Text Miner 文本矿工用户操作指南

作者：吴林瀚 201707020322 计科1704

1. 系统概述

1.1 目标

在现实世界中，可获取的大部分信息是以文本形式存储在文本数据库中的，由来自各种数据源的大量文档组成，如新闻文档、研究论文、书籍、数字图书馆、电子邮件和Web页面。由于电子形式的文本信息飞速增涨，文本挖掘已经成为信息领域的研究热点。

面对大量的文本数据，人们迫切需要有文本挖掘系统来进行文本数据挖掘和可视化分析，更简便更直观地提取文本信息。本系统 Text Miner 文本矿工的目标正是满足人们简便快捷地完成文本挖掘的需求，通过本系统可以做到实时完成文本数据的信息挖掘和数据可视化分析，更加快速高效地完成工作。

1.2 背景

文本挖掘是信息挖掘的一个研究分支，用于基于文本信息的知识发现。文本挖掘是抽取有效、新颖、有用、可理解的、散布在文本数据中的有价值知识，并且利用这些知识更好地组织信息的过程。目前，很多文本挖掘算法已经通过 python 实现，并且将代码打包开源，例如 jieba，snownlp 和 wordcloud 等优秀的文本挖掘开源第三方库包。本系统将基于这些优秀的开源第三方库包完成对文本数据的挖掘工作。

数据可视化主要旨在借助于图形化手段，清晰有效地传达与沟通信息，是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。目前已经存在很多基于 python 的数据可视化工具和开源第三方库包，例如 Streamlit 等数据可视化工具已被广泛使用。本系统将基于 Streamlit 搭建数据可视化平台并完成文本挖掘的实时可视化呈现。

1.3 功能概述

本系统可进行的文本挖掘功能如下：

- **关键词抽取**：对长文本的内容进行分析，输出能够反映文本关键信息的关键词。
- **情绪判断**：对文本进行情感倾向判断，将文本情感分为正向、负向、中性。用于口碑分析、话题监控、舆情分析。
- **生成词云**：通过形成"关键词云层"或"关键词渲染"，对文本中出现频率较高的"关键词"的视觉上的突出。通过过滤掉大量的文本信息，使浏览者只要一眼扫过词云就可以领略文本的主旨。

2. 安装方法

2.1 方法一：联网运行（推荐）

直接访问系统网页：https://wulh-textminer.herokuapp.com/

这是最推荐的运行方法，因为你不需要进行任何的环境配置和源代码获取，就能运行本系统并且使用全部功能。

本系统 Text Miner 文本矿工已经部署在 Heroku 云平台上。Heroku 是一个支持多种编程语言的云平台即服务。通过在 Heroku 上创建一个新的 APP 并将 Text Miner 文本矿工进行部署，我们可以在任何一台已联网并可以浏览网页的设备上（电脑、手机、iPad等设备）运行本系统。

2.2 方法二：本地运行

2.2.1 环境配置

在运行本系统 Text Miner 文本矿工之前必须完成一定的环境配置，由于本系统是基于 Python 3 和一些 Python 的第三方模块完成的，所以需要事先配置好 Python 3 环境和 pip 软件包安装程序，进而可以使用 pip 从 Python 软件包索引和其他索引安装软件包，从而导入 Python 的第三方模块。

完成 Text Miner 文本矿工系统时所使用的 Python 3 版本和 pip 版本仅供参考：

- Python 3.7.2
- pip 20.3.1

2.2.2 运行方法

获取 Text Miner 文本矿工系统源代码的方法：

1. 通过 GitHub Repository : https://github.com/WilliamWuLH/TextMiner Fork、clone 或者 Download 系统源代码
2. 可以通过 http://www.wlhan.top/#contact 中的联系方式或者发送邮件到 wulhan@outlook.com 和系统开发者（我本人）取得联系。

获取到 Text Miner 文本矿工系统的源代码之后，运行 TextMiner_Runner.py 程序，命令如下：

```
cd TextMiner
py TextMiner_Runner.py
```

