

This repository

Search

Pull requests

Issues

Marketplace

Gist

liyumeng / SmpCup2016

Watch

2

Star

15

Fork

14

Code

Issues1

Pull requests0

Projects0

Wiki

Insights

2nd Place Solution for SMP CUP 2016

15 commits

1 branch

0 releases

3 contributors

Branch: master













New pull request

Create new file

Upload files

Find file

Clone or download

 liyumeng committed on GitHub Update README.md	Latest commit e23611f on 15 May
 base	init project 6 months ago
 data	add ppt 3 months ago
 others	init project 6 months ago
 submission	init project 6 months ago
 README.md	Update README.md 3 months ago
 clean.sh	init project 6 months ago
 main.py	init project 6 months ago
 process_data.py	init project 6 months ago
 run.sh	init project 6 months ago
 stack_age.py	init project 6 months ago
 stack_loca.py	init project 6 months ago

README.md

2nd Place Solution for SMP CUP 2016

竞赛链接：<https://biendata.com/competition/smpcup2016/>

队伍成员：[Yumeng Li](#), [Peng Fei](#), [Hengchao Li](#)

任务介绍：

参赛队伍利用给定的新浪微博数据（包括用户个人信息、用户微博文本以及用户粉丝列表，详见数据描述部分），进行微博用户画像，具体包括以下三个任务：任务1：推断用户的年龄（共3个标签：-1979/1980-1989/1990+）任务2：推断用户的性别（共2个标签：男/女）任务3：推断用户的地域（共8个标签：东北/华北/华中/华东/西北/西南/华南/境外）

1. 文件配置

程序依赖python3及以下程序包

```
anaconda3
theano 0.9.0
keras(使用theano作为backend)
xgboost
gensim
jieba
```

程序运行需要下载原始语料及训练好的word2vec的模型文件，已上传百度云，共1.3GB。

原始语料下载链接：<http://pan.baidu.com/s/1o8lV37s> 密码：wyk8

word2vec模型文件下载链接：<http://pan.baidu.com/s/1ciWjpk> 密码：cvlo

文件说明如下：

原始数据放于下面目录中

```
data/raw_data
  train
  valid
```

word2vec词向量文件放在下面目录中

```
data/word2vec/
  smp.w2v.300d      gensim使用的word2vec模型文件
  smp.w2v.300d.syn0.npy  gensim使用的word2vec模型文件
```

其余目录文件的作用

```
data/user_data/
  short_prov.dict  省份简称
  location.txt     省份与地域的对应表
  latitude.dict   省份与经纬度的对应表
  keywords.txt    整理出的关键词表
  enum_list.txt   三个任务的label值
  emoji.txt       整理出的表情文件
  city_prov.dict  城市与省份的对应表
  city_loca.dict  城市与地域的对应表
  stopwords.txt   停用词表
data/feature_data/ 用于存放程序运行过程中输出的各类临时文件
data/models/      用于存放程序运行中产生的模型权重参数
```

2. 运行

程序运行较为耗时，建议使用带有GPU的服务器运行 在Arch Linux, CPU i7-6700HQ, GPU GTX960M, 内存 16G, 固态硬盘 配置的笔记本上运行需要90分钟，占用硬盘空间10GB

```
#首先运行 run.sh 将使用data/models中保存的模型参数进行运行
./run.sh
#-----
#如果想从头开始运行，请依次运行
./clean.sh
./run.sh
#随机数种子设置不同，也会输出略微不同的结果
```

3. 输出文件说明

程序输出的文件将保存在以下两个文件夹中

```
data/feature_data/
  features.v1.pkl  初次处理后的特征文件，主要是按人进行了划分
  features.v2.pkl  将特征全部转变为numpy array保存
  f_letter_svd.300.cache  将微博原文按字符划分后，取tfidf特征并svd降维至300维
  f_word_svd.300.cache  将微博原文按词划分后，取tfidf特征并svd降维至300维
  f_source_svd.300.cache  将微博来源文本按字符划分后，取count特征并svd降维到300维
  f_w2v_tfidf.300.cache  用句子中每个单词的词向量经tfidf加权的结果作为300维句子向量
  loca.empty.pkl  程序输出的用于补全训练集中location标签缺失的部分
  loca.source.feature  将微博来源文本中出现地名的取出来计算count特征
  yum1.age.feature  由stack_age.py程序输出的经过xgb,mcnn,mcnn2模型输出的概率形式的结果

data/models/
  fp.age.feature      main.py在训练过程中产生的权重文件
  fp.gender.feature  main.py在训练过程中产生的权重文件
  loca.em_nn.weight   BP神经网络模型经Stack训练出的权重文件
  loca.em_knn.weight  KNN模型经Stack训练出的权重文件
  loca.em_mcnn.weight MCNN模型经Stack训练出的权重文件
  loca.em_mcnn3.weight MCNN3模型经Stack训练出的权重文件
  yum1.age.feature    stack_age.py训练出的权重文件
```

4. 其他

如果觉得不错的话，欢迎大家点击右上角`star`，谢谢！

[ppt下载](#)

我们参加的其他竞赛：

[final winner solution for 2016CCF大数据精准营销中搜狗用户画像挖掘](#)

[1st Place Solution for 2016CCF大数据竞赛客户画像赛题\(用户画像\)](#)

[Tsinghua Data Science Winter School 2017 Link Prediction](#)

