**开课吧人工智能课学院**

**自然语言片与人工智能第5期 项目四报告书**

项目名：新闻人物言论自动提取

队名：冯・诺依曼队

在本项目中，我们实现了一个新闻人物言论自动提取的构建及优化，并以WEB应用的形式编写了可以与用户进行交互的展示程序。

在实现项目目标的过程中，我们小组运用了多种从NLP课程中学习到的方法和技术，并根据本项目自身的特点，灵活地做了一些调整和优化，成功地达到了预期的目标。

但是，由于项目时间和规模有限，以及我们对自然语言处理技术还处在逐步的学习探索过程中，因此，实现的项目成果还有一些有待改进的地方。

以下，我们简单总结了一下我们小组的项目四成果的优点和不足之处。

**【优点】**

**1、在取得言论的算法精度上下功夫进行了优化**

・取本句子 代表说的动词后面的言论。

・取之前句子的言论和取之后句子的言论。

・找到代词指代的人，合并同一个的言论。

・提取出了动词的否定词，以使结果更精确。例如‘不认为’。

・输出结果优化。过滤部分因为句子解析准确度问题导致的错取。

**2、训练块，迭代快，调整快**

对于本项目的任务场景来说，模型训练所需的时间相对较长是一个很大的挑战。对于需要反复多次进行的【分析现象 -> 提出假设 -> 改变模型做实验 -> 观察是否符合期望】这样的调参过程来说，长时间的模型训练使得调参的结果很难在短时间内得到反馈，从而使得整个模型的迭代效率大大降低。

为了从根本上提高模型的优化迭代速度，我们从关键性的“模型训练时间”着手，采取了多种方法来提高模型训练的效率，有效果地减少了在此环节上的时间浪费，并提高了训练结果的精确度。经过以上努力，我们将训练的时间控制在了合理的范围内，并且使得训练结果的有效性在很大程度上得到了保证。这就为我们后续反复调整参数，优化模型精确度的工作打下了基础。

**3、尝试增加了情感分析/识别负面言论的功能**

结合之前课程学习内容，增加了情感分析/识别负面言论的模块，丰富了本应用程序的功能。

**4、有效运用框架**

在将模型构建成WEB应用程序的过程中，我们使用了「Django」框架，使得前端和后端能在很短时间内搭建起来。这大大节省了搭建WEB应用所需时间精力，使得我们可以集中精力进行模型的持续改进工作中。并且，框架本身的功能结合互联网上的一些资源，使得前端的展示页面能够拥有漂亮简洁的UI界面。

**5、良好的交互效果**

为了提高用户的交互体验，我们设计了清晰友好的输入界面和简单明了操作方式。自动提取出的人物和言论的输出结果也以非常直观的方式显示在WEB页面上。用户可以很方便很直观地确认到提取结果。

为了进一步提高交互效果，我们设计了两种不同的用户交互模式。

第一种是用户手动输入新闻原文。在这种交互方式下，用户可以自行选择输入内容，复制粘贴到输入框，然后按下蓝色按钮即可获得摘要结果。这种交互方式允许用户选择指定的输入内容，从而能达到更好的有方向性的测试效果。

另一种方式是随机生成新闻用于输入。用户在这种模式下不需要手动搜索新闻，也不需要输入文本，只需要点击一下绿色按钮即可得到输出的摘要结果。自动输入的新闻标题和正文文本是由程序自动从新闻网站抓取。这种交互模式对用户非常友好，省时省力。用户可以在段时间内快速查看多次人物言论的提取结果。

**【有待改进的地方】**

1. **在某些情况下，言论提取的效果不是很理想。**

・没有表示说的词，言论提取不到

・连续句子，没有“ ”，言论没提取

・有一些结尾或是开头有XX报道，.... 后面的一些记者信息也会被提取出来，因为有‘报道’，是说的近义词

・某些句子可能存在重复提取的问题，如果句子复杂，分词复杂

1. **模型准确度问题。**

模型是使用bert构造的。但由于训练用的数据的问题，模型的准确度有限。

1. **算力限制**

服务器空间不足，并且GPU也无法使用。这导致了的模型准确度受限于计算资源，无法进一步提高精确度。

1. **代词指代消解问题**

代词指代消解还存在问题。目前没找到合适的中文开源库，代码只是比较粗略的做了。作为今后的课题，此问题还有进一步改善的空间。

以上关于本次项目四成果的分析报告。

谢谢各位老师的指导和辛勤付出。

队名：冯・诺依曼队

成员：李鹏飞、吕汶颖、刘迪

时间：2020年3月18日