《信息检索与 Web 挖掘》实验要求

一、 实验目的

1．了解海量数据分析的基本原理及方法；
2．掌握基础的自然语言处理技术；
3．掌握实验评估的基本方法。

二、 实验内容

问答网站中需要对相似问题进行合并来优化用户体验，而目前的合并工作都是通过人工识别或者字符完全匹配进行，效率低下。问答网站 Quora（https://www.quora.com/）最近发布了一个公开数据集（https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs），数据集中每一行包括两个部分，根据一定规则抓取的相似问题以及两个问题是否重复的人工标注结果，共有 400000 多行数据。

请下载相关数据，并设计算法实现重复问题的识别，以十折交叉验证下的 F1 值作为算法的评估指标。

三、 合作

3～4 人为一组，鼓励合作。

四、 提交

实验报告应该至少包含如下内容：
1．实验目的、内容与要求；
2．系统设计思路及总体框架；
3．算法描述及具体实现的流程分析；
4．实验结果分析（最好图表说明）：
5．实验总结：系统优缺点、待改进的地方；在实验过程中遇到的问题，实验的心得体会；

五、 评估

从实验报告内容和实验效果两部分进行评估。