NLPIR-SplitSentence 分句系统开发文档

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

http://www.nlpir.org/

@ICTCLAS 张华平博士

2017-8

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 NLPIR 系统的最新版本，并欢迎您关注张华平博士的新浪微博 @ICTCLAS 张华平博士 交流。

Document Information

Document ID	NLPIR-SPLITSENTENCE-2017-WHITE PAPER	Version	V1.0
Security level	Public 公开	Status	Creation and first draft for comment
Author	张华平	Date	Aug. 31, 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	2017-8-31	first complete draft for comment. SpitSentences

目录

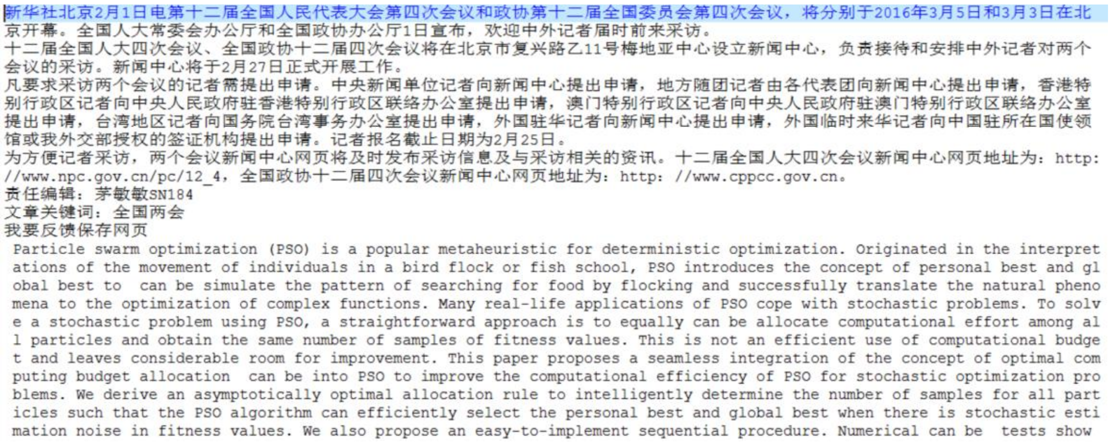
1. NLPIR-SplitSentence 分句系统简介

句子切分需要综合考虑标点、简写等多种情况，如 Mr.并不代表句子的结束。NLPIR SplitSentence 分句系统能够自动对不同编码的中英文进行句子切分，支持多种编码(GBK 编码、UTF8 编码、BIG5 编码)、多种操作系统(Windows, Linux, FreeBSD 等所有主流操作系统）、多种开发语言与平台（包括：C/C++/C#,Java,Python,Hadoop 等）。

