数据科学基础

大作业讲解

2020年12月23日

关于计算社会学

COVID-19背景下的网络社会心态及公众情绪分析

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

获取2019年12月8日至6月中旬新浪、百度、天涯等网站，有关疫情的新闻标题、内容以及重点新闻的评论（重点新闻是指评论量远超（数据分布）其它新闻的新闻）；获取央媒如荔枝网、新华网等网站12月8日-6月中旬有关疫情的新闻标题、内容以及重点新闻的评论；

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

根据标志性事件，将整个疫情发展的时间轴进行阶段性划分，通过分析不同阶段内的互联网数据，掌握不同阶段内社会心态的变化趋势（数据可视化）。

不重视与无奈扩散阶段

2019.12.8-2020.1.22

在这个阶段标志性事件为2019.12.8发现首例肺炎患者和2020.1.22湖北启动公共卫生事件二级响应；

- ✓ 期间央媒的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间较大影响力的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间重点新闻的评论情况（定义重点新闻，获取首发新闻平台的评论内容）

资源缺乏阶段

2020.1.23-2020.2.7

标志性事件为2020.1.23武汉宣布“封城”和2020.2.7“吹哨人”李文亮去世；

- ✓ 期间央媒的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间较大影响力的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间重点新闻的评论情况（定义重点新闻，获取首发新闻平台的评论内容）

严格统一管控和物资配给阶段

2020.2.10-2020.2.13

标志性事件为2020.2.1019省对口支援湖北除武汉外16个市州及县级市和2020.2.13湖北省相关领导的变动；

- ✓ 期间央媒的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间较大影响力的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间重点新闻的评论情况（定义重点新闻，获取首发新闻平台的评论内容）

有序复工阶段

2020.3.10-2020.6

标志性事件为，各省开始有序复工；

- ✓ 期间央媒的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间较大影响力的新闻报道内容（包括新闻标题和新闻内容）；
- ✓ 期间重点新闻的评论情况（定义重点新闻，获取首发新闻平台的评论内容）

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

在此次新型冠状病毒（COVID-19）传播这一重大公共卫生事件情景下，深描中国大众的网络社会**心态**。

心态字典：首先根据分析需求，定义相应的心态词（如冷漠、高兴、怀疑等），再建立一个覆盖较为全面的情绪-心态映射关系，分析文章中的**核心情绪词**（如太棒了->高兴）

