周报-第1周-余连玮

本周完成的事情

1. 阅读论文《网络评论文本的细粒度情感分析研究》并梳理所用技术。
    - 数据处理模块——垃圾评论过滤
        - 构建评论特征
        - 采用朴素贝叶斯分类器进行分类
        - 构建评论可信度Credibility评估函数
    - 数据处理模块——中文分词
        - 采用 NLPIR 分词系统
        - 利用用户词典功能识别新词
    - 情感要素抽取模块——CRFs条件随机场
        - 特征选取
        - 构建标注集
        - 标注意群集
    - 情感要素抽取模块——隐式情感对象识别
    - 细粒度情感分析模块——构建情感歧义搭配词典
        - 挖掘关联规则
        - 利用点互信息挖掘词语间搭配关系强度
        - 构建情感歧义搭配词典
    - 细粒度情感分析模块——对立观点情感强度计算
        - 分析正负向情感强度
        - 生成对立观点情感摘要
2. 阅读论文《食品安全网络公开数据采集技术研究》并梳理所用技术。
    - 数据采集——利用Scrapy框架进行网络爬虫获取数据
    - 文本提取——视频文字提取：利用百度文字识别API识别视频中的字幕
    - 文本提取——音频文字提取：利用百度语音识别API识别音频中的语音信息
    - 图片实体识别：百度图像识别API完成了对图片实体的识别，识别对象主要包含动物和食物
    - 文本处理——利用Jieba分词框架进行中文分词
    - 文本处理——特征选择和提取：利用TextRank算法提取文本中的关键词
    - 文本处理——文本向量化
        - 构建主题词典
        - 再利用权值构建文本向量
    - 文本分类：采用 SVM 算法和贝叶斯算法来进行文本分类，区分文本属于食品安全话题与否。
    - 文本聚类：使用 K-means 方法对向量化后的文本数据进行聚类，得到一段时间内食品安全话题下的热点。

3. 阅读《文本话题识别算法的研究与实现》并梳理所用技术。
    - 文本预处理
        - 利用jieba分词进行中文分词
        - 根据停用词表进行停用词去除
        - 同义词去除：准备完备同义词表，利用大型语料库训练 word2vec 模型用于对多义同义词归并进行确定
        - 关键词抽取：利用 TF-IDF 算法计算每个文档单词的权重 ， 根据TF-IDF 值选取最高的部分单词作为文档的关键词
    - 垃圾信息过滤——利用SVM 进行文本分类
    - 文本建模及相似度度量
        - 主题模型LDA建模
        - VSM建模
        - 相似度计算及加权综合
    - 文本聚类：在相似度的基础上进行层次聚类，返回每一个文档对应的类别数。
4. 运行并理解《食品安全网络公开数据采集技术研究》代码。
    - 启发：增加数据源，采集视频、音频数据以丰富数据集。
    - 发现视频文字提取部分可以换一种方法实现：直接将视频转换为音频在进行语音识别可能效果更好。
5. 运行并理解《食品安全事件检测与分析》情感分析模块的代码。具体代码分析见文档—— `情感分析模块代码分析.md` 。
6. 与王子昂一起梳理所有论文技术，整理整体实训，具体见文档—— `食品安全事件检测与情感分析技术总结.md` 。
7. 修改实训计划，具体见实训计划 `实训计划_王子昂_余连玮.docx` 。

下周的计划

- 寻找新的适合的视频或音频数据源，进行文本提取，形成可调用的api，集成到邓娜代码中。
- 帮助王子昂重构代码。