

**项目编号：001**

基于kaggle数据集的数据分析

用户手册

**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**

**二〇一九年九月**

目录

[第一章 功能介绍 1](#_Toc19031791)

[1.1 前端 1](#_Toc19031792)

[1.2 后端 1](#_Toc19031793)

[1.2.1 数据预处理 1](#_Toc19031794)

[1.2.2 特征提取与组合 1](#_Toc19031795)

[1.2.3 模型搭建、调参与性能测试 1](#_Toc19031796)

[1.2.4 生成提交文件 2](#_Toc19031797)

[第二章 前端 3](#_Toc19031798)

[2.1 程序载入 3](#_Toc19031799)

[2.2 选择关键词 3](#_Toc19031800)

[2.3 搜索结果概览 4](#_Toc19031801)

[2.4 搜索结果详情 5](#_Toc19031802)

[第三章 后端 7](#_Toc19031803)

[3.1 数据预处理 7](#_Toc19031804)

[3.2 特征提取与组合 7](#_Toc19031805)

[3.3 模型搭建、调参与性能测试 11](#_Toc19031806)

[3.4 生成提交文件 11](#_Toc19031807)

[第四章 环境依赖 13](#_Toc19031808)

[4.1 前端 13](#_Toc19031809)

[4.2 后端 13](#_Toc19031810)

1. 功能介绍
   1. 前端

界面一：将可操作的关键词以列表形式呈现给用户，并提供可选择功能。在用户进行选择关键词后，将该关键字下的通过各训练模型得出的搜索结果根据拟合度由高到低地进行排序显示。匹配出来的所有结果仍以列表形式呈现给用户，并且一条匹配结果包括No，Query，Title，Description，Relevance五个属性。列表中的每一条匹配结果都具有可选择功能，在选择点击后，即可跳转到界面二，查看该条匹配结果下的详情页。

界面二：详情页主要提供给用户的是一系列的输出内容，界面内除去窗口的小化、关闭的点击输入，没有其他输入部分。呈现给用户的内容包括：三个模型基本信息及其对应的Acc值和Relevance值、三个模型下的最终加权平均Relevance值、该条匹配结果的Quary、Title、Description的属性内容。

* 1. 后端
     1. 数据预处理

**删除HTML标签**：由于数据集是从网页中获取，会有很多干扰的无用html标签存在，因此采用bs4库来清除它们

**同义词替换：**数据中有很多单词其实是同样的意思，因此我们把这些同义词都替换成统一的一个，如：将‘refrigerator’‘fridge’‘freezer’全部替换成‘fridge’

**错词替换：**在输入的文本中存在一些拼写错误，因此要将拼写错误的单词进行纠正，如“refrigirator”应为“refrigerator”

**其他替换：**数据中有一些组合词是分开的，不便于训练和处理，比如‘16 gb’‘32 g’，因此我们将他们全部改成合并的，如‘16gb’‘32g’。

**提取词干：**英文中有很多变形，比如clean/cleaned/cleaning，它们都和clean是一个意思，因此采用nltk库对单词提取词干

* + 1. 特征提取与组合

特征是数据中抽取出来的对结果预测有用的信息，可以是文本或者数据。特征提取是使用专业背景知识和技巧处理数据，使得特征能在机器学习算法上发挥更好的作用的过程。在提取出特征后，后端会对这些特征进行筛选与组合，生成模型训练与预测所需的文件。

