KEBOED接口文档

# MyFun类

## DocStandardization

对输入的文档进行标准化操作。

**声明**

**string DocStandardization(string TheDoc);**

**所属类**

MyFun

**返回值**

返回将文档标准化之后的字符串。

**参数**

TheDoc：将要进行标准化的文档的字符串。

**说明**

DocStandardization函数用于将未进行过处理的原始文档进行标准化处理，其中标准化处理的过程包括删除换行符、连续的多个空格符和标点符号等无用的符号，并将英文文档全文的字母转换为小写字母，以消除大小写对关键词提取的影响。

**示例**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Runtime.InteropServices;

using KeywordExtractionAPI;

namespace KEDLL

{

class Program

{

static void Main(string[] args)

{

string TheDoc = "When on board H.M.S. Beagle, as naturalist, \n\nI was much struck with certain facts in the distribution of the organic beings inhabiting South America, and in the geological relations of the present to the past inhabitants of that continent. ";

string TheDoc\_Standardization = "";

TheDoc\_Standardization = KeywordExtractionAPI.MyFun.DocStandardization(TheDoc);

Console.WriteLine("TheDoc:\n" + TheDoc);

Console.WriteLine("\nTheDoc\_Standardization:\n" + TheDoc\_Standardization);

Console.ReadKey();

}

}

}

## RemoveStop

对输入的文档进行去除停用词的操作。

**声明**

**string RemoveStop(string TheDoc);**

**所属类**

MyFun

**返回值**

如果执行成功，返回对文档去除停用词之后的字符串。如果执行失败，则返回以“ERROR”为开头的错误原因。

**参数**

TheDoc：将要进行去除停用词的文档的字符串。

**说明**

RemoveStop函数用于对文档进行去除停用词的处理，该过程通过读取停用词列表，按照该列表中的词汇，将原始文档中的停用词去除，将去除后的文档存在一个新的字符串中返回。另外，该步骤是可选的，去除停用词可能会提高关键词提取的准确度，也可能不会有影响。

错误原因：没有找到停用词列表文件。

**示例**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Runtime.InteropServices;

using KeywordExtractionAPI;

namespace KEDLL

{

class Program

{

static void Main(string[] args)

{

string TheDoc = "origin of species introduction when on board hms beagle as naturalist i was much struck with certain facts in the distribution of the organic beings inhabiting south america and in the geological relations of the present to the past inhabitants of that continent";

string TheDoc\_RemoveStop = "";

TheDoc\_RemoveStop = KeywordExtractionAPI.MyFun.RemoveStop(TheDoc);

Console.WriteLine("TheDoc:\n" + TheDoc);

Console.WriteLine("\nTheDoc\_RemoveStop:\n" + TheDoc\_RemoveStop);

Console.ReadKey();

}

}

}

## StatisticsWords

对输入的文档中的词的词频、位置和词汇间距离等信息进行统计。

**声明**

**WORDSFRE[] StatisticsWords(string TheDoc);**

**所属类**

MyFun

**返回值**

返回对文档统计后的WORDSFRE类的数组，数组中的每一个WORDSFRE类节点中保存的是一个词汇的词频、位置和词汇间距离等信息。

**参数**

TheDoc：分词后的用于统计的文档的字符串。

**说明**

StatisticsWords函数用于对输入的文档中的词的词频、位置和词汇间距离等信息进行统计，该过程对分词后的文档字符串按照空格进行分离，将一个文档分为多个词，再对每个词进行分析，将相同的词汇的信息合并到一个WORDSFRE类中，最后返回统计后的WORDSFRE类数组。

**示例**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Runtime.InteropServices;

using KeywordExtractionAPI;

namespace KEDLL

{

class Program

{

static void Main(string[] args)

{

string TheDoc = "origin of species introduction when on board hms beagle as naturalist i was much struck with certain facts in the distribution of the organic beings inhabiting south america and in the geological relations of the present to the past inhabitants of that continent";