TextMiner_Runner.py 程序会自动通过 pip 安装本系统所必要的软件包和第三方模块，相应的软件包和参考版本号如下：

- streamlit==0.72.0
- pandas==0.25.3
- numpy==1.19.3
- jieba==0.42.1
- snownlp==0.12.3
- matplotlib==3.1.1
- wordcloud==1.8.1

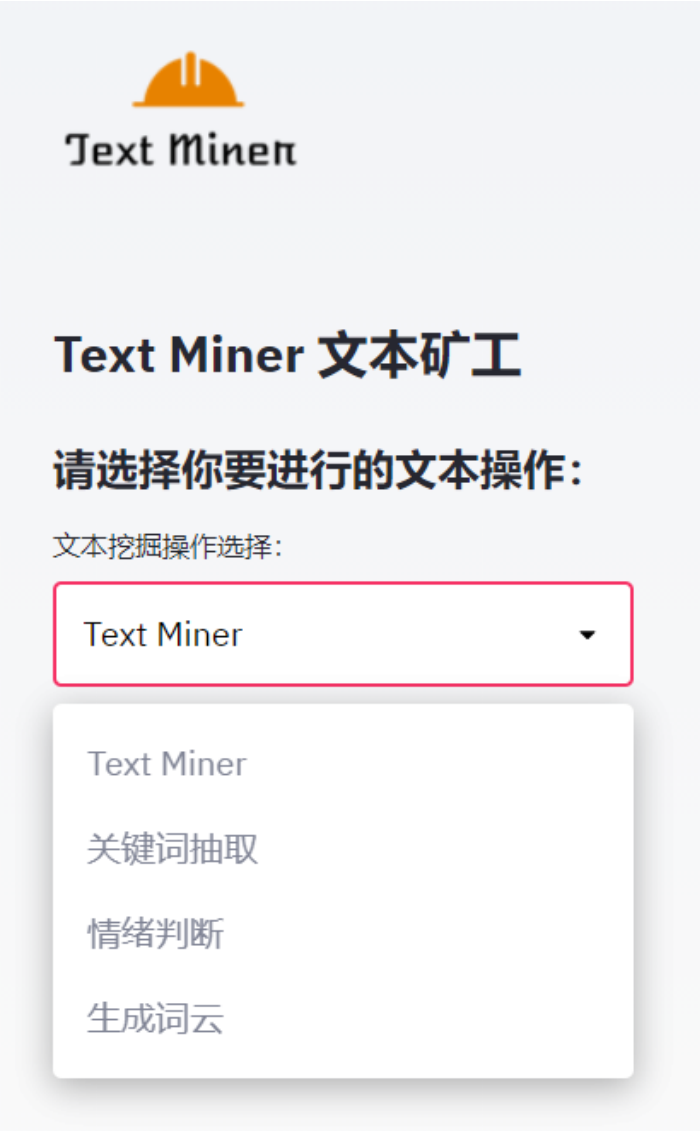
配置好软件包和第三方模块之后，TextMiner_Runner.py 程序会自动启动运行 Text Miner 文本矿工系统，正常运行时会自动在你的默认浏览器中弹出 Text Miner 文本矿工系统界面，你便可以在 Text Miner 文本矿工系统界面中完成各种文本挖掘功能操作。当然，你也可以根据运行系统时命令行中所给的提示自行打开 Text Miner 文本矿工系统界面并完成各种操作。

3. 使用说明

3.1 功能选择操作说明

在 Text Miner 文本矿工系统页面的左侧的选项栏中可以选择我们要完成的文本挖掘操作，可供选择的文本操作有：

1. Text Miner：Text Miner 文本矿工系统主页面，介绍系统的相关信息。
2. 关键词抽取：对输入的文本进行实时的关键词抽取，并实时进行数据的可视化展示。
3. 情绪判断：对输入的文本进行实时的情绪判断，并实时进行数据的可视化展示。
4. 生成词云：对输入的文本进行实时的词云生成，并实时展示。