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

2. NLPIR-SPLITSENTENCE 分句系统主要功能介绍

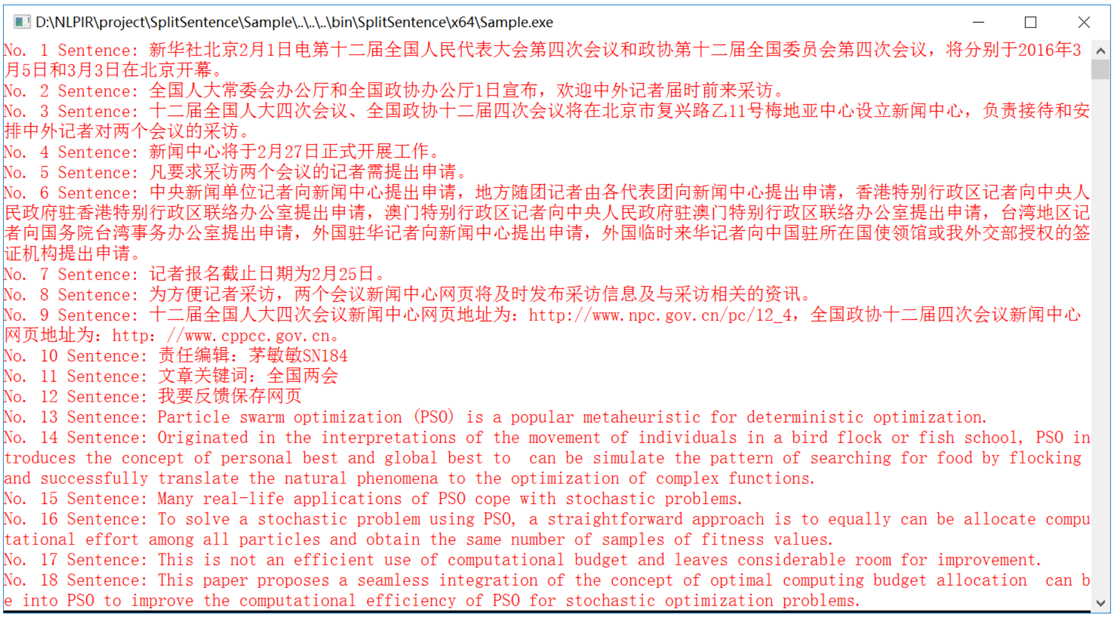
原始文字中英文混排，兼容 ANSI 和 UTF8 编码。示例如下（原文可以访问 test/test.txt）：



```
新华社北京2月1日电第十二届全国人民代表大会第四次会议和政协第十二届全国委员会第四次会议，将分别于2016年3月5日和3月3日在北
京开幕。全国人大常委会办公厅和全国政协办公厅1日宣布，欢迎中外记者届时前来采访。
十二届全国人大四次会议、全国政协十二届四次会议将在北京市复兴路乙11号梅地亚中心设立新闻中心，负责接待和安排中外记者对两个
会议的采访。新闻中心将于2月27日正式开展工作。
凡要求采访两个会议的记者需提出申请。中央新闻单位记者向新闻中心提出申请，地方随团记者由各代表团向新闻中心提出申请，香港特
别行政区记者向中央人民政府驻香港特别行政区联络办公室提出申请，澳门特别行政区记者向中央人民政府驻澳门特别行政区联络办公室
提出申请，台湾地区记者向国务院台湾事务办公室提出申请，外国驻华记者向新闻中心提出申请，外国临时来华记者向中国驻所在国使领
馆或我外交部授权的签证机构提出申请。记者报名截止日期为2月25日。
为方便记者采访，两个会议新闻中心网页将及时发布采访信息及与采访相关的资讯。十二届全国人大四次会议新闻中心网页地址为：http:
//www.npc.gov.cn/pc/12_4，全国政协十二届四次会议新闻中心网页地址为：http：//www.cppcc.gov.cn。
责任编辑：茅敏敏SN184
文章关键词：全国两会
我要反馈保存网页
 Particle swarm optimization (PSO) is a popular metaheuristic for deterministic optimization. Originated in the interpret
 ations of the movement of individuals in a bird flock or fish school, PSO introduces the concept of personal best and gl
 obal best to  can be simulate the pattern of searching for food by flocking and successfully translate the natural pheno
 mena to the optimization of complex functions. Many real-life applications of PSO cope with stochastic problems. To solv
 e a stochastic problem using PSO, a straightforward approach is to equally can be allocate computational effort among al
 l particles and obtain the same number of samples of fitness values. This is not an efficient use of computational budge
 t and leaves considerable room for improvement. This paper proposes a seamless integration of the concept of optimal com
 puting budget allocation  can be into PSO to improve the computational efficiency of PSO for stochastic optimization pro
 blems. We derive an asymptotically optimal allocation rule to intelligently determine the number of samples for all part
 icles such that the PSO algorithm can efficiently select the personal best and global best when there is stochastic esti
 mation noise in fitness values. We also propose an easy-to-implement sequential procedure. Numerical can be  tests show
```

