2025 自然语言处理
课程设计 1

人工智能学院 221300079 王俊童

2025.4.1

综述，首先观察代码结构，逻辑如下：

- 命令行参数解析。有 method，是否 analyze，statistical 里面方法的选取。
- 加载数据和数据分析（需要我们实现数据分析）
- 三个方法的训练：
  - rule: 基于一些规则得到的一个实现。train 有四种纠错规则：
    - $\_extract\_confusion\_pairs$：字符混淆对提取。
    - $\_extract\_punctuation\_rules$：标点符号规则提取
    - $\_extract\_grammar\_rules$：语法规则提取
    - $\_extract\_word\_confusion$：词汇混淆对提取

    然后以上四种错误的纠错发生在 correct 里面。
  - statistical：基于统计学习方法的纠错。这个里面又分为两个模型:
    - ngram 模型：初始化了一堆数据结构，1-4 的 gram 方法，字符混淆矩阵和错误率等
    - ml 模型：用机器学习方法去做。
  - 集成学习方法，在框架代码的 ensemble 部分有留给我们实现。
- 三个方法对应的纠错和评估。跟上面一样了，可以实现很多的 correct 方法，都有对应接口。

可以看出整个代码框架都还是比较整齐的，我们需要完成的 TODO 任务如下:

- 数据的 analyze 分析部分和画图。
- rule：完成规则方法的实现。完成对应规则方法的纠错改正。
- statistical：完成 ngram 和 ml 方法的对应修正和改正。
- main：完成集成学习方法。
- 其余可以加一些深度学习之类的方法实现。

1 实现方法及其简单描述