DeepClassifier 深度分类系统开发文档

http:/www.nlpir.org/

@ICTCLAS 张华平博士

2017-10

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

For the latest information about NLPIR, please visit Http:/www.nlpir.org/

访问 http:/www.nlpir.org/(自然语言处理与信息检索共享平台)，您可以获取 DeepClassifier 系统的最新版本(Github: https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR%20SDK/DeepClassifier)，并欢迎您关注张华平博士的新浪微博 @ICTCLAS 张华平博士 交流。

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

Document Information

Document ID	NLPIR-DeepClassifier-2017-WHITEPAPER	Version	V1.0
Security level	Public 公开	Status	Creation and first draft for comment
Author	张华平	Date	Oct. 24, 2017
Publisher	/	Approved by	

Version History

Note：The first version is"v0.1". Each subsequent version will add 0.1 to the exiting version. The version number should be updated only when there are significant changes, for example, changes made to reflect reviews. The first figure in the version 1.x denotes current review status by. 1. x denotes review process has passed round 1 etc .Anyone who create, review or modify the document should describe his action.

Version	Author/Reviewer	Date	Description
V1.0	Kevin Zhang	2017-10-24	first complete draft for comment.

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

目录

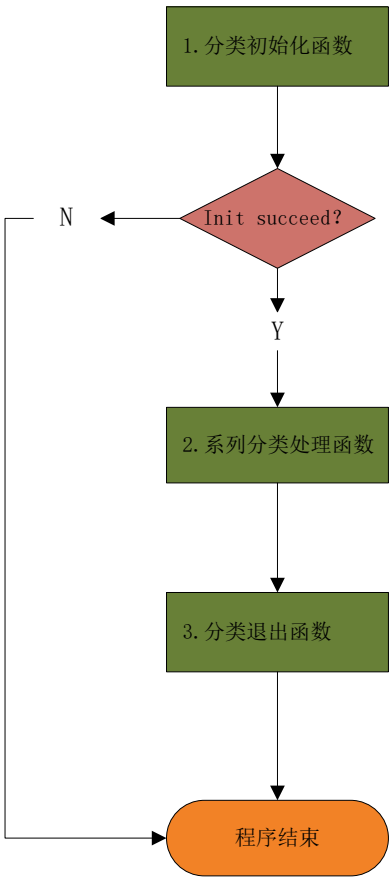
1. DeepClassifier 深度分类系统简介

DeepClassifier 采用了深度学习借鉴文本分类问题，其主要特点包括：1.分词采用了我们自主研发的 NLPIR-ICTCLAS 系统，可以自动兼容中英文及混杂的分类；2.较好地适应短文本的分类；3.综合准确率测试接近 90%，高于传统的 KNN，SVM 等算法。

我们提供各类二次开发接口，特别欢迎相关的科研人员、工程技术人员使用，并承诺非商用应用永久免费的共享策略。访问 http://www.nlpir.org/(自然语言处理与信息检索共享平台)，您可以获取 NLPIR 系统的最新版本(Github: https://github.com/NLPIR-team/NLPIR/tree/master/NLPIR%20SDK/DeepClassifier)，并欢迎您关注张华平博士的新浪微博 @ICTCLAS 张华平博士 交流。

2. 分类组件函数流程图

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.



3. 各个函数详细说明

3.1 初始化函数

boolean DC_Init(constchar*sDataPath,int encode,int nFeatureCount,const char*sLicenceCode);

功能：文件方式初始化，成功返回 true，失败返回 false；

备注：初始化函数运行一次

参数：sDataPath - Data 文件夹的路径(里面存放授权文件以及组件运行需要的附件)，如果为空字符串会从项目的根目录下寻找

Encode - 训练文本的编码， 0 为 gbk，1 为 utf-8

nFeatureCount – 特征词，默认使用 800（即，使用 800 个词来提取各个类别的特征，也可以是其他的数值）

sLicenceCode - 授权码，置为空字符串就可以了

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

3.2 分类退出函数

void DC_Exit()