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

立足此次新型冠状病毒（COVID-19）传播的这一重大公共卫生事件，基于机器学习进行大众网络社会**情绪预测**。

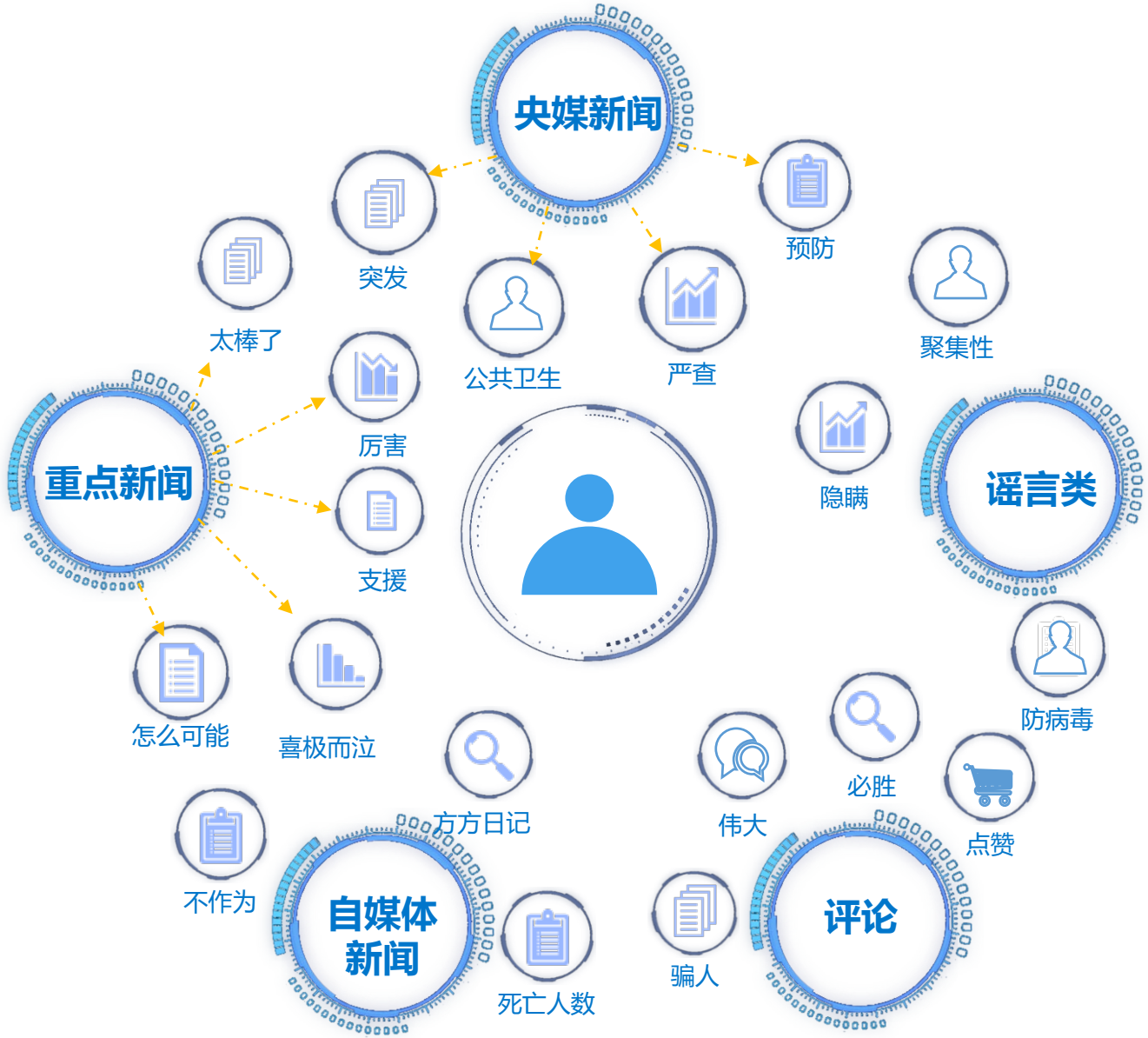
新闻标签：机器学习需要有一定的学习样本，因此在获取到数据后，通过阅读内容根据经验为新闻或评论打上情绪标签，以此进行不同阶段的**情感分析**，可根据所获取数据的总量按一定百分比制定样本；

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

多维情绪解析

Emotion Recognition



如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于心态分析

新闻筛选：Scipy.curve_fit()-最小二乘法拟合高斯分布-大于期望

```
# 关键词提取
document = "水利部水资源司司长陈明忠9月29日在国务院新闻办举行的新闻发布会上透露，" \
           "根据刚刚完成了水资源管理制度的考核，有部分省接近了红线的指标，" \
           "有部分省超过红线的指标。对一些超过红线的地方，陈明忠表示，对一些取用水项目进行区域的限批，" \
           "严格地进行水资源论证和取水许可的批准。"
print(HanLP.extractKeyword(document, 2))
```

提供数据集

关于心态分析

获取2019年12月8日至6月中旬新浪、百度、天涯等网站，有关疫情的新闻标题、内容以及重点新闻的评论（重点新闻是指评论量远超（数据分布）其它新闻的新闻）；获取央媒如荔枝网、新华网等网站12月8日-6月中旬有关疫情的新闻标题、内容以及重点新闻的评论；