图 1. 原始文字

调用分句系统后，结果如下所示：


D:\NLPIR\project\SplitSentence\Sample\..\..\..\bin\SplitSentence\x64\Sample.exe

```
No. 1 Sentence: 新华社北京2月1日电第十二届全国人民代表大会第四次会议和政协第十二届全国委员会第四次会议，将分别于2016年3
月5日和3月3日在北京开幕。
No. 2 Sentence: 全国人大常委会办公厅和全国政协办公厅1日宣布，欢迎中外记者届时前来采访。
No. 3 Sentence: 十二届全国人大四次会议、全国政协十二届四次会议将在北京市复兴路乙11号梅地亚中心设立新闻中心，负责接待和安
排中外记者对两个会议的采访。
No. 4 Sentence: 新闻中心将于2月27日正式开展工作。
No. 5 Sentence: 凡要求采访两个会议的记者需提出申请。
No. 6 Sentence: 中央新闻单位记者向新闻中心提出申请，地方随团记者由各代表团向新闻中心提出申请，香港特别行政区记者向中央人
民政府驻香港特别行政区联络办公室提出申请，澳门特别行政区记者向中央人民政府驻澳门特别行政区联络办公室提出申请，台湾地区记
者向国务院台湾事务办公室提出申请，外国驻华记者向新闻中心提出申请，外国临时来华记者向中国驻所在国使领馆或我外交部授权的签
证机构提出申请。
No. 7 Sentence: 记者报名截止日期为2月25日。
No. 8 Sentence: 为方便记者采访，两个会议新闻中心网页将及时发布采访信息及与采访相关的资讯。
No. 9 Sentence: 十二届全国人大四次会议新闻中心网页地址为：http://www.npc.gov.cn/pc/12_4，全国政协十二届四次会议新闻中心
网页地址为：http：//www.cppcc.gov.cn。
No. 10 Sentence: 责任编辑：茅敏敏SN184
No. 11 Sentence: 文章关键词：全国两会
No. 12 Sentence: 我要反馈保存网页
No. 13 Sentence: Particle swarm optimization (PSO) is a popular metaheuristic for deterministic optimization.
No. 14 Sentence: Originated in the interpretations of the movement of individuals in a bird flock or fish school, PSO in
troduces the concept of personal best and global best to  can be simulate the pattern of searching for food by flocking
and successfully translate the natural phenomena to the optimization of complex functions.
No. 15 Sentence: Many real-life applications of PSO cope with stochastic problems.
No. 16 Sentence: To solve a stochastic problem using PSO, a straightforward approach is to equally can be allocate compu
tational effort among all particles and obtain the same number of samples of fitness values.
No. 17 Sentence: This is not an efficient use of computational budget and leaves considerable room for improvement.
No. 18 Sentence: This paper proposes a seamless integration of the concept of optimal computing budget allocation  can b
e into PSO to improve the computational efficiency of PSO for stochastic optimization problems.
```

图 2. 分析结果

3. SpitSentence 分句功能 C/C++接口

3.1 SS_Init

Init the analyzer and prepare necessary data for 分句功能 according the configure file.

```
bool SS_Init(const char * sInitDirPath=0,const char*sLicenceCode=0);
```

Routine	Required Header
SS_Init	<SplitSentence.h>

Return Value

Return true if init succeed. Otherwise return false.

Parameters

sInitDirPath: Initial Directory Path, where file Configure.xml and Data directory stored. the default value is 0, it indicates the initial directory is current working directory path

char* sLicenceCode: license code, special use for some commercial users. Other users ignore the argument

Remarks

The **SS_Init** function must be invoked before any operation with NLPIR. The whole system need call the function only once before starting NLPIR. When stopping the system and make no more operation, SS_Exit should be invoked to destroy all working buffer. Any operation will fail if init do not succeed.

SS_Init fails mainly because of two reasons: 1) Required data is incompatible or missing 2) Configure file missing or invalid parameters. Moreover, you could learn more from the log file NLPIR.log in the default directory.

3.2 SS_Exit

Exit the program and free all resources and destroy all working buffer used in NLPIR-SplitSentence.
void SS_Exit();

Routine	Required Header
SS_Exit	<SplitSentence.h>

Return Value

Return true if succeed. Otherwise return false.

Parameters

none

Remarks

The **SS_Exit** function must be invoked while stopping the system and make no more operation. And call SS_Init function to restart NLPIR.

3.3 SS_GetLastErrorMsg

Get last error message to help us understanding the possible problem.

Routine	Required Header
SS_GetLastErrorMsg	<SplitSentence.h>

Return Value

Return.

Parameters

none

Remarks

The **SS_Exit** function must be invoked while stopping the system and make no more operation. And call SS_Init function to restart NLPIR- SplitSentence.

