

**项目编号：001**

|  |  |  |
| --- | --- | --- |
| **文件状态**  [ ] 草稿  [ ] 讨论稿  [√] 正式发布 | **文件编号** | 001 |
| **文档类型** | Word |
| **版 本** | V1.0 |
| **完成日期** | 2019年8月30日 |

基于kaggle数据集的数据分析

软件需求规格说明书

**组长：何 钺 1120161980 08111604**

**组员：柴增豪 1120161906 08111602**

**张洛汐 1120161998 08111604**

**周彬韬 1120162003 08111604**

**代雨洁 1120162011 08111605**

**所属技术领域：机器学习，自然语言处理**

**承担部门：padle\_padle**

**项目负责人：何钺**

**项目起止年限：2019.08.26 – 2019.09.13**

|  |  |  |  |
| --- | --- | --- | --- |
| **版本** | **日期** | **更改人** | **描述（注明修改的条款或页）** |
| V1.0 | 2019.08.28 | 何钺 | 第一版 |
| V2.0 | 2019.08.30 | 何钺 | 第二版 |
|  |  |  |  |

**二〇一九年八月**

目录

[第一章 概述 1](#_Toc19134750)

[1.1 目的 1](#_Toc19134751)

[1.2 项目背景 1](#_Toc19134752)

[1.3 范围 1](#_Toc19134753)

[1.4 术语定义 2](#_Toc19134754)

[第二章 系统说明 3](#_Toc19134755)

[2.1 系统概述 3](#_Toc19134756)

[2.2 产品功能 3](#_Toc19134757)

[2.2.1 用户角度 3](#_Toc19134758)

[2.2.2 系统角度 3](#_Toc19134759)

[2.3 局限性 3](#_Toc19134760)

[第三章 软件需求说明 4](#_Toc19134761)

[3.1 功能要求 4](#_Toc19134762)

[3.1.1 项目结构图 4](#_Toc19134763)

[3.1.2 展示系统用例图 4](#_Toc19134764)

[3.1.3 功能模块描述 6](#_Toc19134765)

[3.2 输入输出要求 7](#_Toc19134766)

[3.2.1 数据导入模块 7](#_Toc19134767)

[3.2.2 搜索模块 7](#_Toc19134768)

[3.2.3 模型导入模块 7](#_Toc19134769)

[3.3 故障处理要求 7](#_Toc19134770)

[3.4 可用性 8](#_Toc19134771)

[3.5 可靠性 8](#_Toc19134772)

[3.6 灵活性 8](#_Toc19134773)

[3.7 设计约束 8](#_Toc19134774)

[3.7.1 开发环境 8](#_Toc19134775)

[3.7.2 使用环境 9](#_Toc19134776)

[3.8 界面要求 9](#_Toc19134777)

[3.9 控制 10](#_Toc19134778)

[3.10 设备 10](#_Toc19134779)

[第四章 验收标准 11](#_Toc19134780)

[第五章 参考资料 13](#_Toc19134781)

1. 概述
   1. 目的

* 创建模型：根据Kaggle比赛Crowdflower Search Results Relevance的数据集训练模型，使得模型能对用户输入的关键字和搜索引擎搜索结果的相关度进行评估。
* 对模型进行展示：用户输入关键字，系统根据模型对资料库中的所有产品进行相关度的评估，按照相关度从高到低返回搜索结果。通过这种方式，让用户直观地感受模型的性能。
* 协助评估搜索算法的性能：本项目为开源项目，通过模型对搜索引擎的返回结果进行评估，此项服务可以协助开发搜索算法的企业评估搜索算法的性能。
  1. 项目背景

Kaggle是由联合创始人、首席执行官安东尼·高德布卢姆（Anthony Goldbloom）2010年在墨尔本创立的，主要为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台。该平台已经吸引了80万名数据科学家的关注，这些用户资源或许正是吸引谷歌的主要因素。该项目的数据集正是来源于Kaggle平台的比赛Crowdflower Search Results Relevance。

搜索算法在日常生活中是极其常见的。无论我们是想在视频网站搜索相关的视频，还是在淘宝上购买衣服，搜索结果与搜索关键字的相关性通常是影响用户体验的关键原因之一。目前，小型在线企业没有很好的方法来评估其搜索算法的性能，这使他们难以提供卓越的客户体验。

本次项目的目标是依据比赛Crowdflower Search Results Relevance提供的数据集，创建一个开源模型，用于衡量搜索结果的相关性。 通过这样做，我们将帮助小型企业所有者为用户提供与资源更丰富的竞争对手提供的服务相同的体验。 它同时可以为更成熟的企业提供一个可以测试的模型。

