计算机科学
与技术系

实验一　中文分词

老师：戴新宇
助教：欧阳亚文，龙思宇
{ouyangyw, longsy}@smail.nju.edu.cn
或者在课程 QQ 群中联系

概要

- 实验任务
- 实验评分
- 实验提交
- 注意事项

实验任务

- 任务：中文分词
- 要求：只能使用规则分词算法
- 评测：f1
  - 举例来说，如果标签是[我，爱，中国]，预测是[我爱，中国]，此时只预测对了”中国”，所以精度是0.5，召回是0.33，f1是0.4。
  - 注意，计算f1时同时会考虑顺序信息，比如标签是[我我，我]，预测是[我，我我]，此时f1则为0。
- 数据：通过QQ群发送给大家，包括train, dev, test,大家需要提交test集分词结果。

实验评分

- 实验总分为100分
- **分词性能（70%）**
  - 采用打榜的方式。
  - 前十名满分（准备presentation）。
  - 后面的会根据排名和不同排名之间分数的差距递减。
- **实验报告及代码风格、实现方式等（30%）**
- 选做部分（bonus 30%）
  - 超过预先设定的SOTA结果（30%）

实验提交

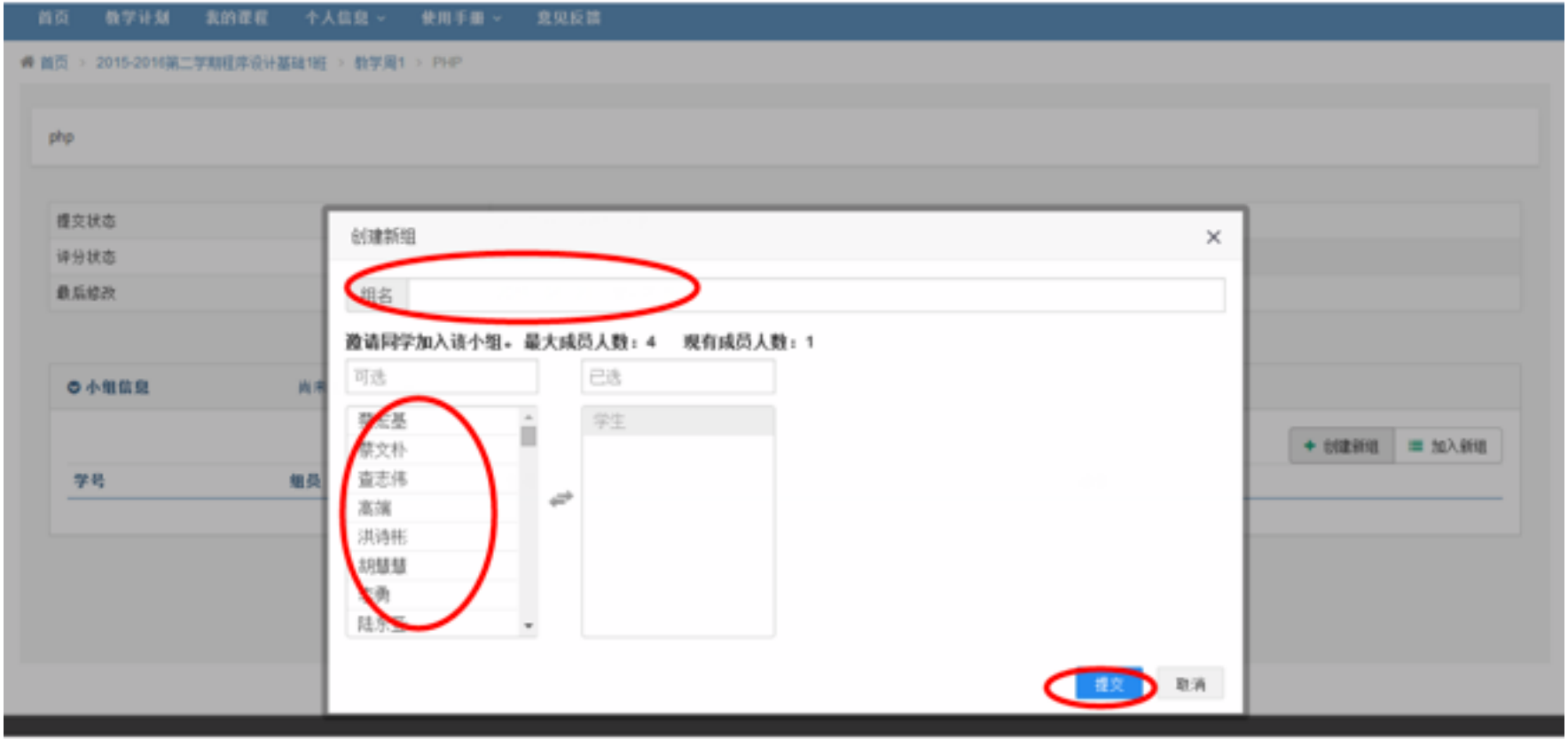
- 排行榜提交： http://114.212.189.62:12345/
- 代码和报告提交：http://cslabcms.nju.edu.cn
  - 使用个人账号登录，独立完成
  - 提交第二周的 Project 1 中文分词项目
  - 所有内容打包并压缩，命名为学号.zip/rar/tar.gz，如181220001.zip，重复提交需要先删除旧版本
  - 提交后请确认作业状态为“已经提交”、“已提交等待评分”，而不是“草稿(未提交)”
- 上述过程中的任何问题请联系助教或老师，无特殊情况不接受其他提交方式，截止日期：**10 月 1 日，23:59:59**，请尽量不要在此之前的几分钟提交，网络有风险

代码和报告提交

- 点击小组信息，加入一个小组或创建一个小组

- 实验设置为该课程系统中的项目作业，但要求独立完成，所以请创建一个只有自己一个成员的小组后提交

代码和报告提交



- 注意：只有在加入某个组之后才能提交作业！

代码和报告提交

- 压缩包内容至少包含
  - 源代码（推荐使用Python）
    - 尽量遵循优秀代码规范，有必要注释
  - 实验报告
    - 推荐 pdf， docx 也可以
    - 报告内容包括个人信息，实现的方法，运行方式，实验总结等
    - **不要贴大段代码，篇幅不超过 4 页**
  - 使用的额外数据，无需提交训练测试数据集
  - README 可选

注意事项

- 若无法复现榜单结果的将取消成绩，所以如果你使用了随机算法，请固定随机种子
- 请不要将任何开源分词器的结果作为提交文件，一旦发现将取消成绩，但允许使用它们的结果作为思考加入规则的依据
- 允许使用额外数据，但需要将数据获取的方式及时在QQ群内分享
- 参考网上的任何代码请指明出处！区别参考与抄袭，任何形式的代码抄袭都是不允许的！被确认的抄袭者与被抄袭者本次实验都记为0分！

Thank you~
Q & A