汉语电商评论关键词提取系统

📖 华北电力大学 数理学院 **“数智化”赋能乡村振兴实践团** 尹喆勋

📖 指导老师：雍雪林

> 前言：伴随着电子商务的高速发展，评论的数量达到一个新量级后，不论是对商家还是用户，都很难再进行有效的分析。对于商家来说，巨大的数据量导致处理起来很艰难。对于用户来说，评论内容冗杂，导致浏览评论费时费力。同时，这也会影响平台的用户体验，增加平台的用户流失风险。因此针对评论内容的关键词提取，对于评论的分析有重大的意义。

针对上述存在的问题，本项目针对电商评论数据的特征提取问题，建立了基于关键词提取的评论分析系统，设计并创新了一种基于 `word2vec` 与 `textrank` 的关键词提取方法。首先，通过基于 `Trie 树` 结构与动态规划查找实现的 `jieba分词` 系统将单句评论进行汉语分词，经由分词后的词汇集合再通过 `textrank` 算法推得单句摘要

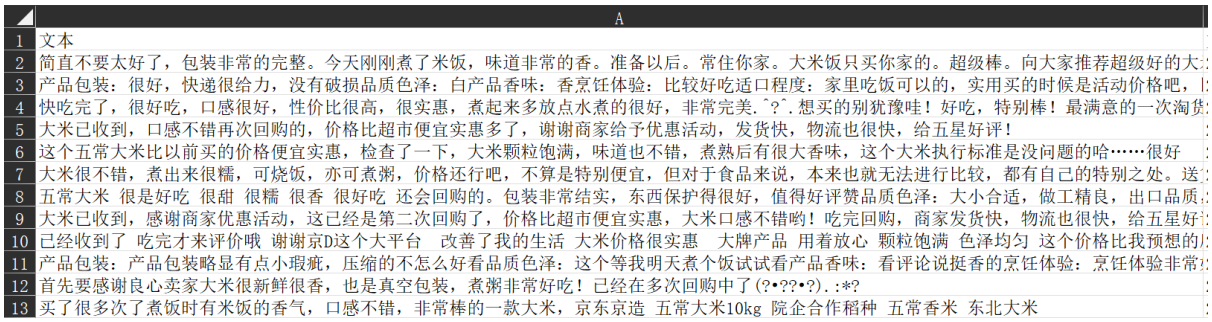
训练时，将摘要词汇通过 `word2vec` 转化为多维向量再引入多层神经网络模型来实现对其他评论的关键词提取。

使用说明

1. 首先准备需要进行关键词提取的评论数据表

   具体要求是文件需为标准excel表格

   要求**其中表格A列为评论数据**，如图



	A
1	文本
2	简直不要太好了，包装非常的完整。今天刚刚煮了米饭，味道非常的香。准备以后。常住你家。大米饭只买你家的。超级棒。向大家推荐超级好的大
3	产品包装：很好，快递很给力，没有破损品质色泽：白产品香味：香烹饪体验：比较好吃适口程度：家里吃饭可以的，实用买的时候是活动价格吧，
4	快吃完了，很好吃，口感很好，性价比很高，很实惠，煮起来多放点水煮的很好，非常完美.^?^.想买的别犹豫哇！好吃，特别棒！最满意的一次淘货
5	大米已收到，口感不错再次回购的，价格比超市便宜实惠多了，谢谢商家给予优惠活动，发货快，物流也很快，给五星好评！
6	这个五常大米比以前买的价格便宜实惠，检查了一下，大米颗粒饱满，味道也不错，煮熟后有很大香味，这个大米执行标准是没问题的哈……很好
7	大米很不错，煮出来很糯，可烧饭，亦可煮粥，价格还行吧，不算是特别便宜，但对于食品来说，本来也就无法进行比较，都有自己的特别之处。送
8	五常大米 很是好吃 很甜 很糯 很香 很好吃 还会回购的。包装非常结实，东西保护得很好，值得好评赞品质色泽：大小合适，做工精良，出口品质
9	大米已收到，感谢商家优惠活动，这已经是第二次回购了，价格比超市便宜实惠，大米口感不错哟！吃完回购，商家发货快，物流也很快，给五星好
10	已经收到了 吃完才来评价哦 谢谢京D这个大平台 改善了我的生活 大米价格很实惠 大牌产品 用着放心 颗粒饱满 色泽均匀 这个价格比我预想的
11	产品包装：产品包装略显有点小瑕疵，压缩的不怎么好看品质色泽：这个等我明天煮个饭试试看产品香味：看评论说挺香的烹饪体验：烹饪体验非常
12	首先要感谢良心卖家大米很新鲜很香，也是真空包装，煮粥非常好吃！已经在多次回购中了(?•??•?).:*?
13	买了很多次了煮饭时有米饭的香气，口感不错，非常棒的一款大米，京东京造 五常大米10kg 院企合作稻种 五常香米 东北大米

