中期报告

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：调用多个api进行了不同api翻译下的数据比较。但是目前存在的问题是数据扩增方法单一，且并未考虑到文本生成指标，后续将会补足这一短板，并且进行多个数据扩增方法间的文本比较。

案例分析：

对政务领域的文本经行扩增是一项很有意义的工作，它可以提升样本数量，提高文本匹配效果。同时由于我们选取的技术方案难度不大，可以较为轻松的完成文本扩增的工作，在实际的应用中也有一定的价值