大作业报告

小组信息：人数：3人

成员：黄琰琛 211250100 [211250100@smail.nju.edu.cn](mailto:211250100@smail.nju.edu.cn)

职责：数据扩增

左皓升 211250074 [211250074@smail.nju.edu.cn](mailto:211250074@smail.nju.edu.cn)

职责：数据分析及筛选

柏帅 211250071 [211250071@smail.nju.edu.cn](mailto:211250071@smail.nju.edu.cn)

职责：扩增检验

研究问题：如何通过有效的方法将国家公示的政务信息进行扩增。

研究方法：

数据获取：在中国政府网站上，通过爬虫等技术手段获得了近10天内的政务公示文章，并保存在txt文件中。然后通过Python文件处理方式对txt内的数据进行处理。具体方式是将txt内的数据按行读出，根据句号将数据进行拆分，然后逐行写入另一个dest.txt文件当中。

数据筛选：以Python代码处理和人工判断结合的方式，对于字数较少无扩增意义的数据以及无效冗余的数据进行清洗和筛选。

数据扩增：主要使用回译的方式进行数据扩增。主要采用用Python调用翻译网站的api，对dest.txt文件内的数据进行两次翻译，以此获取扩增之后的数据。之所以选择回译的方式是因为该方法较易于操作，实现较为方便，且获得的新语料质量较高。具体操作为在main.py中创建了Translation类，该类下声明了一个静态方法translate用以进行中英及英中的翻译。然后在runTranslation中调用该静态方法，对dest.txt文件中的内容逐行回译，并最终写入result.txt文件中。

数据检验：使用jieba包获得原文本和扩增后文本的余弦向量，再计算出余弦相似度来度量扩增效果。具体操作是在main.py中创建了Judge类，在该类下声明了get\_word\_vector静态方法，然后在runTranslation中调用该静态方法，我们在result文件中对每一条语句按照原文本、扩增后文本、余弦相似度顺序写入。值得注意的是，这种方法在某些情况下并不可以完全反映出相似度。某些我们认为较为成功的扩增却计算出了较低的相似度。对此，我们也在思考更好的检验方法，在暑期学校中刚刚学习了一些更好的方法，但是还未有充足的时间来实现我们的想法。

案例分析：

本次实验主要动手实践了NLP模型训练中的一个重要环节，数据库的建立，也就是用有限的数据扩增出相对庞大的数据，从而优化模型训练过程，得到更加符合预期的模型。在实践过程中，小组成员相继学习了数据爬取、数据筛选和数据检验，并对NLP模型构建和实验方法有了初步的了解，结合数据科学基础课程中学习的概率论和统计学相关知识，加深了对大数据以及深度学习人工智能的认识。本次实验是从0到1的过程，虽然时间紧迫，过程中的技术细节被大幅度省略，但我们依旧对大数据研究方向产生了浓厚的兴趣。数据科学作为新时代互联网+技术的根基，有着至关重要的奠基作用，是未来信息社会发展的基石。我们小组成员想借此机会进一步探索NLP技术，对现在司空见惯但不明就里的搜索引擎和翻译工具等技术领域产生了浓厚的兴趣，希望有机会加入陈振宇老师的实验小组亲身体验大数据现代化发展的魅力。