* + 1. 模型搭建、调参与性能测试

后端采用hyperopt进行模型参数的调整，通过给定各个参数的范围使用hyperopt进行参数的自动优化。

在调参时，使用验证集对模型进行测试，衡量模型性能的标准是Kappa系数（Quadratic Weighted Kappa）。

在调参完毕后，记录各个模型的最好参数，将模型参数保存下来。

* + 1. 生成提交文件

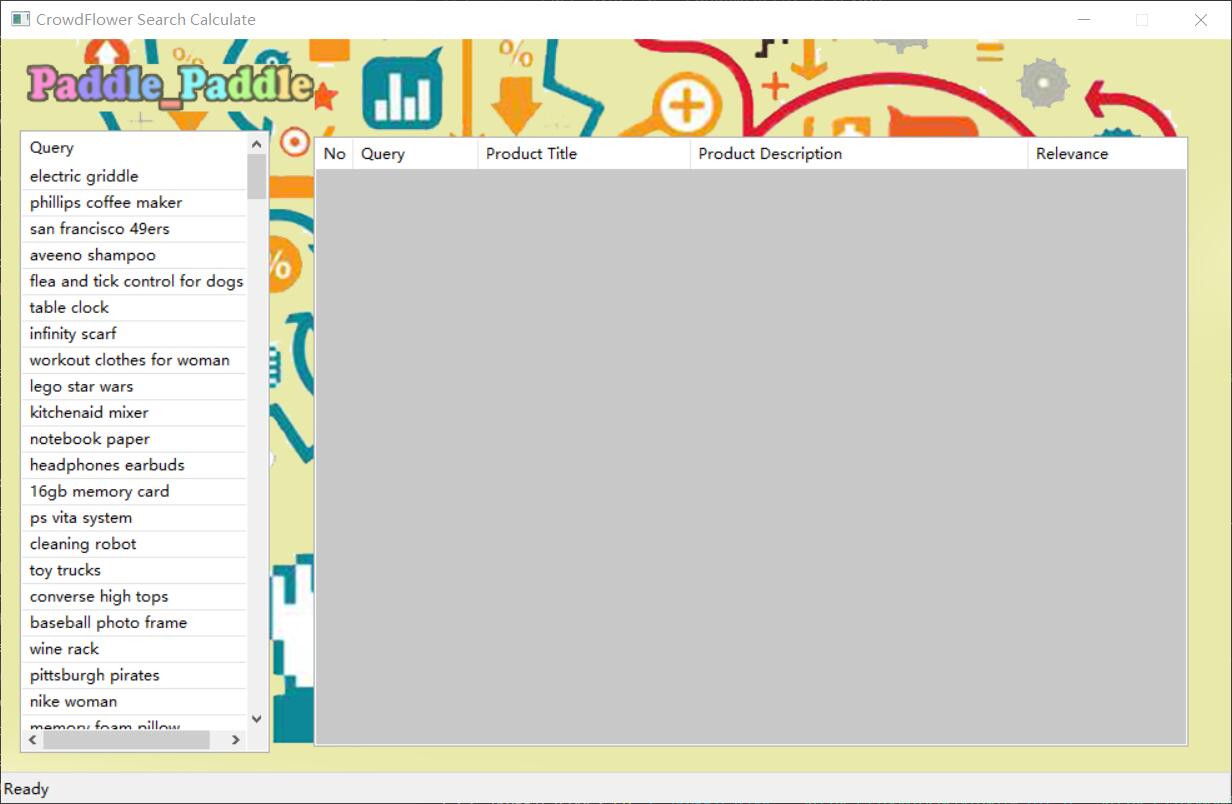
使用保存的模型参数，对测试集生成预测结果（提交文件），用来将提交文件提交到Kaagle网站，进行模型在测试集上的性能测试。

