作业说明

1 任务说明

复述检测是指判别一对比较的文本是否是对同一事件或事实的不同描述。

例如，以下两个句子就是一对具有复述关系的句子。

Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.

Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.

复述涉及语义近似度的计算，是一项有挑战性的工作，广泛应用在 NLP 很多高层任务中，如文本摘要、信息检索、信息抽取、问答系统、人机交互等。

本次试验任务是句对的复述检测，训练语料和测试语料都来自于新闻题材。这些训练语料及测试的 Gold standard 都由人工标注而来。

语料规模为：

训练集： 3876

测试集： 1725

开发集： 200

评价指标： 准确率 P = (correct anwers)/(testcase count)

2 数据格式说明

在作业的文件夹中，已给出训练集 train_data.txt、开发集 dev_data.txt、开发集答案 dev_gold.txt 和测试集 test_data.txt，文件编码格式为 utf8。

训练数据格式为：

Answer \t Sen1_ID \t Sen2_ID \t Sen1 \t Sen2

其中 Answer 为答案，1 代表两个句子具有复述关系，0 代表没有复述关系。Sen1_ID 和 Sen2_ID 为两个句子的 ID。Sen1 和 Sen2 是两个句子的文本。所有字段以\t 分割。

开发集和测试集的格式为:

Sen1_ID \t Sen2_ID \t Sen1 \t Sen2

字段含义与训练集相同，但是没有给出答案。

开发集答案数据格式为:

Answer \t Sen1_ID \t Sen2_ID

字段含义与训练集相同。

最终作业需要以测试集的数据作为输入，输出测试集的答案。测试集答案格式与开发集答案格式相同。

3 作业要求

两个人组队完成，也可一个人完成，自愿组队。最后打包提交以下内容:

- **实验报告** 需要详细说明使用的方法、资源、工具和参考文献，分析实验结果，列出组员分工。
- **程序源代码** 统一放入命名为 src 的文件夹内，要求程序风格良好、注释详细，可运行，并给出程序的运行方式。程序应能在 Linux 环境下运行，运行方式应简单明了(如以 makefile 的形式给出编译，并给出程序运行脚本)
- **测试数据答案** 与 dev_gold.txt 的格式相同，注意以 utf8 格式编码

课堂报告：各组如果对自己组的方法比较满意，想将方法分享给其他同学的，可以自愿选择课堂报告，形式为 oral presentation，各组做好 ppt 派代表进行报告，报告时间为 10 分钟。课堂报告的组会有相应加分。课堂报告将在最后一堂课进行，想要进行课堂报告的同学请提前报名。

最终成绩评定由测试数据准确率、模型方法难度、组员分工、实验报告、代码、课堂报告综合评定。

4 作业提交方式

信科同学提交到: williampei1988@126.com
软微同学提交到: girlhpp@163.com
截止日期：2014.12.29