1. **项目主要进展（附已发表论文首页复印件）**
2. **项目背景**

对新闻网站进行数据分析文本爬取并制作为训练语料，并利用分词器进行语料的清洗和分词处理，并使用Gensim模块训练词向量。将训练后得到的模型应用于错别字检测。

1. **项目进展**
2. **进展概述**

能够对提取出来的语料库进行进一步的清洗，训练得到相关词向量模型

能够利用训练好的词向量进行基本的错别字检查

1. **实现过程**

2.1语料Word2Vec训练

2.1.1获取语料库资源

从互联网中爬取相应的语料，本项目的在实现过程中对维基百科语料进行爬取并从原始的xml文件当中提取出标题和正文。

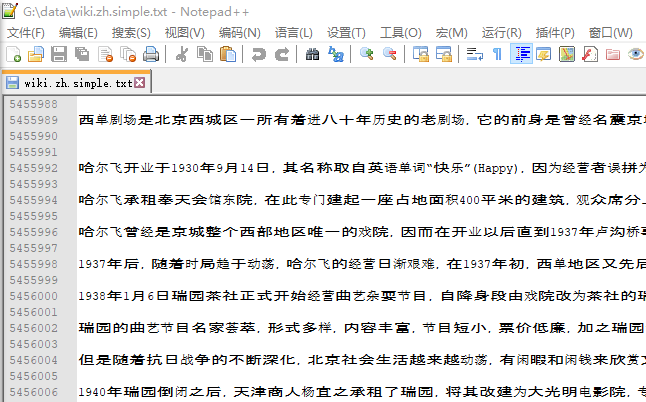
2.1.2化繁为简

提取出来的语料库中，存在一些繁体文字，会对训练产生影响，我们使用

开源的中文简繁转换项目Open Chinese Convert（OpenCC）命令语句

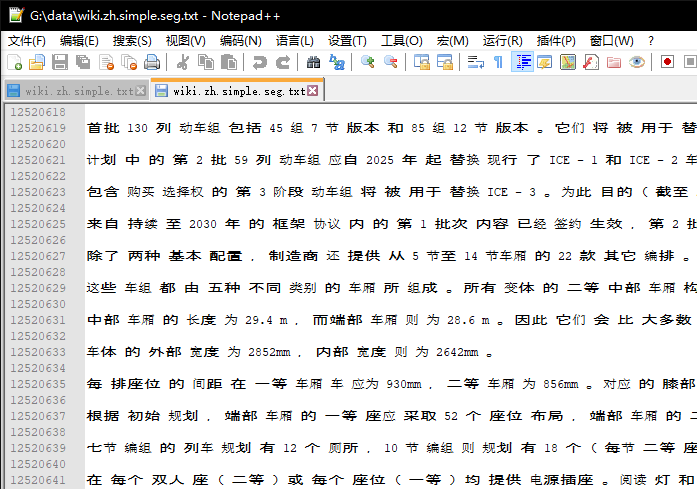
opencc -i input\_filename -o output\_filename -c t2s.json

将第二步得到的文本进行处理得到简体的语料：



* + 1. 分词操作

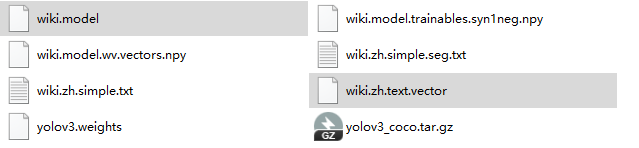
上一步得到的简化文本中还存在一些数字，标点符号，乱码字符等等不利于训练的信息，对此我们对上述语料进行非中文字符串的清理和分词。导入中文停用词列表，该表可以过滤出没有实际意义的功能词缩小从而训练范围提高训练精度。再分词过程中采用结巴分词方法，基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。分词结果如下：



* + 1. gensim训练词向量

gensim中封装了包括了word2vec, doc2vec等模型，word2vec采用了

CBOW(Continuous Bag-Of-Words，连续词袋模型)和Skip-Gram两种模型。我们采用word2vec模型进行训练，训练模型，设置词向量的维度大小为300， 迭代次数为8即将语料训练词向量8次，采用skip-gram模型。经过训练最终得到两个文件，Gensim中默认格式的词向量文件model，以及原始c版本的词向量文件vector。



