 Pubtator 所注释出的实体为 Zn 和 Ni

两实体在句子中的依存距离为 2

依存路径为 Zn -> conj -> Ti -> conj -> Ni

Zn, Ti, Ni 为两个实体最短依存路径上的三个节点

其中 Conj 为 依存类型关系 其余关系可在

Stanford typed dependencies manual (https://link.jianshu.com/?t=http://nlp.stanford.edu/software/dependencies_manual.pdf)中找到