3.2 关键词抽取功能操作说明

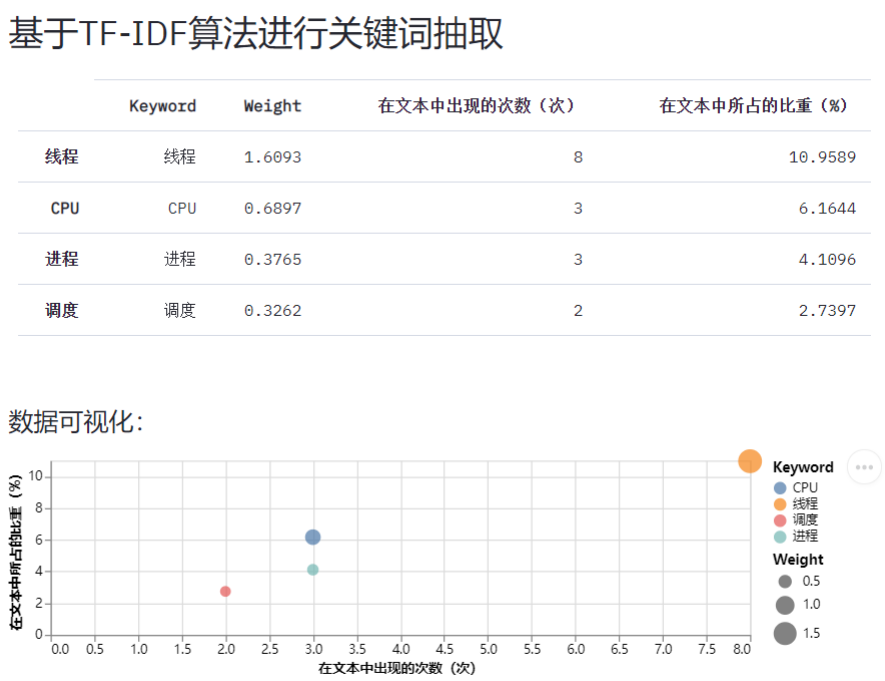
在 Text Miner 文本矿工系统页面的左侧的选项栏中选择 “关键词抽取”，可以进入关键词抽取的界面。

关键词抽取功能的输入：原始文本、要抽取出的关键词数量

使用关键词抽取功能需要输入原始文本，原始文本最好是中文文本，在系统的关键词抽取界面已经提供了原始文本样例，并作为预设输入。使用关键词抽取功能还需要输入要抽取出的关键词数量，在系统中的预设输入是 4，并且最小值为 1。

关键词抽取功能的输出：基于TF-IDF算法的关键词抽取、基于TextRank算法的关键词抽取

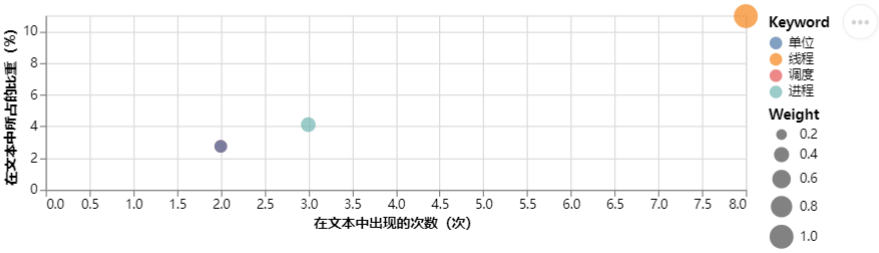
根据输入的原始文本和要抽取出的关键词数量，能够基于TF-IDF算法和基于TextRank算法分别实时进行关键词抽取，输出两种算法抽取出的关键词和对应的数据信息，并且实时进行数据可视化展示。



基于TextRank算法进行关键词抽取

	Keyword	Weight	在文本中出现的次数（次）	在文本中所占的比重（%）
线程	线程	1	8	10.9589
进程	进程	0.3789	3	4.1096
调度	调度	0.2847	2	2.7397
单位	单位	0.2793	2	2.7397

数据可视化：



3.3 情绪判断功能操作说明

在 Text Miner 文本矿工系统页面的左侧的选项栏中选择 “情绪判断”，可以进入情绪判断的界面。

情绪判断功能的输入：要进行情绪判断的所有句子

使用情绪判断功能需要输入要进行情绪判断的所有句子，输入的句子格式有所要求，不同句子之间需要用 $ 隔开，在系统的情绪判断界面已经提供了输入样例，并作为预设输入，并且系统会实时展示你所输入的句子。