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于心态分析

词频（Term Frequency, TF）表示关键词w在文档Di中出现的频率：

$$TF = \frac{\text{关键词的出现次数}}{\text{文档中所有词的数量}}$$

逆文档频率（Inverse Document Frequency, IDF）反映关键词的普遍程度——当一个词越普遍（即有大量文档包含这个词）时，其IDF值越低；反之，则IDF值越高。IDF定义如下：

$$IDF_{\mathrm{w}} = \log \frac{N}{\sum_{\mathrm{i}=1}^{N} I(w, D_i)}$$

其中，N为所有的文档总数，I(w,Di)表示文档Di是否包含关键词，若包含则为1，若不包含则为0。

关于心态分析

根据标志性事件，将整个疫情发展的时间轴进行阶段性划分，通过分析不同阶段内的互联网数据，掌握不同阶段内社会心态的变化趋势（数据可视化）。

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于情绪分析

新闻筛选：Scipy.curve_fit()-最小二乘法拟合高斯分布-大于期望

```
public static void main(String[] args) throws IOException
{
    IClassifier classifier = new NaiveBayesClassifier(); // 创建分类器，更高级的功能请参考IClassifier的接口定义
    classifier.train(CORPUS_FOLDER);                     // 训练后的模型支持持久化，下次就不必训练了
    predict(classifier, "前台客房服务态度非常好！早餐很丰富，房价很干净。再接再厉！");
    predict(classifier, "结果大失所望，灯光昏暗，空间极其狭小，床垫质量恶劣，房间还伴着一股霉味。");
    predict(classifier, "可利用文本分类实现情感分析，效果还行");
}

private static void predict(IClassifier classifier, String text)
{
    System.out.printf("《%s》 情感极性是 【%s】\n", text, classifier.classify(text));
}
```

提供数据集

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

https://github.com/hankcs/HanLP/wiki/%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB%E4%B8%8E%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90#%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90

关于情绪分析

获取2019年12月8日至6月中旬新浪、百度、天涯等网站，有关疫情的新闻标题、内容以及重点新闻的评论（重点新闻是指评论量远超（数据分布）其它新闻的新闻）；获取央媒如荔枝网、新华网等网站12月8日-6月中旬有关疫情的新闻标题、内容以及重点新闻的评论；

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于情绪分析

```
/**
 * 加载数据集
 *
 * @param folderPath  分类语料的根目录.目录必须满足如下结构:<br>
 *                    根目录<br>
 *                    ├── 分类A<br>
 *                    │    └── 1.txt<br>
 *                    │    └── 2.txt<br>
 *                    │    └── 3.txt<br>
 *                    ├── 分类B<br>
 *                    │    └── 1.txt<br>
 *                    │    └── ...<br>
 *                    └── ...<br>
 *                    文件不一定需要用数字命名,也不需要以txt作为后缀名,但一定需要是文本文件.
 * @param charsetName 文件编码
 * @return
 * @throws IllegalArgumentException
 * @throws IOException
 */
IDataSet load(String folderPath, String charsetName) throws IllegalArgumentException, IOException;
```

关于情绪分析

根据标志性事件，将整个疫情发展的时间轴进行阶段性划分，通过分析不同阶段内的互联网数据，掌握不同阶段内社会心态的变化趋势（数据可视化）。

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

在对文本切词时，是否应该要使用停词表去除那些无意义的词。使用停词表后，文本数据必然大幅度降低，然后对于计算TF-IDF也会产生相应的影响，但是基于TF-IDF的计量假设和计算原理来看，我觉得不能使用停词表（因为这些新闻里不会出现政治敏感词，但会有很多人民，你，我，她这种）又觉得可以使用停词表，因为对于不同心态词之间的比例是不会改变的。所以如果使用了停词表，到时候进行数据对比时，是可以通过以某个词的TF-IDF作为标准，求它们之间的比率。但这个比率其实也就是频数以及逆文本频数积的比率。

关于计算社会学

如果我们通过爬虫爬取了其他地方的评论，比如一些自媒体，那我们也要将自媒体的文章爬下来对应起来，否则这数据似乎就不太就有可信度。当然，保留这些评论的来源（url）或许是一个不错的选择?

大作业PDF给出来的那些新闻API中绝大多数都没有评论，然后大作业要求需要有相关的评论。否则不能作为大众心态的判断，毕竟新闻媒体并不能实实在在的反映大众内心真实的想法。所以关于新闻的评论这个我们应该如何处理呢?