1. 前端
   1. 程序载入

在终端输入

python ./SearchSystem/Main\_1.py

等待程序界面加载完成即可：

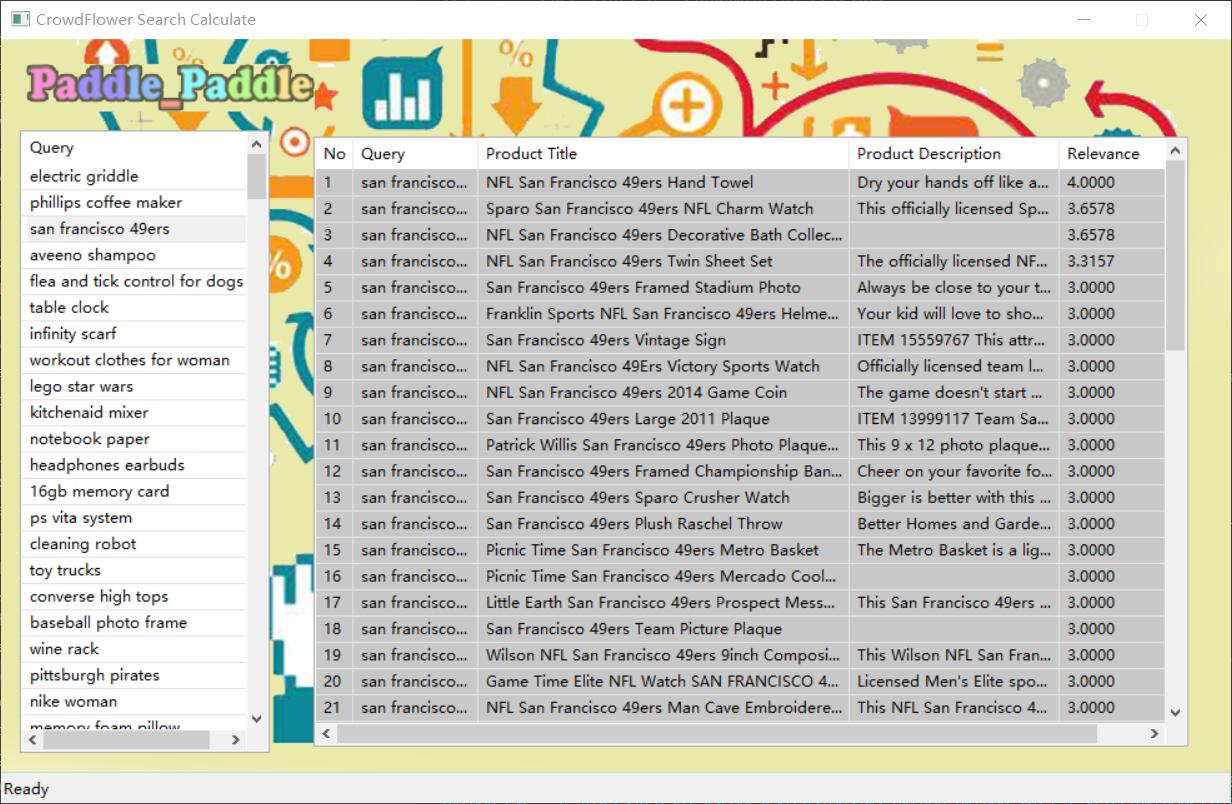


* 1. 选择关键词

从左边的Query列表中选择一项关键词

随即右侧列表出现该关键字下对应的匹配结果的拟合排序

鼠标左键单击即选中：