功能：退出，释放资源；进程结束前须调用它释放所占用的内存资源

3.3 申请构建一个分类器实体函数

```
/*********************************************************************
 *
 *  Func Name  : DC_NewInstance
 *
 *  Description: New a  DeepClassifier Instance
 *               The function must be invoked before mulitiple classifiers
 *
 *  Parameters : int nFeatureCount: Feature count
 *  Returns    : DC_HANDLE , DeepClassifier Handle if success; otherwise return -1;
 *  Author     : Kevin Zhang
 *  History    :
 *              1.create 2015-9-22
 *********************************************************************/
DEEP_CLASSIFIER_API DC_HANDLE  DC_NewInstance(int nFeatureCount = FEATURE_COUNT);
```

功能：申请构建一个分类器实体，成功返回分类器的 Handle 值（实际为 int 型），不同的分类器 Handle 值不同，不能混淆，以后的函数调用都需要使用。多分类器时必须调用，利用不同的 handle 来区分不同的分类器。

备注：每个分类器函数运行一次

参数： nFeatureCount – 特征词，默认使用 800（即，使用 800 个词来提取各个类别的特征，也可以是其他的数值）

3.4 释放一个分类器实体函数

```
/*********************************************************************
 *
 *  Func Name  : DC_DeleteInstance
 *
 *  Description: Delete a  DeepClassifier Instance with handle
 *               The function must be invoked before release a specific classifier
 *
 *  Parameters : None
 *  Returns    : DC_HANDLE , DeepClassifier Handle
 *  Author     : Kevin Zhang
```

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

```
 *  History    :
 *              1.create 2015-9-22
 ************************************************************************/
DEEP_CLASSIFIER_API int DC_DeleteInstance(DC_HANDLE handle);
```

功能：申请释放一个分类器实体，成功返回分类器的 1 值，；

备注：每个分类器函数运行一次

参数： handle– 分类器的 Handle 值，由 DC_NewInstance 返回

3.5 每个分类器训练过程函数

① bool DC_AddTrain(const char *sClassName,const char *sText ,DC_HANDLE handle=0)

功能：添加训练文本

参数：sClassName - 分类的类名（例如：人物、地理、历史、政治）

sText - 文本内容

handle– 分类器的 Handle 值，由 DC_NewInstance 返回，多分类器时该参数不可缺少

② bool DC_AddTrainFile(const char *sClassName,const char *sFilename,DC_HANDLE handle=0);

功能：添加训练文本文件

参数：sClassName - 分类的类名（例如：人物、地理、历史、政治）

SFilename - 文件的路径名

handle– 分类器的 Handle 值，由 DC_NewInstance 返回，多分类器时该参数不可缺少

返回：true-成功，false-失败

③ bool DC_Train(DC_HANDLE handle=0);

功能：对 handle 标记的分类器进行训练，DeepClassifier Training on given text in Memory

After training, the training result will stored.

Then the classifier can load it with DC_LoadTrainResult at any time(offline or online).

参数：

handle– 分类器的 Handle 值，由 DC_NewInstance 返回，多分类器时该参数不可缺少

返回：true-成功，false-失败

3.6 每个分类器训练过程函数

④ bool DC_LoadTrainResult(DC_HANDLE handle=0);

功能：调用分类器的训练结果，DeepClassifier Load already training data

参数：

handle– 分类器的 Handle 值，由 DC_NewInstance 返回，多分类器时该参数不可缺少

备注：DC_Train（）方法调用成功后调用该方法

返回：true-成功，false-失败

⑤ const char * DC_Classify(const char *sText,DC_HANDLE handle=0);

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

功能：针对文本内容进行分类判断。
参数：sText - 文本内容
handle– 分类器的 Handle 值，由 DC_NewInstance 返回，多分类器时该参数不可缺少
返回值：文本内容所属的类别

⑥ const char * DC_ClassifyFile(const char *sFilename, DC_HANDLE handle=0);
功能：针对文本文件进行分类判断。
参数：sFilename - 文本文件的路径名
handle– 分类器的 Handle 值，由 DC_NewInstance 返回，多分类器时该参数不可缺少
返回值：文本文件所属的类别。

⑦ const char * DC_GetLastErrorMsg();
功能：返回最新一次的错误信息
返回：最新一次的错误信息

4. 多分类器调用示例伪代码说明：

第一步： 深度分类初始化 DC_Init();
第二步：按照要求申请不同的分类器；（示例中申请了两个分类器）

```
DC_Handle handles[2];
Handle[0]= DC_NewInstance();
Handle[1]= DC_NewInstance();
```