关于计算社会学

我们访问了数据源，但是api中只有新浪新闻可以查询到指定日期的新闻，其他的网站均不可。请问有没有其他的办法可以获取指定日期内的新闻呢？或者只获取新浪新闻的数据也是可以的吗？

另外关于“心态词典”，这个是需要我们自己建立，还是是已经建立了的呢？如果已经建立，又在哪里可以找到呢？

关于计算社会学

发现机器学习是大作业中很重要的技术，但是通过百度，知乎，b站等一些途径好像没有找到和机器学习有关的很好的教程，所以想问下助教有推荐的机器学习的教程吗？或者是相关的资料。 谢谢！

关于计算社会学

1. 心态词典主要基于一些据有明确情感的词语，比如“太好了”对应于高兴，但是在荔枝网/人民日报等诸多网站上都以官方新闻和正式的新闻为主，文字风格中立，很少能找到这样的词语，对于这类新闻应该怎么建立心态字典呢?

2. 机器学习相关的知识有点多，而且学习起来没有头绪，请问助教能否帮助缩小范围，或者提供一些参考资料甚至指导?

3. 心态字典最开始应该怎么建立？采用什么样的技术手段?

4. “新闻”和“评论”的比例难以达成一种平衡，比较便于获得的信息大部分是新闻，主要代表了官方的意志，能拿到的大众的数据其实很少。现阶段我们能够找到的数据来源中，鲜有“评论”这一方面。而微信朋友圈和QQ空间的说说几乎获取不到，微博内容和b站弹幕的获取难度也较大。怎么保证数据能够代表“大众”？或者说如何定义重点新闻?

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于计算社会学

5. 如果方便，能否提供些许关于TF-IDF技术的作用和技术指导?

6. 请问“大数据”的数据量是否有具体的量（或者数量级），不知道大概要获取到多少数据才能有效支持我们的分析结果。

7. 课上听陈老师说要给我们一个有关神经网络相关的api，届时是否能有相关使用示范之类?

8. 想了解关于具体如何运用课内知识，对于课内知识点的运用量如果不是很大，是否会对最终结果造成一定影响？（如课本中的：参数估计、假设检验等感觉套不进去，请问能否有个大体方向指导？）