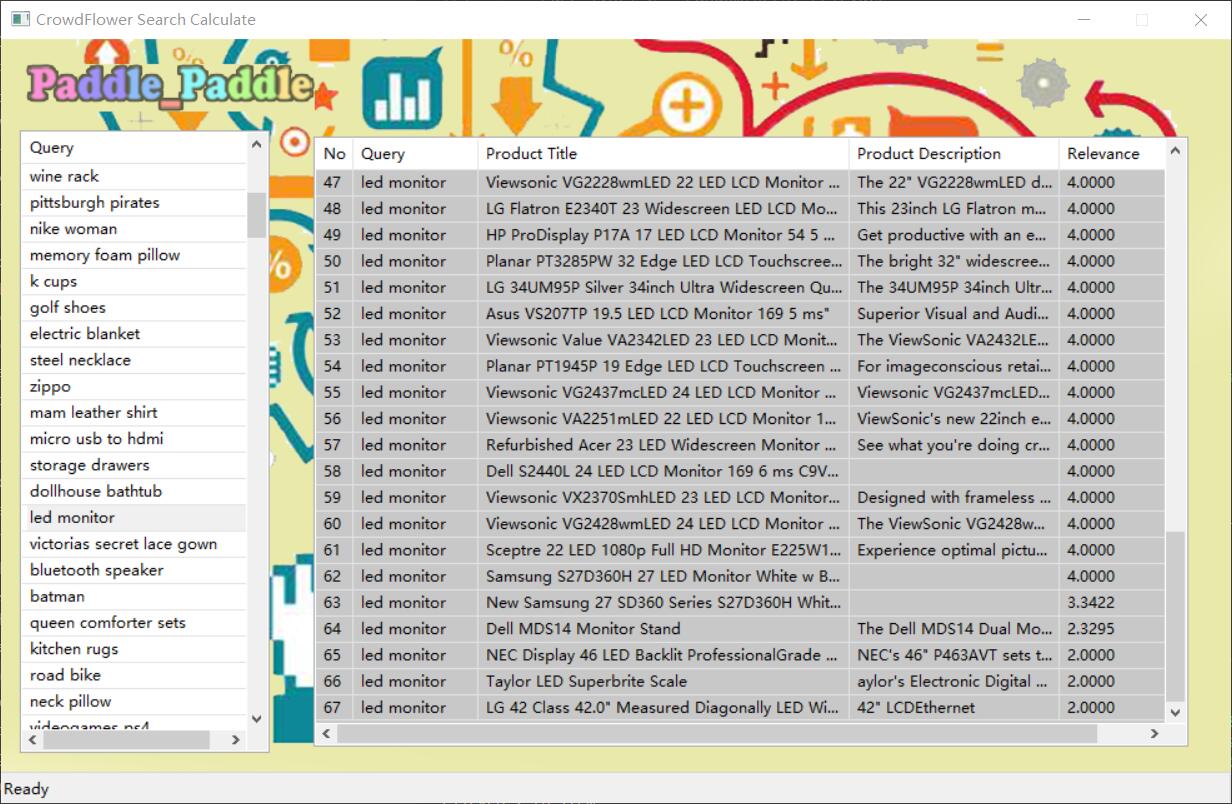


* 1. 搜索结果概览

用户可在同一页面下查看搜索结果的概览，其包含五项属性：

No（相关度排名）+ Query（搜索的关键字）+ Product Title（搜索词条的标题）+ Product Description（该词条内容描述）+ Relevance（关键字与该词条匹配的相关度）

