

软件-软件工程综合实践 2019级

题目1: 英语词汇量估算工具

题目2: 汉语认字量估算工具

张少宏

题目1：英语词汇量估算工具

- 要求：
 - 分组设计，按照软件工程思路设计 英语词汇量估算工具；
 - 提交内容：
 - 数据（如词汇表等）、算法思路、具体设计文档（报告）、代码等；
 - 验证思路；
 - 有实际创新加分
 - 有扩展功能加分
 - 分组责任参考（1-8人）：
 - 总体设计；
 - 算法设计（主要是词汇量测试算法和验证方法）；
 - 前端选择和UI设计（web、桌面程序、app、小程序等都可以）。
 - 简单数据库选择和设计；（不限定数据库）
 - 演示测试：两种测试，一个GUI演示测试，一个是后台批处理测试。
 - 报告撰写。 报告必须注明成员工作量分配（总和100%，如果5个人工作量一样的话就都是20%）

主要功能：（单机模式或者客户端模式）

- 收集词汇列表等不同辅助数据，设计一至多种用户词汇量估算算法；
- 设计验证方法：即 衡量你的算法，估算出来的词汇量到底有多准确？
 - 可与业内产品做比较比如：
 - <http://testyourvocab.com/> （首选），比较可行的方法 浏览器模仿自动测试，抓取结果。
 - 百词斩词汇测试
 - 扇贝单次词汇测试
- 界面设计：可用web页面、桌面程序、app、小程序等。
- 后台批处理测试结果举例：
 - 可考虑 输入一个单词列表，直接算法后台计算结果
 - 输入列表格式：词A， 认识； 词B， 认识； 词C， 不认识； 词D， 不认识； 。
 - 输出结果：估算词汇量
- 界面实例测试结果
 - 找不同学生，每个人测试3-5次，汇报测试结果（分组有多人必须包括这个功能）
 - 主要数据：姓名、四级六级成绩、多次测试结果、 四六级和测试词汇量的相关性。

后台批处理验证方法

- 词汇量估计的 验证方法 是一个 开放问题，没有准确答案。可以上网参考别人的思路，但是不能直接调包。
- 举例：
 - 找一个 设计中 词汇量估计方法没用到的公认词汇列表A 对列表A 多次随机采样 生成 不同长度 不同比例的 词汇测试列表（按比例，比如90%词汇认识，10%词汇不认识）然后多次平均 看看词汇量估计结果的合理性（不同比例，不同长度，多次随机抽样 统计估计结果的平均值、方差值）
 - 不同比例，不同长度 举例：比如 比例 [10% 20% 30%] 长度 [200 300 400]，这样就有9种组合，每个组合测100次， $9 \times 100 = 900$ 次了所以一定要有后台批处理测试方法。

扩展功能：（C/S模式 或者微服务模式）

- 辅助数据和估算算法程序可以根据不同考虑放在服务器端或者客户端；
- 可用服务器端的数据、算法更新客户端相应的数据、算法；
- 发送学生测试实例结果到服务器端数据库
 - 主要数据：学号姓名（如有隐私考虑代号）、四级成绩、六级成绩、测试时间、测试结果。
- 其他更复杂的改进，例如
 - 同一服务不同客户端的交互（比如web和app端）
 - 不同用户的交互（例如：单词对战）

题目2：汉语认字量估算工具

- 要求：
 - 分组设计，按照软件工程思路设计，类似前面的英语词汇量估算工具；
 - 提交内容：
 - 数据（如《现代汉语常用字表》等分级汉字字表等）、算法思路、具体设计文档（报告）、代码等；
 - 有实际创新加分
 - 有扩展功能加分
 - 分组责任参考（1-8人）：
 - 总体设计；
 - 算法设计（主要是词汇量测试算法和验证方法）；
 - 前端选择和UI设计（web、桌面程序、app、小程序等都可以）。
 - 简单数据库选择和设计；（不限定数据库）
 - 演示测试：两种测试，一个GUI演示测试，一个是后台批处理测试。
 - 报告撰写。报告必须注明成员工作量分配（总和100%，如果5个人工作量一样的话就都是20%）