《互联网数据挖掘》项目作业

中文智能问答系统

任务

- 用户给出事实型问题，系统自动回答
- **封闭测试**：只能从给定的维基百科语料上分析获取答案，不允许利用互联网搜索引擎或问答系统返回的答案提供帮助。
- **开放测试**：可使用互联网任意搜索引擎或问答系统结果。

数据

- 样例数据
  - http://www.icst.pku.edu.cn/lcwm/course/WebDataMining2017/data/wdm_assignment_3.rar（包含wiki数据及样例，优先使用）
  - 中文维基百科所有完整内容文件 (2017-10-20 版本)
    - https://dumps.wikimedia.org/zhwiki/20171020/zhwiki-20171020-pages-articles-multistream.xml.bz2
  - 200 对问题-答案样例
    - http://159.203.142.63/ftp/wdm_assignment_3_samples.txt
- 数千个测试问题（作业截止日期前一周放出）

要求

- 自由分组，建议且最多四人一组
  - 没能成功组队的同学将个人信息发给助教
- 方法要求
  - 不限制开发环境、算法
  - 不得人工修改计算结果
  - 封闭测试只能从给定的维基百科语料上分析获取答案
    - 注意：测试数据中可能存在此版本wiki中不存在答案的问题，对于此类问题可回答null或任意错误的答案；**如在此测试中对此问题做出了正确的回答，将给予作弊嫌疑**，查实后按情节轻重扣分。
  - 开放测试可使用互联网任意搜索引擎或问答系统结果
    - 提示：wiki中不存在答案的问题，在此测试中可能存在答案

要求

- 结果
  - 提供close.txt 和 open.txt文件，分别为封闭测试答案以及开放测试答案
  - 每个文件按问题原顺序每行输出对应答案，答案控制在20个以内汉字及其他字符，过长算作错误回答。
    - 文件格式： UTF-8编码，文本文件（.txt）
    - 每行为对应输入此行的答案，即：答案\n
- 会根据历届学生代码查重，严禁使用以往学生的代码

提交材料

- 问题答案（.txt文件）
- 源代码
- 系统说明文档
  - 作者信息
  - 分工情况
  - 编译/运行环境
  - 系统架构 & 关键技术
  - 使用的方法/资源
    - 给出必要的计算公式
  - 参考文献
  - A4: 5-7页

提交方式

- 所有材料打包发送至邮箱：
  webdatamining2017@163.com
  - 姓名+学号+第三次作业.rar|zip
- 截止时间
  - 2017年12月17日（周日）24:00

QA作业提示

- 参考英文QA系统技术，注意中英文在语言细节上的差别
- 普通架构：问题分析+段落检索+模板匹配
- 高级架构：问题分析+段落检索+候选答案选择+基于机器学习的答案确定
- 可考虑机器学习与深度学习：比如分类与排序学习，对候选答案综合考虑多种特征进行最后选择
- 可利用搜索引擎结果进行模板学习等过程
- 细节决定成败，可结合统计方法与规则