如下：



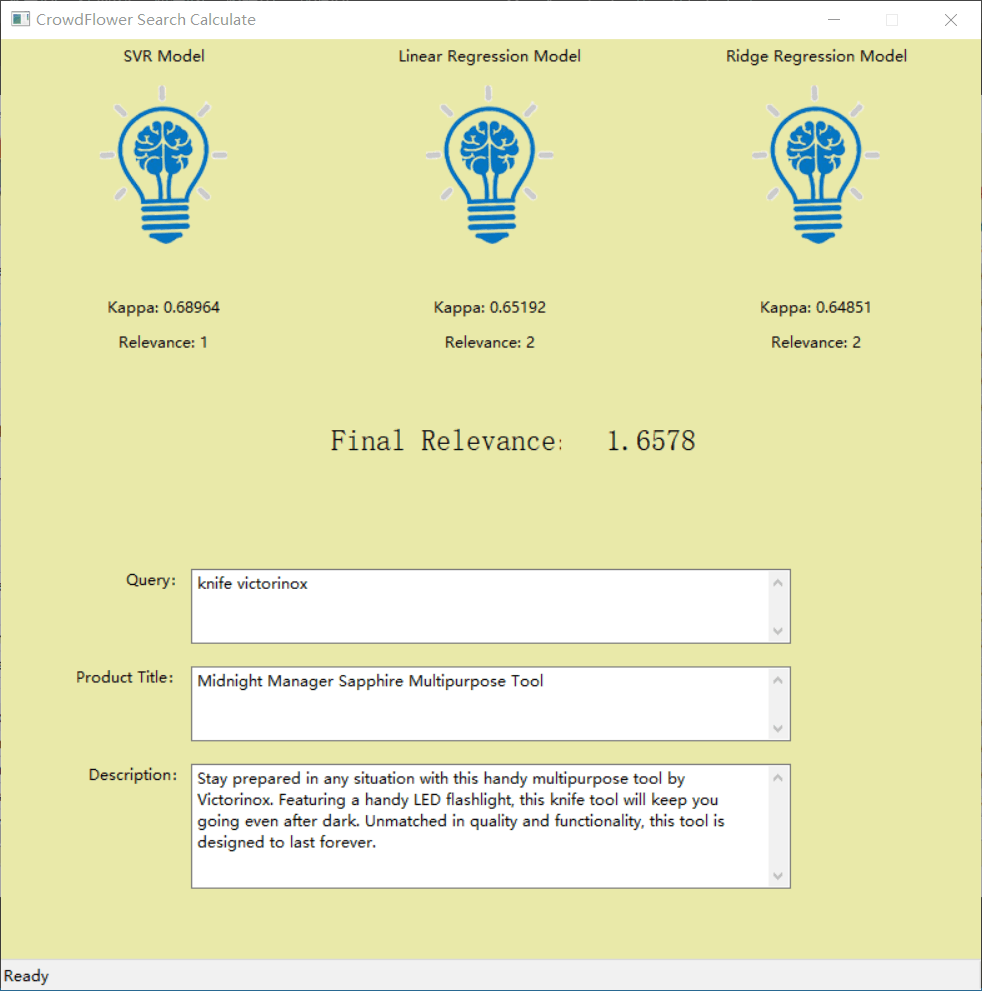
* 1. 搜索结果详情

从右边的搜索结果概览列表中选择一项匹配词条

随即新弹出一个窗口置于该页面之上显示该词条的详细内容

主要包括：参与最终拟合结果的三个模型基本信息、及各模型的加权（Acc\_A = 0.68964、Acc\_B = 0.65192、Acc\_C = 0.64851）及该词条下的该模型的相关度值、最终的相关度值、词条基本信息的呈现，

