第2周周报——王子昂

本周完成的工作

梳理代码

- 分析邓娜数据处理（清洗、分词）、话题提取、事件检测部分的代码

数据处理

- 将爬取到的新闻信息存储在csv文件中，包括序号、时间、新闻内容等
- 将每一条新闻的内容抽取为单独的txt文本，方便进行下一步的处理
- 对每条新闻内容，利用停用词表，去除停用词、制表符、html标签，分词，将分词结果存放到语料库中，同时将每篇文章的分词结果保存到对应的txt文件中

话题提取

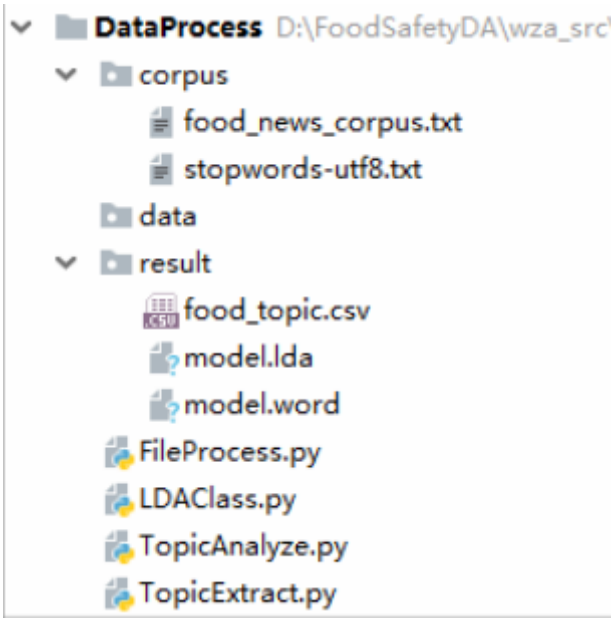
- 主题困惑度分析，进行测试集、训练集划分，构建语料库，进行向量转化，得到对应主题数下的困惑度，得到困惑度最小时的主题数
- 利用Python lda包中的LDA模型输入语料、主题数进行训练，并将模型保存
- 提取每个话题对应的TOP-N的主题词，保存到csv文档中
- 对爬取的新闻信息进行主题的标注
- 根据主题不同，将不同主题的新闻事件进行归类，将相同主题的新闻信息放到一个文件夹下

事件检测

- 从新闻中抽取一部分数据，利用YEDDA工具人工标注新闻中出现的实体，其中食品专有名词标记为FOOD，其余为OTHER
- 将标记好的数据划分为训练集、开发集、测试集，送入命名实体识别模型中进行训练，得到一个能对食品安全专有名词进行标注的模型
- 在运行时发现，一方面由于标记的数据量太小，再加上送入的句子文本过长，导致精度很低，基本不能识别出实体，需要进行改进。
- 利用食品专有名词，以及新闻文本进行相似度计算，得到话题中存在的事件。

代码重构

- 对LDA模型部分的数据处理、模型训练进行重构



- 抽取出LDA类进行模型训练、关键词提取，使结构更加清晰，抽取参数，增加代码的可重用性，同时对构建出的模型进行保存，避免重复训练

```
class LDAClass:
    corpus = []

    '''读取txt文件里面的内容保存为语料'''

    @staticmethod
    def get_corpus(self, filepath):...

    '''打印lda主题词'''

    @staticmethod
    def print_top_words(model, feature_names, n_top_words):...

    '''lda+kmeans'''

    def lda_kmeans(self):...

    ''' 将每篇doc对应的topic存储'''

    @staticmethod
    def save_topic(readfile, writefile):...

    '''将topic存入单独的csv文件'''

    @staticmethod
    def get_topic_word(top_n):...

    """lda train"""

    def train(self):...
```

- 对于数据处理部分，抽取参数，去掉重复、无用代码，并添加必要注释

```python
class FileProcess:
    """
    读取csv保存为单独的txt文件
    将csv中的每一行保存为txt/*/nlp_test_i.txt文件
    """

    @staticmethod
    def csv_to_single_txt(readfile, writefile):...

    # 遍历指定目录，显示目录下的所有文件名
    @staticmethod
    def display_file(filepath):...

    # 读取文件内容并打印
    @staticmethod
    def read_file(filename):...

    '''读取文件目录下所有文件组合成语料并且将所有txt文档集合在一个文档中
    #1.进行分词
    #2.进行停用词的处理
    3.返回语料
    4.并将处理的txt文档集合在一个文档中
    '''

    @staticmethod
    def make_corpus_from_dir(root, write_path):...

    '''添加新列'''

    @staticmethod
    def add_cols(readfile, writefile, content):...
```