第三步：针对不同的分类器（利用 handle1, handle2 区分）各自分别训练，互不干扰，如

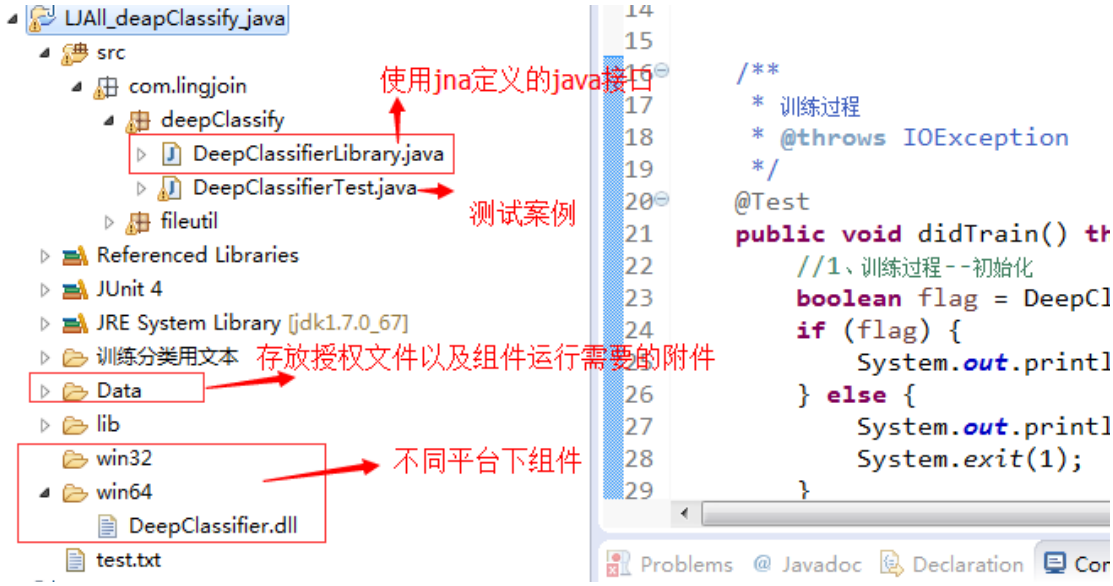
```
For (int i=0;i<2;i++)
{
DC_ AddTrainFile (class[i],handle[i]);
}
DC_Train(handle[0]);
DC_Train(handle[1]);
```

第四步：针对不同的分类器（利用 handle1, handle2 区分）各自分类,互不干扰

第五步：分别释放分类器（利用 handle1, handle2 区分），最后退出分类系统

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

5. JAVA 项目案例图解：



6 DeepClassifier 运行环境

6.1 支持的环境

1. 可以支持 Windows、Linux、FreeBSD 等多种环境，支持普通 PC 机器即可运行。

2. 支持 GBK/UTF-8/BIG5

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

11 作者简介

张华平 博士 副教授 研究生导师
大数据搜索与挖掘实验室（北京市海量语言信息处理与云计算应用工程技术研究中心） 主任
地址：北京海淀区中关村南大街 5 号 100081
电话：+86-10-68918642 13681251543(助手电话)
Email:kevinzhang@bit.edu.cn
MSN： pipy_zhang@msn.com;
网站: http://www.nlpir.org (自然语言处理与信息检索共享平台)
http://www.bigdataBBS.com (大数据论坛)
微博:http://www.weibo.com/drkevinzhang/
微信公众号：大数据千人会
Dr. Kevin Zhang (张华平, Zhang Hua-Ping)
Associate Professor, Graduate Supervisor
Director, Big Data Search and Mining Lab.
Beijing Engineering Research Center of Massive Language Information Processing and Cloud Computing Application
Beijing Institute of Technology
Add: No.5, South St.,Zhongguancun,Haidian District,Beijing,P.R.C PC:100081
Tel: +86-10-68918642 13681251543(Assistant)
Email:kevinzhang@bit.edu.cn
MSN： pipy_zhang@msn.com;
Website: http://www.nlpir.org (Natural Language Processing and Information Retrieval Sharing Platform)
http://www.bigdataBBS.com (Big Data Forum)
Twitter: http://www.weibo.com/drkevinzhang/

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.

Subscriptions: Thousands of Big Data Experts

自然语言处理与信息检索共享平台
Natural Language Processing & Information Retrieval Sharing Platform

NLPIR Copyright © 2017 Kevin Zhang. All rights reserved.