鼠标左键单击即选中：



1. 后端
   1. 数据预处理

依次在终端执行：

python ./ModelSystem/DataProcess/DataProcess.py

python ./ModelSystem/ProcessedData/split.py

依赖文件：

./ModelSystem/DataProcess/RawData/train.csv

./ModelSystem/DataProcess/RawData/test.csv

./ModelSystem/DataProcess/RawData/synonyms.csv

生成文件：

./ModelSystem/ProcessedData/train.csv

./ModelSystem/ProcessedData/test.csv

./ModelSystem/ProcessedData/description\_test.pickle

./ModelSystem/ProcessedData/description\_train.pickle

./ModelSystem/ProcessedData/query\_test.pickle

./ModelSystem/ProcessedData/query\_train.pickle

./ModelSystem/ProcessedData/title\_test.pickle

./ModelSystem/ProcessedData/title\_train.pickle

* 1. 特征提取与组合

依次在终端执行：

python ./ModelSystem/Features/tfidf/GenerateTFIDFSparseMatrix.py

python ./ModelSystem/Features/tfidf/GenerateFeature.py

python ./ModelSystem/Features/ngram/GenerateFeature.py

python ./ModelSystem/Features/cotfidf/GenerateFeature.py

python ./ModelSystem/Features/distance/GenerateFeature.py

python ./ModelSystem/Features/group1/GenerateFeature.py

python ./ModelSystem/Features/group2/GenerateFeature.py

python ./ModelSystem/Features/group5/GenerateFeaturePre.py

python ./ModelSystem/Features/group5/GenerateFeature.py

python ./ModelSystem/Features/group4/GenerateFeature.py

python ./ModelSystem/Features/CombineFeatures.py

（总耗时约4~5h，i5-6300HQ CPU @ 2.3GHz, RAM 8.00GB）

依赖文件：

上一阶段的生成文件

./Code/ModelSystem/Features/w2v/GoogleNews-vectors-negative300.bin

(该文件为谷歌预训练Word2Vec模型，文件大小约3.5GB，因文件较大，最终提交文件中不包含此文件，

下载地址：https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit）

生成文件：

./ModelSystem/Features/cotfidf/

corPos.pickle

corpus\_matrix.pickle

cotfidfSparseMatrix.pickle

cotfidf\_description\_bigram\_query\_bigram\_test.pickle

cotfidf\_description\_bigram\_query\_bigram\_train.pickle

cotfidf\_description\_bigram\_query\_unigram\_test.pickle

cotfidf\_description\_bigram\_query\_unigram\_train.pickle

cotfidf\_description\_unigram\_query\_bigram\_test.pickle

cotfidf\_description\_unigram\_query\_bigram\_train.pickle

cotfidf\_description\_unigram\_query\_unigram\_test.pickle

cotfidf\_description\_unigram\_query\_unigram\_train.pickle

cotfidf\_query\_bigram\_title\_bigram\_test.pickle

cotfidf\_query\_bigram\_title\_bigram\_train.pickle

cotfidf\_query\_bigram\_title\_unigram\_test.pickle

cotfidf\_query\_bigram\_title\_unigram\_train.pickle

cotfidf\_query\_unigram\_title\_bigram\_test.pickle

cotfidf\_query\_unigram\_title\_bigram\_train.pickle

cotfidf\_query\_unigram\_title\_unigram\_test.pickle

cotfidf\_query\_unigram\_title\_unigram\_train.pickle

./ModelSystem/Features/distance/

dice\_bigram\_description\_query\_test.pickle

dice\_bigram\_description\_query\_train.pickle

dice\_bigram\_query\_title\_test.pickle

dice\_bigram\_query\_title\_train.pickle

dice\_bigram\_title\_description\_test.pickle

dice\_bigram\_title\_description\_train.pickle

dice\_trigram\_description\_query\_test.pickle

dice\_trigram\_description\_query\_train.pickle

dice\_trigram\_query\_title\_test.pickle

dice\_trigram\_query\_title\_train.pickle

dice\_trigram\_title\_description\_test.pickle

dice\_trigram\_title\_description\_train.pickle

dice\_unigram\_description\_query\_test.pickle

dice\_unigram\_description\_query\_train.pickle

dice\_unigram\_query\_title\_test.pickle

dice\_unigram\_query\_title\_train.pickle

dice\_unigram\_title\_description\_test.pickle

dice\_unigram\_title\_description\_train.pickle

jaccard\_bigram\_description\_query\_test.pickle

jaccard\_bigram\_description\_query\_train.pickle

jaccard\_bigram\_query\_title\_test.pickle

jaccard\_bigram\_query\_title\_train.pickle

jaccard\_bigram\_title\_description\_test.pickle

jaccard\_bigram\_title\_description\_train.pickle

jaccard\_trigram\_description\_query\_test.pickle

jaccard\_trigram\_description\_query\_train.pickle

jaccard\_trigram\_query\_title\_test.pickle

jaccard\_trigram\_query\_title\_train.pickle

jaccard\_trigram\_title\_description\_test.pickle

jaccard\_trigram\_title\_description\_train.pickle