* 1. 范围

本文档的预期读者：

1. 项目经理：项目经理可以根据该文档了解预期产品的功能，以进行项目管理，规划项目进度。
2. 设计人员：对需求进行分析，并设计出系统，设计包括：UI设计、系统架构设计等。
3. 开发人员：配合文档，了解系统功能，进行编码，实现功能。
4. 测试人员：根据本文档编写测试用例，并对软件产品进行功能性测试和非功能性测试。
5. 客户：根据本文档，判断是否符合自己的需求，并且协助制定验收标准，为之后的项目验收提供依据。
   1. 术语定义

|  |  |
| --- | --- |
| **术语** | **解释** |
| 用户 | 该产品的主要服务对象。 |
| 前端 | 用以展示用户界面，接收用户输入，展示搜索结果。 |
| 模型 | 通过Kaggle提供的训练集数据训练得到的模型。 |
| 资料库 | 由Kaggle提供训练集及测试集数据合并而来，当用户搜索关键词时，检索该资料库并返回结果。 |
| 后台/后端 | 对用户输入进行处理，运用模型以及资料库得到用户搜索结果并返回数据至前端。 |

1. 系统说明
   1. 系统概述

本项目是一个基于Python3.6的系统，运用TensorFlow, scikit-learn等机器学习库训练得到模型，通过模型对搜索结果进行相关度评估；同时，系统以“用户输入关键词，系统返回排序后的搜索结果”的形式，让用户能直接地感受到模型的准确性。

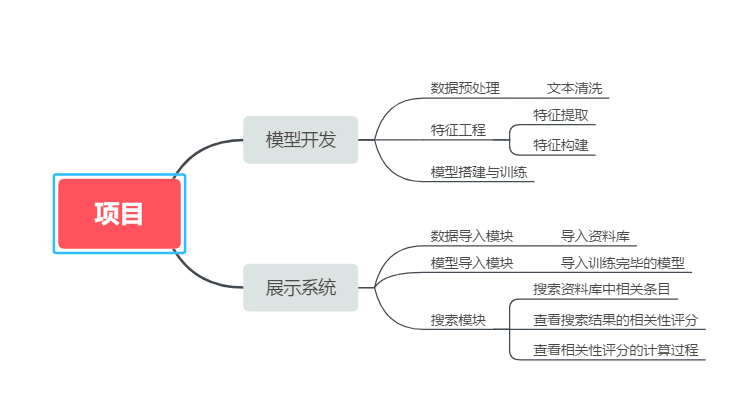
* 1. 产品功能
     1. 用户角度

1. 用户能搜索资料库，得到资料库中与搜索关键词相匹配的结果；
2. 对于每个搜索结果，用户能查看搜索结果与搜索关键词的相关性评分；
3. 对于每个搜索结果，用户能查看相关性评分的计算过程。
   * 1. 系统角度
4. 能支持用户完成其相应的业务
5. 能储存、管理资料库
   1. 局限性
6. 该系统的搜索功能旨在直观展示模型的性能，与真正的搜索功能本质上不一样。
7. 对于不具备一定专业知识的用户来说，需要一定的时间来学习使用和理解该产品。
8. 软件需求说明
   1. 功能要求

基础功能：能够从不同的关键字匹配出相对应的搜索结果，并给每个搜索结果赋予一个经过模型预测分析后的相关度的值。

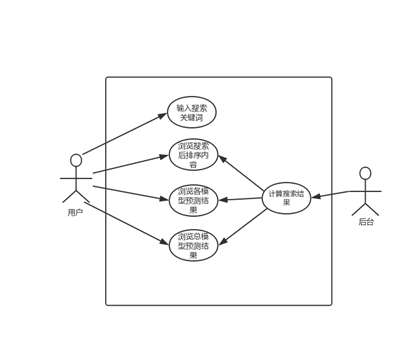
扩展功能：将关键字以列表形式统一展现出来，并能够将它的所有搜索结果逐一列出，每个搜索结果有一个总的相关度给分，每个分数还能查看使用过的每个模型得分。

* + 1. 项目结构图



* + 1. 展示系统用例图

**用例图**



**参与者**

本系统的参与者主要指使用该系统观察预测crowdflower词条匹配数据的人群，“后台”是一种抽象，完成由系统自动处理的功能，譬如数据和训练模型导入的功能，反馈给前端不同词条的计算结果，将前端展示给用户。