关于计算社会学

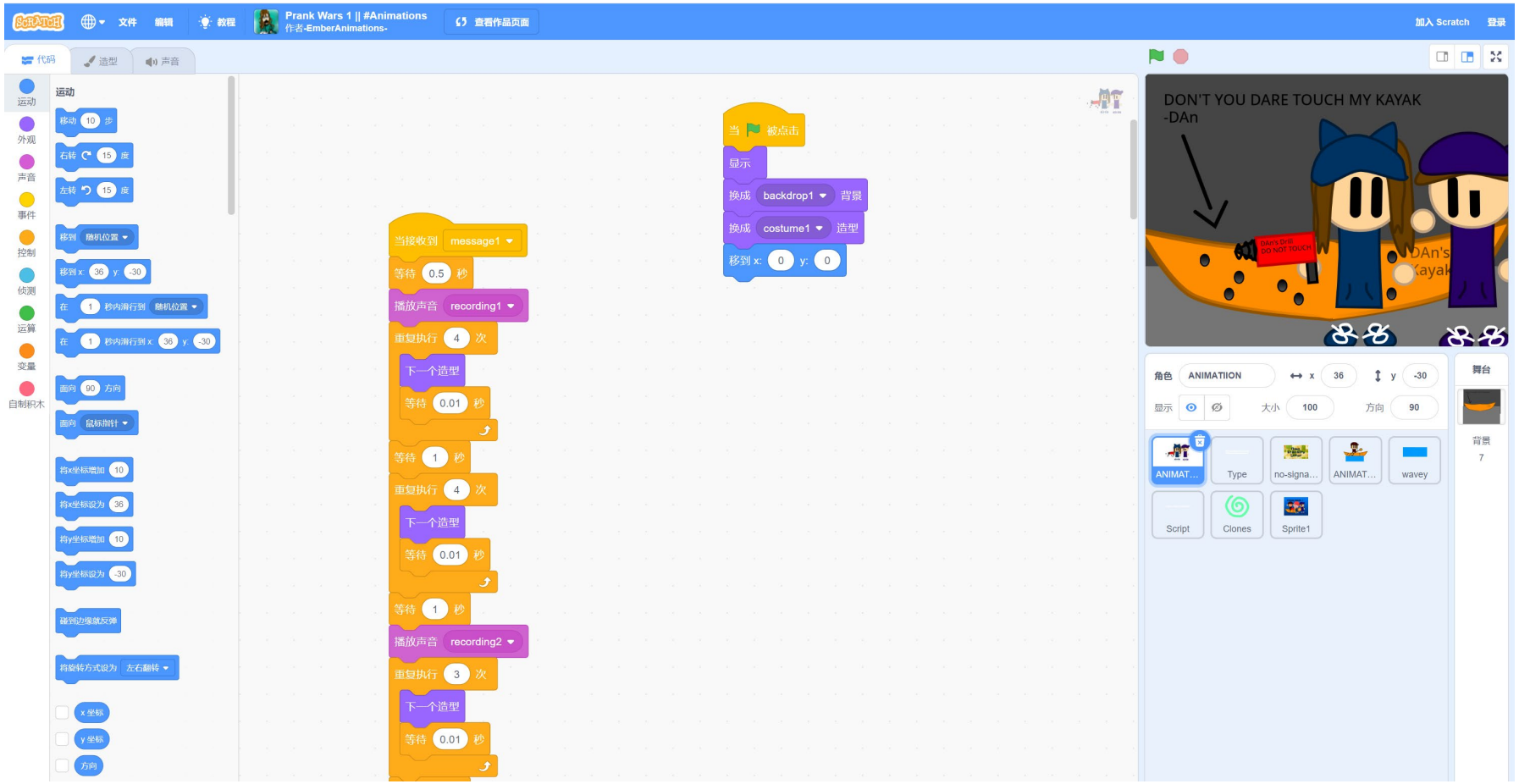
1、百度新闻的爬取，如果按照关键字来搜索的话，“疫”字的搜索结果只能到2020年7月份左右，而与我们想分析的时间段相差太远。

2、关于微博的爬虫，我在github上发现了一些项目。我们是可以直接拿来用吗。还是说比自己写的分要少一点儿。

3、我们对于“预测”这个词理解不是很深入，我们是要根据已有数据预测出下一次重大事件大众会有什么心态吗？预计一些事件比如确诊人数公布对大众心态的影响？我问过的一些同学他们的回答更像是在说我们拿到一篇文章，能推测出这篇文章的情感态度。所以关于预测，我希望能得到更清楚的解释。

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于算法合成



如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

关于算法合成

Scratch	算法合成
指定语言	多语言
分类明确	无明确分类
粒度明确	粒度不定
功能明确	功能不定
手动拖拽	自动处理

关于算法合成

多语言：C/C++, Java, Python……

分类：可视化相关操作、安全相关操作、通信相关操作……

粒度：行、方法、代码块、类……

功能：加密、排序、分类、通信….

形式：自动提示、手动触发

其他：性能、易用性……

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

特征与分类

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

特征与分类

如何判断编程语言？

如何进行功能分类？

特征与分类

通过对源代码文件的后缀进行语言判断

通过代码的相似度判定进行功能判断

特征与分类

基于文本的检测方法是最早的检测代码相似性的技术。首先,预处理代码段，如除去空格、注释等；接着，将代码段转换成字符，如果两个代码段的字符相同，则两段代码相同。

基于词法的检测方法也称为基于Token 的检测方法。首先,将代码段解析成一个字符串序列(Token序列)；接着，检测不同代码段中的Token序列，如果存在相同的Token子序列，说明存在代码克隆现象。

基于度量值的检测方法只针对对代码进行固定粒度检测的情况。首先，将代码切分成固定粒度的比较代码单元；接着，从比较代码单元中抽取度量值,进而确定是否进行了代码克隆操作。度量值包括代码变量、参数、返回值等。

特征与分类

基于语法的检测方法也称为基于树的检测方法。首先，通过对代码进行词法和语法分析，来构建源程序的一棵抽象语法树；接着，比较相同或相似的子树，进而确定是否进行了代码克隆操作。