- 命名实体识别模型部分，利用标点符号以及文字长度对数据集中的新闻进行切分，重新进行训练，准确率得到了较大的提升，下一步可以考虑对更多的数据进行标注，来提高准确率，同时除了标记食品名词，还可以添加对地点、人物的标注

```
2019-06-25 19:45:50,564 - log\train.log - INFO - iteration:8 step:20/740, NER loss: 0.386200
2019-06-25 19:45:54,071 - log\train.log - INFO - iteration:8 step:120/740, NER loss: 0.310605
2019-06-25 19:45:57,473 - log\train.log - INFO - iteration:8 step:220/740, NER loss: 0.348355
2019-06-25 19:46:01,435 - log\train.log - INFO - iteration:8 step:320/740, NER loss: 0.366574
2019-06-25 19:46:04,995 - log\train.log - INFO - iteration:8 step:420/740, NER loss: 0.356825
2019-06-25 19:46:08,805 - log\train.log - INFO - iteration:8 step:520/740, NER loss: 0.356400
2019-06-25 19:46:12,384 - log\train.log - INFO - iteration:8 step:620/740, NER loss: 0.347280
2019-06-25 19:46:16,165 - log\train.log - INFO - iteration:8 step:720/740, NER loss: 0.418124
2019-06-25 19:46:16,872 - log\train.log - INFO - evaluate:dev
2019-06-25 19:46:26,549 - log\train.log - INFO - processed 257650 tokens with 1934 phrases; found: 4393 phra

2019-06-25 19:46:26,549 - log\train.log - INFO - accuracy:  92.27%; precision:  15.21%; recall:  34.54%; FB1

2019-06-25 19:46:26,549 - log\train.log - INFO -              Food: precision:  15.21%; recall:  34.54%; FB1

2019-06-25 19:46:26,556 - log\train.log - INFO - evaluate:test
2019-06-25 19:46:34,789 - log\train.log - INFO - processed 217787 tokens with 3448 phrases; found: 4131 phra

2019-06-25 19:46:34,789 - log\train.log - INFO - accuracy:  95.08%; precision:  42.65%; recall:  51.10%; FB1

2019-06-25 19:46:34,789 - log\train.log - INFO -              Food: precision:  42.65%; recall:  51.10%; FB1
```

讨论之后开发方向

- 丰富数据源：添加社交媒体（微信公众号、微博）、视频新闻（央视网、优酷）、新闻网站的数据
- 优化命名实体识别模型的构建，对更多的数据进行标注，增加标记的数据类型
- 进行食品安全热点事件的趋势分析

实训计划

内容：食品安全事件检测与舆情分析

1 / 第一周

- 余连玮：理清《网络评论文本的细粒度情感分析研究》和《文本话题识别算法的研究与实现》论文所用到的技术；
- 王子昂：理清《基于网络数据的食品安全事件检测与分析》和《社交网络推文情感分类系统的设计与实现》所用到的技术，将论文中涉及的技术整合到一个框架中，确定之后的研究和开发方向
- 两人共同完成：讨论并将上述论文的技术整合到一个框架，确定后续深入的方向。
- 输出：得到一个整合上述所有技术的大纲。

2 / 第二周

- 王子昂：配置环境，整理命名实体识别代码（标注、识别）、话题提取部分、事件检测部分的代码，理解代码结构和实现方式。基于源代码的基础上进行代码的重构，添加必要注释，修改其中存在的问题，使整体结构清晰，代码

易于重用。结合得到的新的数据源，重新进行数据处理，命名实体识别、构建新的库语料、训练词向量、得到新的 LDA 模型。

- 余连玮：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。
- 输出：得到更加丰富的数据集、并对新的数据集进行数据处理、得到重新训练的可用模型。

3 / 第三周

- 数据源：社交媒体（微博、微信公众号）、新闻网站、视频网站（央视网、优酷）
- 共同完成：爬取数据
- 输出：丰富的数据

4 / 第四周

- 对爬取的数据进行预处理，包括清洗、去停用词、分词，得到语料库，进行话题提取
- 从爬取的新闻中抽取一部分数据利用 YEDDA 进行命名实体的标记（时间、地点、食品专有名词）利用这些数据进行模型训练
- 输出：LDA、Bi-CRF 模型训练结果

5 / 第五、六周

- 共同完成：全部的事件按时间分类，对每一个月的新闻数据进行主题提取、事件检测，选出热点最高的 TOP-N 事件，绘制趋势图（根据事件发生的地点不同，可以针对地点、时间进行趋势分析）
- 输出：得到一个完整的食品安全事件检测与分析系统：采集文本、视频、音频数据，进行事件检测与提取，对不同时期相关的社交网络推文进行情感分析，得到食品安全热点事件的变化趋势。