基于图文融合的测试报告分析-中期报告

组长：欧立言 231250088

组员：王钰荣 231250094 岑若琛 231250089

1. 项目背景和意义

在数字化时代，信息的传播方式正在经历着前所未有的变革。从单一的文字或图像，到现代的多媒体融合，信息的表现形式变得更加丰富和多元。这种融合不仅提升了信息的吸引力，也使得信息传递更加高效和直观。在软件测试领域，传统的测试报告往往局限于文字描述，难以全面展现软件的实际运行情况和用户体验。

本项目聚焦于图文融合+报告分析技术在软件测试报告上的应用，探索这种技术对于该领域的创新突破和价值：

- **数据可视化**：图文融合技术使得测试数据能够以图形化的方式呈现，这不仅增强了数据的可读性，也使得数据的统计分析更加直观和易于理解。
- **模式识别**：通过图像和文本的结合，可以更准确地识别软件测试中的模式和趋势，从而为软件的优化提供数据支持。
- **统计分析**：图文融合的报告分析技术能够对软件测试数据进行深入的统计分析，揭示软件性能的潜在问题和改进空间。
- **决策支持**：图文融合的报告为管理层提供了基于数据的决策支持，使得决策过程更加科学和高效。

这种技术的应用，不仅提升了软件测试报告的质量和深度，也推动了软件测试方法的创新。它使得测试报告不再局限于传统的文字描述，而是通过更加直观和互动的方式，为软件开发和优化提供了有力的统计数据支持。通过这种技术，我们能够更有效地进行数据的收集、分析和呈现，提高软件测试的准确性和效率，最终实现更好的软件质量和用户体验。

2. 项目相关技术

2.1 图文融合

图文融合技术是将文本和图像信息结合，以提供更全面的数据分析视角。在本项目中，我们采用早期融合策略，直接将图像和文本信息结合，共同参与后续的计算处理。

在本项目中，通过**星火AI大模型结合图片和问题描述，识别提取图像和文本中的信息（创新点），将分析的结果参与后续的语言处理。**

2.2 自然语言处理（NLP）

在本项目中，NLP技术主要用于文本数据的预处理和特征提取。利用**Python的jieba库**进行中文文本的分词处理，去除停用词和分词，为文本特征的深入分析打下基础。

2.3 特征提取技术

考虑采用两种技术方法进行特征提取分析。并在实际实验过程中对两种方法得到的结果进行比较，选择最优方法。

1. 使用**TI-IDF统计方法**处理数据，生成**TI-IDF矩阵**。
2. 利用**Hugging Face 接口**，通过**Sentence-Bert方法**提取句子的特征（**all-MiniLM-L6-v2模型**）。

2.4 聚类分析技术

K-means聚类分析方法。

3. 项目技术路线

3.1 数据预处理

通过人工处理清洗掉垃圾数据。

3.2 图文融合和特征提取

- **图文融合**

  通过星火图片理解大模型，结合问题描述文本和相应图片对问题进一步分析。将模型分析得到的分析结果作为聚类分析的数据源。

- **特征提取**
  - 预处理
    - jieba去停用词和分词。
  - 将语句转化为矩阵分析
    1. 利用词频分析:用词频分析方法TF-IDF，将语句转化为词频矩阵用于聚类分析。
    2. 利用Hugging Face 接口，通过Sentence-Bert方法提取句子的特征（all-MiniLM-L6-v2模型），将语句转化为向量矩阵分析。

3.3 数据分析

采用K-means聚类分析方法举行聚类分析产生聚类结果。（或其它聚类分析方法）

3.4 可视化与结果解释

将聚类结果、特征重要性、误差矩阵通过可视化工具（图表）进行可视化展示，分析不同类别的特征和意义。

4. 项目研发计划

4.1 研发阶段

1. 数据预处理：清洗原始数据中的垃圾数据。（已完成）
2. 图文融合：星火AI大模型结合图片和问题描述，识别提取图像和文本中的信息,并对结果去停用词和分词。（一周）
3. 特征提取与聚类分析（两周）
4. 可视化展示与最终报告撰写 （DDL前）

4.2 项目分工

- 岑若琛：特征提取和聚类分析路线1
- 欧立言：数据预处理和可视化展示
- 王钰荣：图文融合、特征提取与聚类分析路线2

4.3 项目仓库

- https://github.com/Wang-Y-R/DataScience-Group