基于语义的检测方法,也称为基于图的检测方法。首先，通过对代码的语法结构、上下文环境等进行分析，来构建源程序的程序依赖图；接着，通过匹配算法和程序切片得到相同或相似的子图同构的PDG，进而确定是否进行了克隆操作。

特征与分类

检测分类	中间表现形式	匹配算法	优点	缺点
基于文本	字符	字符匹配	算法实现简单，几乎可以检测所有编程语言的源代码	不能识别程序的语法、语义等信息，检测准确率较低
基于词法	Token序列[1,10-11]	LCS、后缀树[24]、语义索引、Karp-Rabin指纹算法	使用轻量级工具，可扩展到对多种编程语言的代码和纯文本的检测，同时相对于复杂算法具有更低的时空复杂度	不能识别程序的语法、语义等逻辑信息，检测准确率较低
基于语法	抽象语法树[12-14]	子树匹配[12]	可识别程序的语法信息，检测准确率较高	构造AST的代价较大，子树匹配算法的复杂度较高
基于语义	程序依赖图[15-20]	子图匹配[22]、程序切片[17,29]	可识别程序的语义逻辑信息，检测准确率较高	构造PDG和子图同构的PDG的代价较大；随着程序规模的扩大，时间复杂度和空间复杂度也提高
基于度量值	程序属性[30-33]	直接比较[30]、欧氏距离[31]、度量值比较[30,32]	检测准确率高，便于代码重构	局限于固定粒度的检测，如果粒度太大，则漏检率会很高

http://www.jsjkx.com/CN/article/openArticlePDF.jsp?id=18909

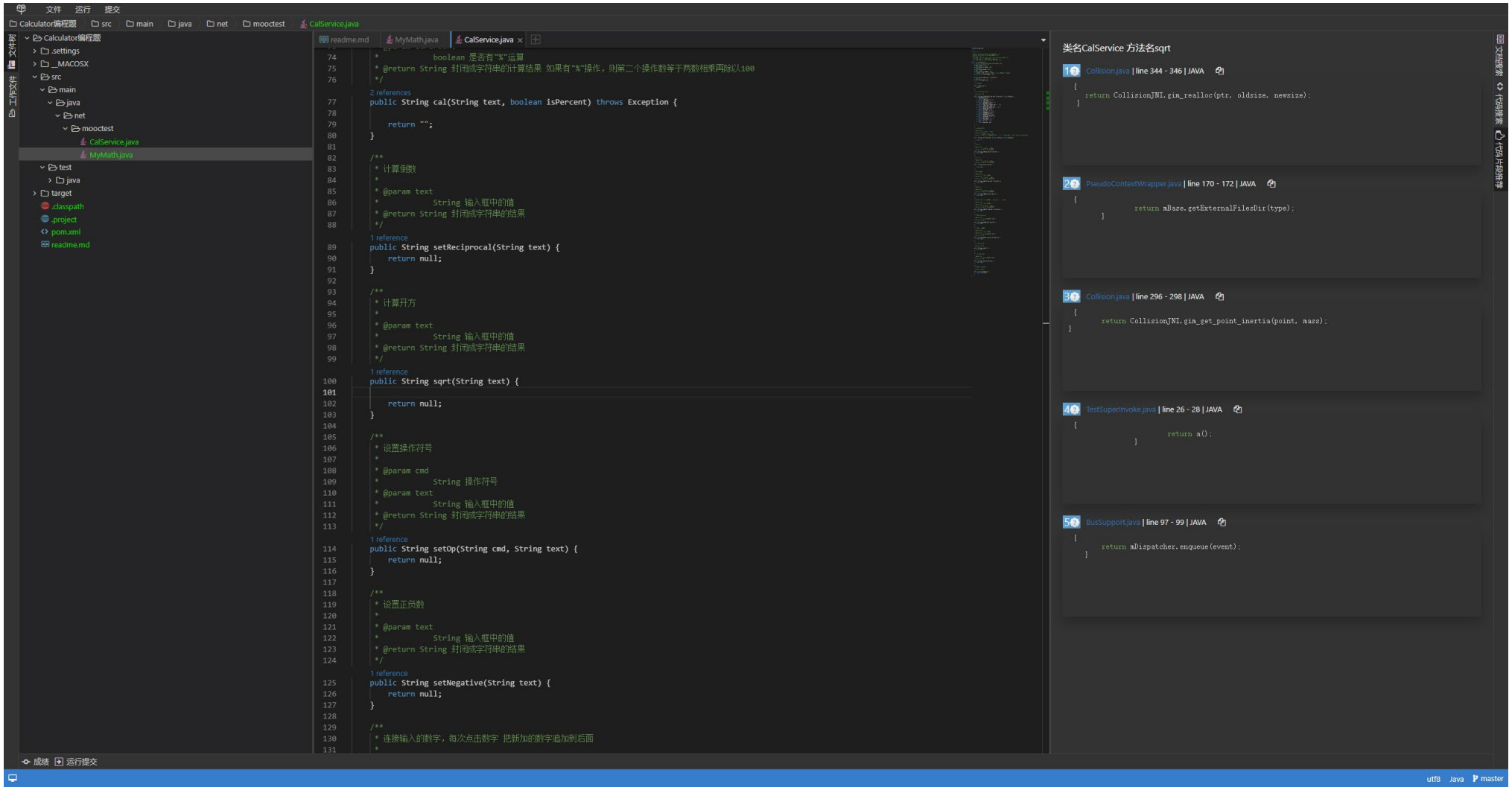
如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