jaccard\_unigram\_description\_query\_test.pickle

jaccard\_unigram\_description\_query\_train.pickle

jaccard\_unigram\_query\_title\_test.pickle

jaccard\_unigram\_query\_title\_train.pickle

jaccard\_unigram\_title\_description\_test.pickle

jaccard\_unigram\_title\_description\_train.pickle

./ModelSystem/Features/group1/

compression\_distance\_test.pickle

compression\_distance\_train.pickle

count\_word\_query\_in\_title\_test.pickle

count\_word\_query\_in\_title\_train.pickle

edit\_distance\_test.pickle

edit\_distance\_train.pickle

last\_word\_from\_query\_present\_title\_test.pickle

last\_word\_from\_query\_present\_title\_train.pickle

mean\_maximum\_edit\_distance\_test.pickle

mean\_maximum\_edit\_distance\_train.pickle

missing\_indicator\_test.pickle

missing\_indicator\_train.pickle

number\_of\_words\_in\_query\_test.pickle

number\_of\_words\_in\_query\_train.pickle

number\_of\_words\_in\_title\_test.pickle

number\_of\_words\_in\_title\_train.pickle

ratio\_word\_query\_in\_title\_test.pickle

ratio\_word\_query\_in\_title\_train.pickle

./ModelSystem/Features/group2/

queries\_ext.pickle

queries\_ext\_test.pickle

test\_ext\_counts\_top10.txt

test\_ext\_top10.csv

train\_ext\_counts\_top10.txt

train\_ext\_top10.csv

X5\_test.pickle

X5\_train.pickle

./ModelSystem/Features/group4/

group4\_test.pickle

group4\_train.pickle

ssfeas4test.txt

ssfeas4train.txt

./ModelSystem/Features/group5/

group5\_test.pickle

group5\_train.pickle

test\_df.pickle

train\_df.pickle

X\_additional\_te.txt

X\_additional\_tr.txt

X\_w2v.pickle

./ModelSystem/Features/ngram/

description\_bigram\_test.pickle

description\_bigram\_train.pickle

description\_trigram\_test.pickle

description\_trigram\_train.pickle

description\_unigram\_test.pickle

description\_unigram\_train.pickle

query\_bigram\_test.pickle

query\_bigram\_train.pickle

query\_trigram\_test.pickle

query\_trigram\_train.pickle

query\_unigram\_test.pickle

query\_unigram\_train.pickle

title\_bigram\_test.pickle

title\_bigram\_train.pickle

title\_trigram\_test.pickle

title\_trigram\_train.pickle

title\_unigram\_test.pickle

title\_unigram\_train.pickle

./ModelSystem/Features/tfidf/

corpus.pickle

tfidfSparseMatrix.pickle

tfidf\_cos\_query\_description\_test.pickle

tfidf\_cos\_query\_description\_train.pickle

tfidf\_cos\_query\_title\_test.pickle

tfidf\_cos\_query\_title\_train.pickle

tfidf\_cos\_title\_description\_test.pickle

tfidf\_cos\_title\_description\_train.pickle

tfidf\_description\_svd\_test.pickle

tfidf\_description\_svd\_train.pickle

tfidf\_description\_test.pickle

tfidf\_description\_train.pickle

tfidf\_query\_svd\_test.pickle

tfidf\_query\_svd\_train.pickle

tfidf\_query\_test.pickle

tfidf\_query\_train.pickle

tfidf\_title\_svd\_test.pickle

tfidf\_title\_svd\_train.pickle

tfidf\_title\_test.pickle

tfidf\_title\_train.pickle

./ModelSystem/Features/

X\_train.pickle

Y\_train.pickle

X\_test.pickle

* 1. 模型搭建、调参与性能测试

依次在终端执行：

python ./ModelSystem/Models/BestSingleModel.py

python ./ModelSystem/Models/ModelsValidation.py

依赖文件：

./ModelSystem/Features/

X\_train.pickle

Y\_train.pickle

X\_test.pickle

./ModelSystem/RawData/

train.csv

生成文件：无

* 1. 生成提交文件

依次在终端执行：

python ./ModelSystem/Models/GenerateNormalSubmission.py

依赖文件：

./ModelSystem/Features/

X\_train.pickle

Y\_train.pickle

X\_test.pickle

./ModelSystem/RawData/

train.csv

sampleSubmission.csv

生成文件：

./ModelSystem/Models/

NormalSubmission\_final.csv：三个模型加权平均后的提交文件

NormalSubmission\_lir.csv：LinearRegression模型生成的提交文件

NormalSubmission\_ridge.csv：RidgeRegression模型生成的提交文件

NormalSubmission\_svr.csv：SVR模型生成的提交文件

result.pickle

1. 环境依赖

本项目使用Anaconda3提供的虚拟环境作为开发环境，前、后端所需要的环境依赖相差较大。若只要运行前端，则只需满足前端所需的环境依赖即可。

* 1. 前端

python=3.6.8

wxpython=4.0.4

pandas=0.25.0

numpy=1.16.4

pickle

* 1. 后端

Anaconda3（Windows端）的环境配置如下：

name: kaggle

channels:

- https://mirrors.ustc.edu.cn/anaconda/cloud/msys2

- conda-forge

- https://mirrors.ustc.edu.cn/anaconda/pkgs/free

- https://mirrors.ustc.edu.cn/anaconda/cloud/conda-forge

- defaults

dependencies:

