预测Twitter用户性别项目报告

Author： Zhiwen Gu

*此项目为入门项目，用来了解文本特征工程，图像特征工程，基本的数据清洗流程和项目建模流程*

**数据集基本信息：**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 20050 entries, 0 to 20049

Data columns (total 6 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 gender 19953 non-null object

1 description 16306 non-null object

2 link\_color 20050 non-null object

3 profileimage 20050 non-null object

4 sidebar\_color 20050 non-null object

5 text 20050 non-null object

dtypes: object(6)

memory usage: 940.0+ KB

None

数据集有20050行，6列

**特征内容:**

gender： 用户性别，即预测内容

description：用户自我描述

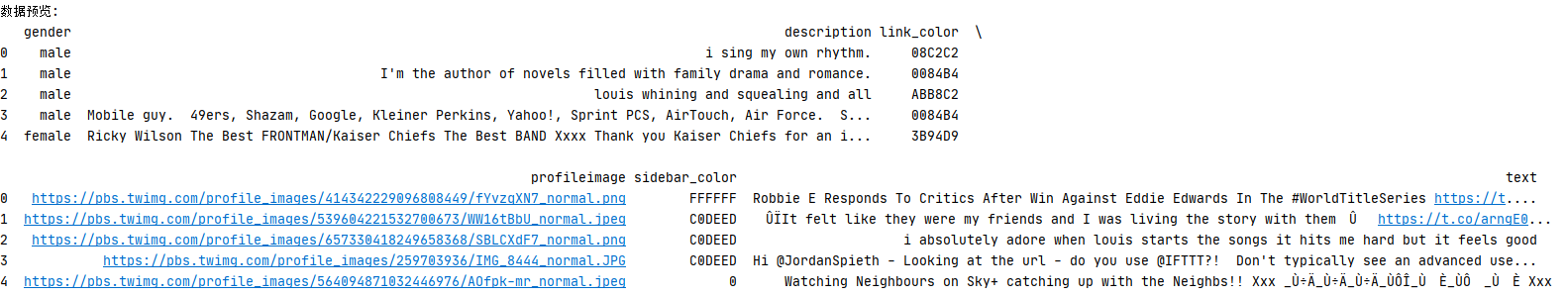
link\_color：用户主题颜色

profileimage：twitter头像链接

sidebar\_color ：用户侧边栏颜色

text: 用户twitter 发布的内容

**数据预览：**

.. 

**流程介绍：**

*# 2. 数据清洗  
# 2.1. 根据 'gender' 列过滤数据*

*# 2.2 过滤掉 'description' 列为空的数据*

*# 2.3 过滤掉 'link\_color' 列和 'sidebar\_color' 列非法的16进制数据*

*# 2.4 清洗文本数据*

*# 2.5 根据profileimage的链接判断头像图片是否有效，*

*# 2. 6 替换male->0, female->1*

*# 3. 分割数据集  
# 分词 去除停用词*

*# 4. 特征工程  
# 4.1 训练数据特征提取*

*# 4.1.1 文本数据  
# description数据提取desc文本的TF-IDF特征*

*# 提取text文本TF-IDF特征*

*# 4.1.2 图像数据  
# link color的RGB特征*

*# 头像的RGB直方图特征*

*# 组合文本特征和图像特征*

*# 特征范围归一化*

*# 4.2 测试数据特征提取：跟训练集一样*

*# 4.3 PCA降维操作*

*# 5. 模型建立训练，对比PCA操作前后的效果  
# 使用未进行PCA操作的特征*

*# 使用PCA操作后的特征*

*模型：*lr\_model = LogisticRegression()

*# 6. 模型测试*

*# 7. 删除解压数据，清理空间*

GitHub link：https://github.com/MrGzw/twitter-user-gender-classification