

****

信 息 学 院

《社交网络技术与应用》

期末大作业 项目报告

**题 目 医疗知识智能问答系统**

**组 员 22920225705728 栗斌**

**34520192201617 徐荪睿**

**3、项目具体要求**

**1）收集主题社交媒体问答数据**

**可自行选定社交媒体网站，例如豆瓣，微博，知乎等，爬取大作业主题相关的问答**

**数据**

**要求至少爬取 10 万条数据**

**2）数据预处理与存储**

**a) 问答数据处理**

**数据预处理（中文分词、去停用词）与关键词提取**

**数据索引**

**b) 数据存储**

**将处理后的数据存储进数据库设计好的表里**

**3）实现社交媒体跨模态问答系统**

**a) 在网站前端实现聊天窗口**

**在页面上实现一个聊天窗口，允许用户进行问答**

**b) 实现文本问答功能**

**问题理解**

**答案检索或推理**

**基于提供的测试集分析问答算法的优劣**

**c) 实现视觉问答功能**

**视觉问题理解**

**基于图像生成对应描述**

**答案检索或推理**

**基于提供的测试集分析问答算法的优劣**

**4）实现文档处理与分析系统**

**a) 在网站前端聊天窗口实现提示功能**

**实现提示语句，允许用户进行文档处理与分析**

**b) 实现文档处理与分析功能**

**允许上传文档**

**对文档内容进行词性标注并显示部分结果**

**对文档内容进行实体识别并显示部分结果**

**实现文档摘要功能并显示摘要**

**允许用户下载结果文件**

**c) 实现文本挖掘功能**

**允许上传数据集压缩包**

**对数据集进行文本聚类并显示 t-SNE 的结果，分类显示词云图结果**

**允许用户下载结果文件**

**5）实现复杂文档处理系统**

**a) 在网站前端聊天窗口实现提示功能**

**实现提示语句，允许用户进行复杂文档处理**

**b) 实现文档处理与分析功能**

**允许上传复杂类型文档（doc、pdf、png 等）**

**对复杂文档内容进行布局分析（图像、表格、文本、公式的布局）**

**对文档内容进行至少 2 种类型提取并显示部分结果**

**根据布局和提取结果转换成用户想要的文档**