情绪判断

请输入您要判断的所有句子，不同句子之间请用 $ 隔开！

输入样例（包含有 8 个句子）：

这部电影真心棒，全程无尿点$这部电影简直烂到爆$这部电影真不错$太差劲了吧这个$我觉得不行啊$这部电影我爱了$这部电影不太行$这部电影一般般吧

请输入您要判断的所有句子：

这部电影真心棒，全程无尿点$这部电影简直烂到爆$这部电影真不错$太差劲了吧这个$我觉得不

您输入的所有句子如下：

	0
0	这部电影真心棒，全程无尿点
1	这部电影简直烂到爆
2	这部电影真不错
3	太差劲了吧这个
4	我觉得不行啊
5	这部电影我爱了
6	这部电影不太行
7	这部电影一般般吧

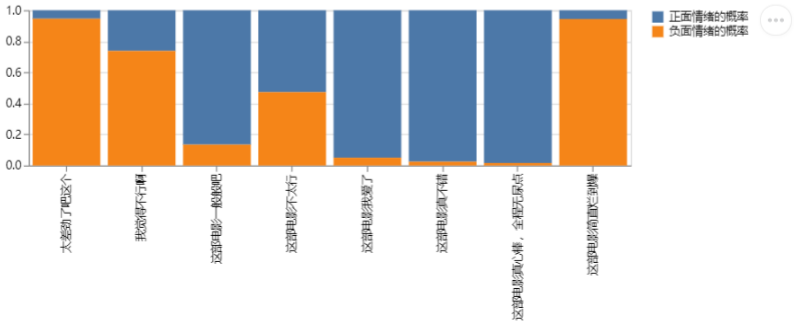
情绪判断功能的输出：每个句子的情绪判断概率结果、每个句子的情绪评价

根据输入的要进行情绪判断的所有句子，系统可以对每一个句子进行实时的情绪判断，输出情绪判断的概率结果并且实时完成数据可视化，同时系统还可以完成对每一个句子的情绪评价，并且实时进行数据可视化展示。

情绪判断

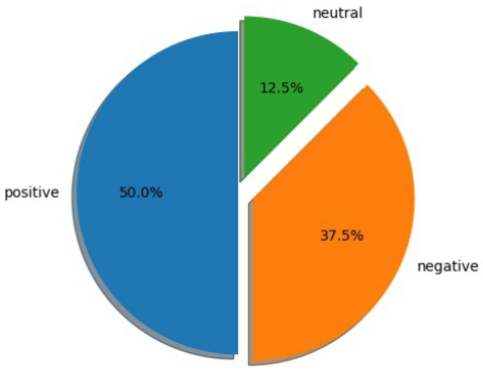
	正面情绪的概率	负面情绪的概率
这部电影真心棒，全程无尿点	0.9843	0.0157
这部电影简直烂到爆	0.0567	0.9433
这部电影真不错	0.9757	0.0243
太差劲了吧这个	0.0545	0.9455
我觉得不行啊	0.2617	0.7383
这部电影我爱了	0.9498	0.0502
这部电影不太行	0.5280	0.4720
这部电影一般般吧	0.8657	0.1343

数据可视化：



	情绪评价
这部电影真心棒，全程无尿点	正面评价
这部电影简直烂到爆	负面评价
这部电影真不错	正面评价
太差劲了吧这个	负面评价
我觉得不行啊	负面评价
这部电影我爱了	正面评价
这部电影不太行	中立评价
这部电影一般般吧	正面评价

数据可视化：



3.4 生成词云功能操作说明

在 Text Miner 文本矿工系统页面的左侧的选项栏中选择 “生成词云”，可以进入生成词云的界面。

生成词云功能的输入：要生成词云的文本

使用生成词云功能需要输入要生成词云的文本，在系统的生成词云界面已经提供了输入样例，并作为预设输入。

生成词云功能的输出：生成的词云图

根据输入的要生成词云的文本，可以实时生成文本所对应的词云图，实时生成词云图的过程比较缓慢，需要耐心等待。