**开课吧人工智能课学院**

**自然语言片与人工智能第5期 项目一报告书**

项目名：非监督文本自动摘要模型的构建

队名：冯・诺依曼队

在本项目中，我们实现了一个非监督文本自动摘要模型的构建，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一成果的优点和不足之处。

**【优点】**

**1、使用多种方法的融合实现了较高的准确性**

为了提高自动摘要结果的准确性，我们除了使用了项目指导书中提示的词向量训练、SIF句子向量化、KNN平滑等方法以外，还根据本项目自身的特点和实际需要，在数据预处理部分和相关度计算的部分中，灵活地使用了“停用词处理”、“句子向量和标题的相关度计算”、“文章关键字提取”等方法，在一定程度上有效地提高了自动摘要的输出结果的准确性。

**2、高适应性和高容错性**

为了使得我们构建的自动摘要模型能够适用于不同长度的输入文章，我们将输出摘要的长度设计成了根据输入文章长度自动变化的变量。并且，还将两个重要参数也设计成了根据输入文本长度自动调节的变量。一个参数是用在句子向量化处理等部分的参数「学习率α」，另一个参数是KNN平滑处理中用到的「n\_neibours值」。

从而，当输入文章的长度变化时，我们构建的模型可以在无需修改代码和无需手动调整参数的情况下，做到输出的摘要长度能够自动匹配当前输入文本长度，并且通过参数的自动调整来最大限度地获得较高的准确度。

除此之外，考虑到输入文章有时可能出现“没有标题”等与模型处理的预期输入不符的情况，我们增加了一些容错处理，以使得在输入文章无标题等一些预期外的情况下，也能保证程序不出现运行异常，并能自动生成相对准确的摘要。

**3、考虑了语义连贯性和输出摘要的可读性**

为了提高摘要的语义连贯性，我们使用了项目指导书里提示的KNN平滑等方法，并且通过优化处理逻辑以及反复地调参达到了相对理想的结果。

在最初的模型调试阶段，输出的摘要是没有标点符号的，这影响了摘要的可读性。为了增加输出摘要的可读性，我们做了一些工作，使得摘要中能够包含位置准确的标点符号。

摘要中加标点的功能主要是通过先将原文各个句子所含标点及其位置记录，再在摘要生成阶段重新添加到句子中实现的。另外对于一些特殊情况和特殊位置，则是根据规则自动生成标点符号。

**4、性能较高、响应快速**

由于模型的构建采用的都是简单但有效的方法的组合，并且在程序设计阶段做了架构上的一些调优，最终的自动摘要系统性能高、响应快速。

在测试环境下，程序加载启动速度较快，只有5秒。并且程序处理速度极快，在按下摘要生成按钮后，几乎是直接返回结果，响应时间估算大概小于0.2秒。

对于自动搜索新闻并随机输入的功能来说，经过对新浪新闻、网易新闻等新闻源网站的测试，摘要生成的速度也很理想。

**5、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。

**6、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。自动摘要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摘要结果。

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的标题和正文。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摘要生成结果。

**【有待改进的地方】**

1. **语言模型训练用的语料库的数据量不足。**

受限于训练语言模型用的机器性能，用来训练语言模型的语料库的数据量比较有限。如果能够使用更大的数据量来训练模型的话，可能会对提高摘要准确度有帮助。

1. **句子的相关度计算模型中标题与文本的权重值未完全达到最优。**

在句子的相关度（Ranking值）计算模型中，同时计算了句子对于标题与正文内容的相关度，然后利用权重进行了加权计算。当前标题相关度的权重是0.1，正文相关度的权重是0.9。当前的这个权重值设置是根据多次调参的结果最终确定下来的相对来说比较合适的值。

但是，从调参的过程来看，即使是权重值的微小调整也会在很大程度上使得最终的摘要准确度发生很大的变化。由于更精确地调整权重值的困难性，我们没能继续优化权重值或者调整模型。我们认为这部分还存在可以改进的余地。

**3、框架的运用和前端展示及交互性方面的不足。**

Flask、Bottle和 Bootstrap框架是在人工智能领域较多使用的框架。使用这些框架可以快速地将AI模型转化为可以演示的WEB应用程序，并且可以轻松地设计出良好的用户交互界面以及美观的UI界面。但是，在完成本次项目的过程中，由于项目时间有限，而且我们小组成员并不十分熟悉这些推荐的框架，因此，我们把主要精力放在了模型的设计和完善方面，只使用了较为熟悉的Django作为前端的框架。虽然使用Django框架也能够快速构建出符合基本展示要求的WEB应用程序，但是对于未能学习并尝试Flask、Bottle和Bootstrap框架来探索更多的展示及交互方面的可能性这一点来说，还是稍微有一些遗憾。希望能在后续的项目改进或者项目二、项目三中，可以有机会弥补这个遗憾。

1. **语义连贯性问题仍然存在。**

虽然我们通过KNN平滑、相关值计算模型优化等多种方法尽可能地提高了摘要的语义连贯性，但是，从输出的摘要的结果来看，始终还是无法完全做到每一句都像人写那么自然连贯。在语义连贯性方面，我们认为还存在着改进的空间。

以上关于本次项目一成果的分析报告。

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：2019年12月2日