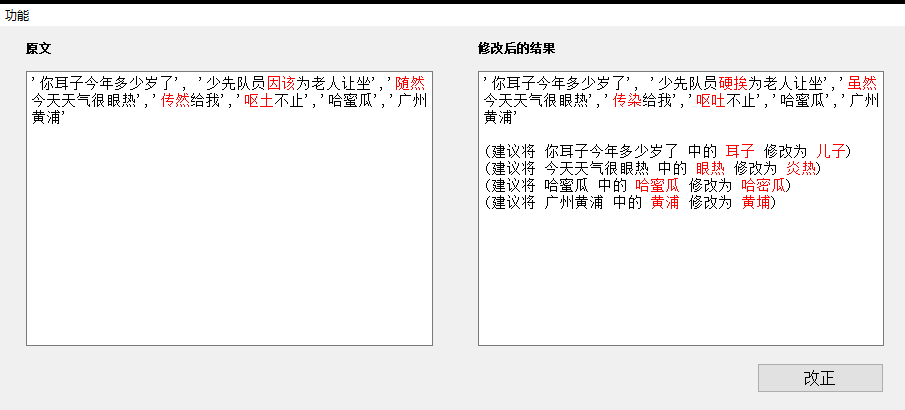
2.2利用得到的模型进行纠错

2.2.1 准备停用词表和词向量模型载入程序中，对于输入程序的语句进行停用词和分词的拆分。计算出词之间的距离，如果小于一定的值那么就判定错误，并将该词或字利用Python的pypinyin库进行转化，在词典中找到选词进行修改

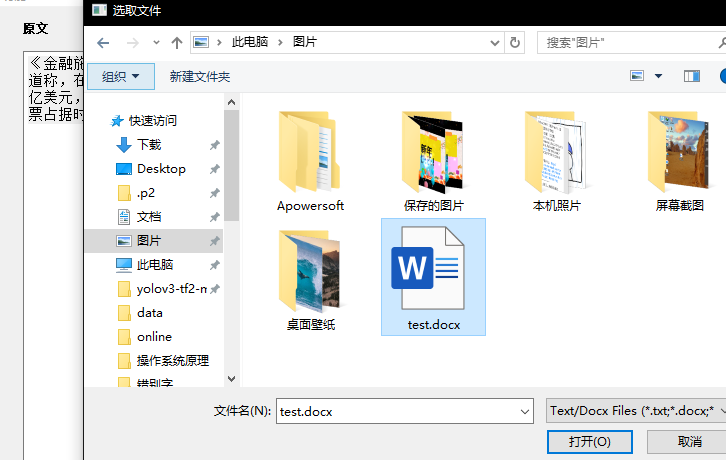
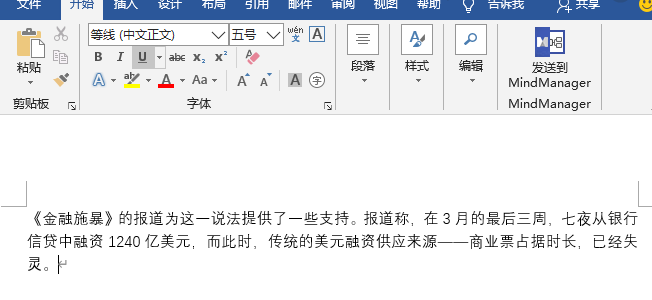
2.2.2程序界面

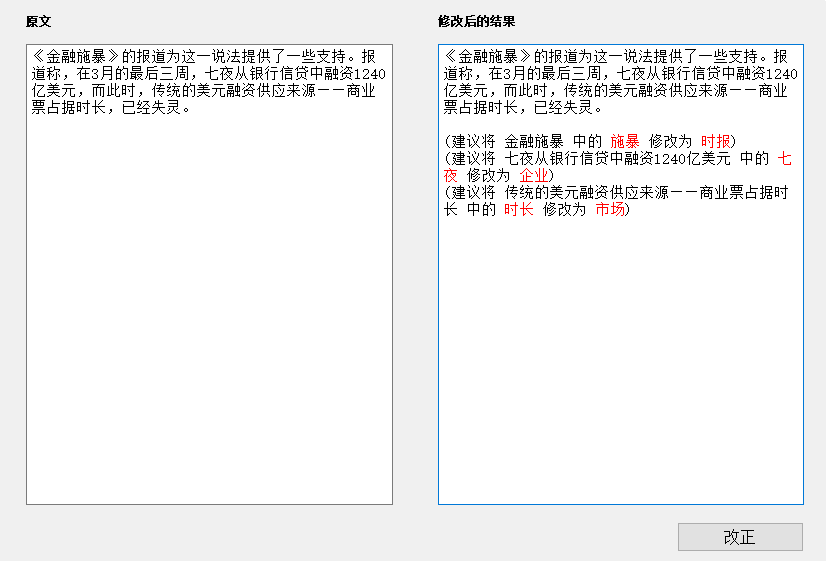
利用Python的PyQt5库进行程序的界面设计，可以实现对用户输入的句子和文档进行纠错。

2.3程序示例：







1. **实现工具**

**编译工具：Pycharm Python==3.7.4**

**Python库： jieba==0.42.1 pypinyin==0.37.0 PyQt5==5.13.0**

**genism==3.8.1**

**二. 下一步工作计划**

1. 扩大训练的语料库，增加训练量
2. 改善纠错方法，提高纠错率，增加纠错的方法
3. 对于词向量的使用进行改善
4. 对程序的界面进行进一步的设计
5. **经费使用情况和经费安排计划**

**资料费（资料查询，打印复印费用）：549元**

1. **存在问题、建议及需要说明的情况**
2. 纠错时对于连续错误的词语不能很好的识别出来
3. 还没有改进对于字形错误和一些逻辑错误
4. 对于文本量太大的情况，程序可能出现崩溃的情况