算法合成

如何触发算法合成？

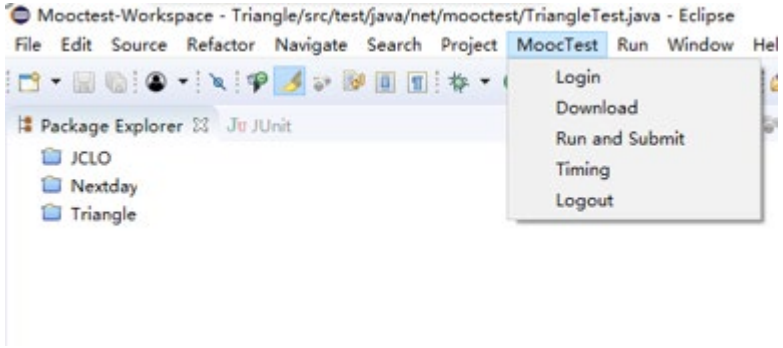
主动触发？定时轮询？

算法合成



如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

算法合成



条件和循环 - 成绩等级判断

Python3 运行(Ctrl+Enter) 提 交 分数： 0 分

题目描述

编写程序读取多个成绩，然后按照如下方案对成绩分级： 若成绩>=90, 则级别为A； 若成绩>=80, 则级别为B； 若成绩>=70, 则级别为C； 若成绩>=60, 则级别为D； 否则成绩为F。

输入描述

输入是一个成绩列表，每一个成绩都大于0且小于等于100。

输出描述

输出为一个成绩级别列表。

测试样例

样例1: 输入-输出

```
[90, 85, 20]
```

```
['A', 'B', 'F']
```

算法合成

如何评估生成代码的质量？

算法合成

可用性	能否成功运行
可读性	代码规范评分
效率	代码的复杂度

需要完成的工作

- 提取代码的语言属性和功能属性
- 对代码进行分类（至少包括语言和功能）
- 对代码质量进行评估
- 选择合理的可解释的算法合成方法
- 提供多个代码推荐，并排序展示
- 合成一个完整的可运行代码

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

评分标准

基本分：完成算法的语言和功能分类；触发算法合成后，能够推荐多份代码，合成后可运行

加分项：给定代码，能够自动进行语言和功能分类；（上限10分）

采用综合方法进行分类；（上限15分）

采用多维指标对代码推荐进行排序；（上限15分）

自己实现基于语义或者语法的相似度对比；（上限25分）

使用插件方式或者在线方式进行交互；（上限25分）

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问

如有疑问，可发邮件至zhaoyuan@smail.nju.edu.cn提问