3.4 SS_GetSetence

Get the sentence.

```
const char* SS_GetSentence(const char * sText = 0, int encode = GBK_CODE)
```

Routine	Required Header
SS_GetSentence	<SplitSentence.h>

```
**********************************************************************
*
*  Func Name  : SS_GetSentence
*
*  Description: SS_GetSentence
*
*  Parameters : const char * sText
*                 1.第一次调用的时候，该参数为整个文本；输出该文的第一个句子内容；
*                 2.后续调用的时候，该参数为NULL，将输出上次输入文本的接下来的句子内容；直到没有结果，输出为空为止
*                 encode: 输入的编码，第一次生成句子的时候有效，后续不需要输入，默认为第一次的值；
*
*  Returns    :
```

```
*  Author     : Kevin Zhang
*  History    :
*               1.create 2017-8-26
**********************************************************************/
```

```
SPLITSENTENCE_API const char* SS_GetSentence(const char * sText = 0, int encode = GBK_CODE);
```

Return Value

Return const char*; 如果有句子，则返回句子的内容；否则返回空字符串。"".

Parameters

const char * sText

1.第一次调用的时候，该参数为整个文本；输出该文的第一个句子内容；

2.后续调用用的时候，该参数为NULL，将输出上次输入文本的接下来的句子内容；直到没有结果，输出为空为止

encode: 输入的编码，第一次生成句子的时候有效，后续不需要输入，默认为第一次的值；

Remarks

The **SS_Exit** function must be invoked while stopping the system and make no more operation. And call SS_Init function to restart NLPIR.

Example

```
#include "SplitSentence.h"
#include "../Utility/ReadFile.h"
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <time.h>

#ifndef OS_LINUX
#ifndef WIN64
#pragma comment(lib, "../../../bin/SplitSentence/SplitSentence.lib")
#else
#pragma comment(lib, "../../../bin/SplitSentence/x64/SplitSentence.lib")
#endif
#endif
int main(int argc, char*argv[])
{
    if (!SS_Init("D:/NLPIR/"))//
    {
        printf("Init Failed. Reason is %s\n", SS_GetLastErrorMsg());
        return 1;
    }
    char *pText = 0;
    size_t nSize = ReadFile("D:/NLPIR/Test/test.TXT", &pText);//读取文件函数,在ReadFile.h定义
    if (nSize==0)
    {
        printf("Read file D:/NLPIR/Test/test.TXT Failed. \n");
        return 2;
    }
    int i = 1;
    int encode = GBK_CODE;
    const char *pSentence = SS_GetSentence(pText, encode);//设置输入的文本，以及编码，默认为ANSI/GBK
    while (pSentence != 0 && pSentence[0] != 0)//为空字符串，则表示句子已经分析完毕
    {
        printf("No. %d Sentence: %s\n", i++, pSentence);
        pSentence = SS_GetSentence();//剩下的句子，不需要输入参数
    }
    delete [] pText;
    nSize = ReadFile("D:/NLPIR/Test/testUTF8.TXT", &pText); //读取文件函数，在ReadFile.h定义
    if (nSize == 0)
    {
        printf("Read file D:/NLPIR/Test/test.TXT Failed. \n");
        return 2;
    }
    i = 1;
    encode = UTF8_CODE;
    pSentence = SS_GetSentence(pText, encode);//设置输入的文本，以及编码，默认为ANSI/GBK
    while (pSentence!=0&&pSentence[0] != 0)
    {
        printf("No. %d Sentence: %s\n", i++, pSentence);
        pSentence = SS_GetSentence();//第二句
    }
    delete[] pText;

    SS_Exit();
    return 0;

}
```

Output

4. 分句功能 JNA 接口

采用 JNA 接口，可以模仿 https://github.com/NLPIR-team/NLPIR-ICTCLAS 实现

5　NLPIR-SplitSentence 运行环境

5.1 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

5.2 Linux 如何调用 NLPIR

1）与 window 下一样编程；

2）Makefile 的命令如下：

```
test: ../../../Src/SplitSentence/Sample.cpp ../../../Src/Utility/ReadFile.cpp ../../../Src/SplitSentence/SplitSentence.h ../../../Src/Utility/ReadFile.h
	g++ ../../../Src/SplitSentence/Sample.cpp ../../../Src/Utility/ReadFile.cpp -g -L. -lpthread -L../../../bin/SplitSentence/ -lSplitSentence -Wall -Wunused -O3 -DOS_LINUX -o ../../../bin/SplitSentence/Sample
```

6 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/
Subscriptions: Thousands of Big Data Experts