**用例图描述**

|  |
| --- |
| 用例编号：UC01 |
| 用例名称：搜索内容 |
| 参与者：用户 |
| 简要说明：搜索关键词。 |
| 基本事件流：用户在搜索栏键入关键词并搜索，可得到资料库中与搜索关键词相匹配的结果。 |
| 扩展事件流：无。 |
| 关系描述：无。 |
| 前置条件： 预测数据集已成功导入系统。 |
| 后置条件：无。 |
| 限制：输入信息仅包括大小写、数字、空格，不支持中文及乱码等。 |

|  |
| --- |
| 用例编号：UC02 |
| 用例名称：浏览相关性评分 |
| 参与者：用户 |
| 简要说明：用户可查看各搜索结果的相关性评分。 |
| 基本事件流：用户在搜索结果页面，可查看到各结果的综合相关性评分。 |
| 扩展事件流： 系统计算保留该搜索结果下的各模型的相关性评分。 |
| 关系描述：无。 |
| 前置条件： 预测模型正常导入，系统正常运行。 |
| 后置条件：无。 |
| 限制： 必须在用户搜索关键字后跳转的页面中，有搜索结果的前提下才能看见相关性评分。 |

|  |
| --- |
| 用例编号：UC03 |
| 用例名称：浏览详情页信息 |
| 参与者：用户 |
| 简要说明：用户在某一搜索结果下可查看其各预测模型下的预测值。 |
| 基本事件流：用户在搜索结果页面，选择搜索结果列表里的某一项搜索结果，可跳转到详情页查看到每个预测模型具体数值以及该项搜索结果的具体信息。 |
| 扩展事件流：无。 |
| 关系描述：无。 |
| 前置条件： 系统正常运行。 |
| 后置条件：无。 |
| 限制：搜索结果页面成功呈现出计算相关度评分结果。 |

* + 1. 功能模块描述

**搜索资料库**

|  |  |
| --- | --- |
| 描述 | UML活动图及时序图 |
| **时序图综述**：该时序图描述了“搜索资料库”的过程，涉及用户、后台两个对象。  **参与者对象描述**： “用户”、“后台”是参与者。  **消息描述**：用户打开搜索页面，选择关键词，传递给后台。后台进行处理，将匹配关键字的资料库内容以及计算的相应的相关度结果返回给用户。 |  |

**获取相关性评分的计算过程**

|  |  |
| --- | --- |
| 描述 | UML活动图及时序图 |
| **时序图综述**：该时序图描述了“获取相关性评分计算过程”的过程，涉及用户、后台两个对象。  **参与者对象描述**： “用户”、“后台”是参与者。  **消息描述**：用户点击任意资料库搜索结果项，后台返回各个模型计算出的相关度以及该搜索结果项的全部信息，最终将这些数据展示给用户。 |  |

* 1. 输入输出要求
     1. 数据导入模块

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段** | **媒介体** | **数据类型** | **数值范围** | **备注** |
| 资料库 | 磁盘读取 | .csv文件 | N/A | 文件内容仅包含数字、大小写英文字母及英文常见标点符号。 |

* + 1. 搜索模块

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段** | **媒介体** | **数据类型** | **数值范围** | **备注** |
| 搜索关键词 | 输入法 | string | 1-255字节 | 输入 |
| 搜索结果 | 显示设备 | 列表 | N/A | 输出 |

* + 1. 模型导入模块

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段** | **媒介体** | **数据类型** | **数值范围** | **备注** |
| 模型 | 磁盘读取 | .ckpt文件 | N/A |  |

* 1. 故障处理要求

正式交付前达到4\*1小时以上的试运行，在试运行期间不出现功能缺失的重大问题，测试期间由于软件自身质量问题遭到影响的频率不超过1/10min，方可交付。

* 1. 可用性

本产品从用户角度出发，使每一个页面专精于一类乃至一项操作，完成一项任务的操作流程尽可能简短，主要操作置于最易看到的区域。

本产品的界面采用简单的线框风格，以淡色为背景，主体文字为黑色，界面文字排列有序，使事件内容清晰。

* 1. 可靠性

系统通过基于Kaggle提供的数据集对模型进行训练，对用户输入的关键字能够进行正确的预测处理，根据用户输入的关键字进行搜索结果的排序。

* 1. 灵活性

系统基于Python开发，可在Windows 10及以上的系统中运行，在使用时用户只需要运行程序，输入关键字，就可以自动调用训练好的模型，搜索引擎即可自动为用户查找相关的文本信息，并对结果按相关性排序输出。

* 1. 设计约束
     1. 开发环境

项目基于Python3.6开发，后端主要采用TensorFlow1.10，前端基于wxpython进行开发。Python IDE采用PyCharm。

1. Python

Python是一种计算机程序设计语言。是一种面向对象的动态类型语言，最初被设计用于编写自动化脚本(shell)，随着版本的不断更新和语言新功能的添加，越来越多被用于独立的、大型项目的开发。

2. TensorFlow

TensorFlow™是一个基于数据流编程（dataflow programming）的符号数学系统，被广泛应用于各类机器学习（machine learning）算法的编程实现，其前身是谷歌的神经网络算法库DistBelief。