2. 启动程序（**目录在use/main.exe**）

   将目标文件选择为准备好的excel评论数据表

   选择好适当的统计信息保存目录与词图保存目录

3. 点击“开始进行评论关键词提取”等待程序运行完毕

   最终会在统计信息表格中记录所有摘要关键词出现次数

   同时会显示最终得到的关键词词图

文件说明

source 文件夹

- main.py 主程序文件，包含GUI界面与完整的模型训练与模型应用代码
- training.py 训练程序文件，包含完整的模型训练程序，可以根据需要自行搭建训练环境进行修改

use 文件夹

- trained_model文件夹 包含全部模型文件与数据集，可经过 training.py 程序生成
- **main.exe 主程序文件**
- 其他文件均为python与第三方库依赖项

训练评论集.xlsx 用于训练测试的评论数据完整版

训练评论集2000.xlsx 用于训练测试的评论数据

额外信息

以下是对神经网络实现的具体解释

`x = GlobalAveragePooling1D()(dot_product):`

这一行代码使用 GlobalAveragePooling1D 层对 dot_product 进行全局平均池化。dot_product 是一个矩阵，每一行对应一个词向量的点乘结果。全局平均池化对矩阵的每一行求平均值，将矩阵转换为一个一维向量 x。这样将每个句子中的词语对应的点乘结果取平均，得到了句子的向量表示。

`x = Dense(64, activation='relu')(x):`

这一行代码定义了一个全连接层 Dense，并对前面得到的句子向量 x 进行变换。全连接层是神经网络中常用的一种层，每个神经元与上一层的所有神经元相连。这里定义了一个包含 64 个神经元的全连接层，并使用 relu 作为激活函数。relu 是一个常用的非线性激活函数，它将负值设为零，保留正值不变。

`output = Dense(1, activation='sigmoid')(x):`

这一行代码再定义了一个全连接层 Dense，用来生成模型的输出。这里的目标是进行二分类任务，因此定义了一个只有一个神经元的全连接层，并使用 sigmoid 作为激活函数。sigmoid 激活函数将输出限制在 0 到 1 之间，可以用来表示概率值。

`model = Model(inputs=[input_target, input_context], outputs=output):`

这一行代码创建了一个 Model 对象，用来定义整个神经网络模型。我们将的 input_target 和 input_context 作为输入层，output 作为输出层，形成一个端到端的模型。这样就构建了一个用于学习句子关键词提取的神经网络模型

这些代码构成了一个简单的神经网络模型，用于学习句子关键词的提取任务。在模型的输出层，我使用 sigmoid 激活函数，因为目标是对每个词语预测是否是句子的关键词（二分类任务）同时在训练过程中，使用了正样本和负样本对来进行监督学习，通过优化损失函数来训练模型。

关于二次开发

本汉语电商评论关键词提取系统全部源代码依据MIT开源协议开放

```
# Copyright(c) 2023, KaoruXun(尹喆勋)
# Developed for the "digital intelligence" empowerment rural revitalization
# practice group of North China Electric Power University
```

Python开发环境配置需要以下第三方库：

openpyxl	用于Excel文件读写
jieba	用于文本分词与特征提取
numpy	用于人工神经网络训练
tensorflow	用于模型训练与复现
pickle	用于对于分词向量数据序列化与反序列化
wordcloud	用于生成词图
tkinter	用于GUI用户界面