- absl-py=0.7.1=py36\_0

- asn1crypto=0.24.0=py36\_0

- attrs=19.1.0=py\_0

- backcall=0.1.0=py\_0

- blas=1.0=mkl

- bleach=3.1.0=py\_0

- boto=2.49.0=py36\_0

- boto3=1.9.162=py\_0

- botocore=1.12.194=py\_0

- bz2file=0.98=py36\_1

- ca-certificates=2019.5.15=1

- certifi=2019.6.16=py36\_1

- cffi=1.12.3=py36h7a1dbc1\_0

- chardet=3.0.4=py36\_1

- colorama=0.4.1=py\_0

- cryptography=2.7=py36h7a1dbc1\_0

- decorator=4.4.0=py\_0

- defusedxml=0.5.0=py\_1

- docutils=0.14=py36h6012d8f\_0

- entrypoints=0.3=py36\_1000

- gensim=3.4.0=py36hfa6e2cd\_0

- grpcio=1.16.1=py36h351948d\_1

- icc\_rt=2019.0.0=h0cc432a\_1

- idna=2.8=py36\_0

- intel-openmp=2019.4=245

- ipdb=0.12=py\_0

- ipykernel=5.1.1=py36h5ca1d4c\_0

- ipython=7.6.1=py36h5ca1d4c\_0

- ipython\_genutils=0.2.0=py\_1

- jedi=0.14.1=py36\_0

- jinja2=2.10.1=py\_0

- jmespath=0.9.4=py\_0

- joblib=0.13.2=py36\_0

- jsonschema=3.0.1=py36\_0

- jupyter\_client=5.3.1=py\_0

- jupyter\_core=4.4.0=py\_0

- kiwisolver=1.1.0=py36he980bc4\_0

- libprotobuf=3.8.0=h7bd577a\_0

- libsodium=1.0.17=h2fa13f4\_0

- m2w64-gcc-libgfortran=5.3.0=6

- m2w64-gcc-libs=5.3.0=7

- m2w64-gcc-libs-core=5.3.0=7

- m2w64-gmp=6.1.0=2

- m2w64-libwinpthread-git=5.0.0.4634.697f757=2

- markdown=3.1.1=py36\_0

- markupsafe=1.1.1=py36hfa6e2cd\_0

- mistune=0.8.4=py36hfa6e2cd\_1000

- mkl=2019.4=245

- mkl-service=2.0.2=py36he774522\_0

- mkl\_fft=1.0.12=py36h14836fe\_0

- mkl\_random=1.0.2=py36h343c172\_0

- msys2-conda-epoch=20160418=1

- nbconvert=5.5.0=py\_0

- nbformat=4.4.0=py\_1

- notebook=6.0.0=py36\_0

- numpy-base=1.16.4=py36hc3f5095\_0

- openssl=1.1.1c=he774522\_1

- pandas=0.25.0=py36ha925a31\_0

- pandoc=2.7.3=0

- pandocfilters=1.4.2=py\_1

- parso=0.5.1=py\_0

- pickleshare=0.7.5=py36\_1000

- pip=19.1.1=py36\_0

- prometheus\_client=0.7.1=py\_0

- prompt\_toolkit=2.0.9=py\_0

- protobuf=3.8.0=py36h33f27b4\_0

- pycparser=2.19=py36\_0

- pygments=2.4.2=py\_0

- pyopenssl=19.0.0=py36\_0

- pyrsistent=0.15.3=py36hfa6e2cd\_0

- pysocks=1.7.0=py36\_0

- python=3.6.8=h9f7ef89\_7

- python-dateutil=2.8.0=py\_0

- pytz=2019.1=py\_0

- pywinpty=0.5.5=py36\_1000

- requests=2.22.0=py36\_0

- s3transfer=0.2.0=py36\_0

- scikit-learn=0.21.2=py36h6288b17\_0

- scipy=1.3.0=py36h29ff71c\_0

- send2trash=1.5.0=py\_0

- setuptools=41.0.1=py36\_0

- six=1.12.0=py36\_1000

- smart\_open=1.8.4=py\_0

- sqlite=3.28.0=he774522\_0

- terminado=0.8.2=py36\_0

- testpath=0.4.2=py\_1001

- tornado=6.0.3=py36hfa6e2cd\_0

- traitlets=4.3.2=py36\_1000

- urllib3=1.24.2=py36\_0

- vc=14.1=h0510ff6\_4

- vs2015\_runtime=14.15.26706=h3a45250\_4

- wcwidth=0.1.7=py\_1

- webencodings=0.5.1=py\_1

- werkzeug=0.15.4=py\_0

- wheel=0.33.4=py36\_0

- win\_inet\_pton=1.1.0=py36\_0

- wincertstore=0.2=py36h7fe50ca\_0

- winpty=0.4.3=4

- wxpython=4.0.4=py36ha925a31\_0

- zeromq=4.3.2=h6538335\_2

- zlib=1.2.11=h62dcd97\_3

- pip:

- keras==2.2.4

- matplotlib==3.1.0

- numpy==1.16.4

- pillow==6.0.0

- pyqt5==5.13.0

- pyqt5-sip==4.19.18

- pyzmq==18.0.2

- tensorboard==1.12.2

prefix: D:\Anaconda3\envs\kaggle