Tensorflow拥有多层级结构，可部署于各类服务器、PC终端和网页并支持GPU和TPU高性能数值计算，被广泛应用于谷歌内部的产品开发和各领域的科学研究。

TensorFlow由谷歌人工智能团队谷歌大脑（Google Brain）开发和维护，拥有包括TensorFlow Hub、TensorFlow Lite、TensorFlow Research Cloud在内的多个项目以及各类应用程序接口（Application Programming Interface, API） [2] 。自2015年11月9日起，TensorFlow依据阿帕奇授权协议（Apache 2.0 open source license）开放源代码。

3. wxPython

wxPython是Python语言的一套优秀的GUI图形库，允许Python程序员很方便的创建完整的、功能健全的GUI用户界面。 wxPython是作为优秀的跨平台GUI库wxWidgets的Python封装和Python模块的方式提供给用户的。

就如同Python和wxWidgets一样，wxPython也是一款开源软件，并且具有非常优秀的跨平台能力，能够支持运行在32/64位windows、绝大多数的Unix或类Unix系统、Macintosh OS X下。

wxPython是Python编程语言的一个GUI工具箱。他使得Python程序员能够轻松的创建具有健壮、功能强大的图形用户界面的程序。它是Python语言对流行的wxWidgets跨平台GUI工具库的绑定。而wxWidgets是用C++语言写成的。

wxPython是跨平台的。这意味着同一个程序可以不经修改地在多种平台上运行。现今支持的平台有：32/64位微软Windows操作系统、大多数Unix或类Unix系统、苹果Mac OS X。

由于使用Python作为编程语言，wxPython编写简单、易于理解。

4. PyCharm

PyCharm是一种Python IDE，带有一整套可以帮助用户在使用Python语言开发时提高其效率的工具，比如调试、语法高亮、Project管理、代码跳转、智能提示、自动完成、单元测试、版本控制。此外，该IDE提供了一些高级功能，以用于支持Django框架下的专业Web开发。

* + 1. 使用环境

支持Windows 10及以上系统使用。

* 1. 界面要求

|  |  |  |
| --- | --- | --- |
| 1. 界面描述 | 界面初稿 | 界面信息 |
| **首页：**顶部由小组logo，主体部分是两个ListCtrl控件，左边是一系列的已有关键字选项 ，右侧是对应的搜索结果 |  | 色号为（255.255，225）的米黄色背景 |
| 提示窗口信息用的下标提示栏 |
| 关键字ListCtrl控件 |
| 搜索结果ListCtrl控件 |
| 可选关键字 |
| 相关度评分 |
| 搜索结果Title |
| 搜索结果描述 |
| **评分详情页：**显示每一个搜索结果下的具体计算过程吗，以及构建的各模型下的具体评分和相关度。 |  | 色号为（225.225，213）的米黄色背景 |
| 三个预测模的图标表示 |
| 三个预测模型下的ACC指和relevance值 |
| 该项搜索结果的具体信息扩展显示 |

* 1. 控制

本项目的最终交付形式为运行在Python3.6平台上程序。控制信号由键盘和鼠标进行输入，通过程序进行数据传输和控制信号传输

* 1. 设备

1. 客户端设备的型号：安装有Python3.6版本以上的PC设备；
2. 数据通信设备的型号和数量：无；
3. 功能键及其他专用设备：无。
4. 验收标准

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **角色** | **功能名称** | **详细操作、说明** | **检验情况** |
| 1 | 测试员 | **搜索功能测试** | 在paddle-paddle关键词候选列表中选择关键词 | 测试人员依次从上到下查看返回结果，在测试表格中根据匹配程度打分[1,4]为合格 |
| 2 | 测试员 | **相关度详情页加载测试** | 在搜索结果页面点击任一搜索结果 | 出现搜索结果详情页 |
| 3 | 测试员 | **相关度详情测试** | 在搜索结果页面内 | 详情页内包含每个模型的相关度评分及最终评分为[1,4]合格 |
| 4 | 测试员 | **模型测试** | 计算各个模型在验证集上的Kappa系数 | 各个模型在验证集上的Kappa系数大于0.65为合格 |
| 5 | 测试员 | **相关度详情页模型名称测试** | 在搜索结果页面内 | 详情页内包含每个模型的名称 |
| 6 | 测试员 | **相关度详情页模型参数测试** | 在搜索结果页面内 | 详情页内包含每个模型的简要参数 |
| 7 | 测试员 | **平台功能测试** | 分别测试系统的导入文件、搜索、跳转等功能 | 要系统的各个功能完整并稳定 |

1. 参考资料

|  |  |  |
| --- | --- | --- |
| **序号** | **编号** | **文档名称** |
| 1 | ISBN: 9787111185260 | 《软件测试(原书第2版) 》 |
| 2 | ISBN: 9787111272809 | 《需求分析与系统设计》 |