哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

自然语言处理

实验二：电商评论观点挖掘

School of Computer Science and Technology

Harbin Institute of Technology

1 实验目标

本次实验目的是对命名实体识别以及分类技术有一个全面的了解，本次实验的过程包括属性词以及情感词识别、情感分类、属性分类等环节。本次实验所要用到的知识如下：

- 基本编程能力（文件处理、数据统计等）
- 训练数据和测试数据的使用方法
- 数据预处理、实体识别技术、文本分类技术

2 实验环境

推荐编程语言为：C++/C#/Python/Java

操作系统：Windows/Linux/Mac OS

其他无特殊要求

3 实验内容及要求

实验是在商品评论中抽取商品属性特征和消费者观点，并确认其情感极性和属性种类。对于商品的某一个属性特征，存在着一系列描述它的观点词，它们代表了消费者对该属性特征的观点。每一组{商品属性特征，消费者观点}具有相应的情感极性（负面、中性、正面），代表了消费者对该属性的满意程度。此外，多个属性特征可以归入某一个属性种类，例如外观、盒子等属性特征均可归入包装这个属性种类。实验给定商品评论文件和对应的标注文件。示例格式如下：

```
# Train_reviews:
# 二元组:(评论ID,评论内容)
233,快递也挺快的，包装的也很好，快递员服务也好，就是遮瑕效果不很好

# Train_labels:
# 九元组:(评论ID, Aspect, A_start, A_end, Opinion, O_start, O_end, Category, Polarity)
233,快递,0,2,挺快的,3,6,物流,正面
233,包装,7,9,很好,11,13,包装,正面
233,快递员服务,14,19,好,20,21,物流,正面
233,遮瑕效果,24,28,不很好,28,31,功效,负面
```

3.1 属性词-情感词识别

属性词-情感词(Aspect-Opinion)识别，输入为商品评论，输出为属性词以及对应的情感词，在该实验中对实体的**偏移不做要求**，另外存在只有 Aspect 或者只有 Opinion 的情况。

两个参考思路：

1） 识别句子中所有的属性词和情感词，然后将属性词和情感词一一对应起来。

2） 识别句子中的属性词，然后遍历识别出来的属性词，根据属性词来识别该属性词对应的情感词。（或者先识别情感词，然后识别每一个情感词对应的属性词）

```
# 输入:
# 二元组: (评论ID,评论内容)
2,发货速度和配送服务还是一流的，还不错

# 输出:
# 三元组: (评论ID, Aspect, Opinion)
2,发货速度,一流的
2,配送服务,一流的
2,_,还不错
```

提交内容：1）对测试数据集预测的结果，格式如上图输出所示。

3.2 属性分类

判断 3.1 中识别 Aspect-Opinion 的 Category，该字段的结果属于以下集合

｛ 包装，成分，尺寸，服务，功效，价格，气味，使用体验，物流，新鲜度，真伪，整体，其他 ｝

```
# 输出:
# 四元组: (评论ID, Aspect, Opinion, Category)
2,发货速度,一流的,物流
2,配送服务,一流的,物流
2,_,还不错,整体
```

提交内容：1）在步骤 3.1 的基础上，对测试数据集预测的结果，格式如上图的输出所示。

3.3 观点极性分类

在步骤二的基础上，判断步骤一中识别 Aspect-Opinion 的 Polarity，该字段的结果属于以下集合

｛ 正面、中性、负面 ｝

```
# 输出:
# 四元组: (评论ID, Aspect, Opinion, Category, Polarity)
2,发货速度,一流的,物流,正面
2,配送服务,一流的,物流,正面
2,_,还不错,整体,正面
```

提交内容：1）在步骤 3.1 的基础上，对测试数据集预测的结果，格式如上图的输出所示。

4 实验报告

要求：以小组为单位，字数不少于 4000 字，采用科技论文的组织方式，内容包括作者信息（姓名、学号、email）、中英文摘要、引言、实体识别相关研究工作、自己所采用的方法和在实验给定数据集上的实验结果、实验结果分析、本次实验的心得收获及相关的参考文献。

实验报告要格式规范、逻辑清晰、内容完整。

5 提交方式(暂定)

截止日期：2019 年 12 月 8 日 24 点

提交方式：三个步骤的结果采用打榜的方式，请将结果提交到指定平台，另外将源代码、实验报告以及队伍信息以附件的形式发送到邮箱(banifeng@126.com)。

6 评分标准

允许组成小组完成实验，每个小组不超过 3 个人，根据小组打榜的排名以及得分确定三个步骤得分，小组内部根据贡献多少确定一个排名，排名权重分别为 1、0.95、0.9。

1） 步骤一 6 分，步骤二、三分别为 3 分。

对于每个步骤的评分采用 F1 值评价。每个步骤预测的元组总个数记为P；真实标注的元组总个数记为G；正确的四元组个数记为S。

精确率：

$$\text{Precision} = \text{S/P}$$

召回率：

$$\text{Recall} = \text{S/G}$$

F1 值:

$$\text{F1} = \frac{2 * Precision * Recall}{Precision + Recall}$$

2） 报告 8 分。

以小组为单位，字数不少于 4000 字，采用科技论文的组织方式，内容包括作者信息（姓名、学号、email）、中英文摘要、引言、实体识别相关研究工作（2 分）、小组所采用的方法和在实验给定数据集上的实验结果、实验结果分析（4 分）、本次实验的心得收获（1 分）及相关的参考文献（1 分）。

7 其他要求

1） 实验允许使用预训练好的语言模型，以及参考开源的源代码，需要在报告中说明，但是不建议直接命令行调用第三方的工具。

2） 实验禁止使用其他的标注数据或第三方的标注工具对测试数据进行标注。

3） 禁止人工对测试数据进行标记。

4） 本数据集为化妆品品类的评论数据。为保护品牌隐私，数据已做脱敏，相关品牌名等用**代替。

5） id 字段作为唯一标识对应 Train_reviews.csv 中的评论原文和 Train_labels.csv 中的四元组标签。一条评论可能对应多个四元组标签。

6） Train_labels.csv 中的 A_start 和 A_end 表示 Aspect 在评论原文中的起始位置；O_start 和 O_end 表示 Opinion 在评论原文中的起始位置。若 Aspect 为"_",则 A_start 和 A_end 为空，Opinion 同理；（注：预测结果不需要位置信息，仅考察四元组的预测情况）。

7） Aspect 和 Opinion 字段抽取自评论原文，与原文表述保持一致。若 Aspect 或 Opinion 为空，则用“_”表示。

8） 数据有版权，仅限本次实验使用，请不要外传。

9） 将三个步骤的结果文件分别提交到指定平台上，注意有最大提交次数限制。平台地址为 https://competitions.codalab.org/competitions/21751。