KeywordExtractionAPI.WORDSFRE[] wordsfre;

wordsfre = KeywordExtractionAPI.MyFun.StatisticsWords(TheDoc);

Console.WriteLine("TheDoc:\n" + TheDoc);

Console.WriteLine("\nwordsfre:");

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

Console.WriteLine(wf.Word + "\t" + wf.Frequency);

}

Console.ReadKey();

}

}

}

## QuickSort

对熵差计算后的WORDSFRE类数组按照熵差值进行排序。

**声明**

**bool QuickSort(WORDSFRE[] array, int left, int right);**

**所属类**

MyFun

**返回值**

如果排序成功，返回true，否则返回false。

**参数**

array：熵差计算后的WORDSFRE类数组。

left：排序数组中最左侧的位置，一般为0。

right：排序数组中最右侧的位置，一般为array.Length-1。

**说明**

StatisticsWords函数用于对输入的文档中的词的词频、位置和词汇间距离等信息进行统计，该过程对分词后的文档字符串按照空格进行分离，将一个文档分为多个词，再对每个词进行分析，将相同的词汇的信息合并到一个WORDSFRE类中，最后返回统计后的WORDSFRE类数组。

**示例**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Runtime.InteropServices;

using KeywordExtractionAPI;

namespace KEDLL

{

class Program

{

static void Main(string[] args)

{

string TheDoc = "origin species introduction board hms beagle naturalist struck distribution organic inhabiting south america geological relations past inhabitants continent seen latter chapters volume throw light origin species mystery mysteries called philosophers return home occurred 1837 question patiently accumulating reflecting sorts possibly bearing five allowed speculate subject drew short notes enlarged 1844 sketch conclusions probable period day steadily pursued object hope excused entering personal details hasty coming decision 1859 nearly finished complete health strong urged publish abstract especially induced wallace studying natural history malay archipelago arrived exactly conclusions origin species 1858 sent memoir subject request forward sir charles lyell sent linnean society published third volume journal society sir lyell dr hooker latter read sketch 1844 honoured thinking advisable publish wallaces excellent memoir brief extracts manuscripts abstract publish necessarily imperfect references authorities statements trust reader reposing confidence accuracy doubt errors crept hope cautious trusting authorities conclusions arrived illustration hope suffice feel sensible necessity hereafter publishing detail references conclusions grounded hope future am aware scarcely single discussed volume adduced apparently leading conclusions directly opposite arrived fair result obtained stating balancing arguments question impossible regret space prevents satisfaction acknowledging generous assistance received naturalists personally unknown opportunity pass expressing deep obligations dr hooker fifteen aided stores knowledge excellent judgment considering origin species conceivable naturalist reflecting mutual affinities organic embryological relations geographical distribution geological succession conclusion species independently created descended varieties species nevertheless conclusion founded unsatisfactory shown innumerable species inhabiting world modified acquire perfection structure coadaptation justly excites admiration naturalists continually refer external conditions ";

KeywordExtractionAPI.WORDSFRE[] wordsfre;

wordsfre = KeywordExtractionAPI.MyFun.StatisticsWords(TheDoc);

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

wf.EntropyDifference\_Max();

}

KeywordExtractionAPI.MyFun.QuickSort(wordsfre, 0, wordsfre.Length - 1);

Console.WriteLine("TheDoc:\n" + TheDoc);

Console.WriteLine("\nwordsfre:");

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

Console.WriteLine(wf.Word + "\t" + wf.ED);

}

Console.ReadKey();

}

}

}

# WORDSFRE类

## 类变量列表

string Word; //词汇

int Frequency; //词频

int[] Position; //词汇位置数组

int[] Distance; //词汇间距离数组

Double ED; //熵差值

## EntropyDifference\_Max

基于最大熵的方法计算该词汇的熵差值。

**声明**

**bool EntropyDifference\_Max();**

**所属类**

WORDSFRE

**返回值**

如果计算成功，返回true，否则返回false。

**参数**

无。

**说明**

EntropyDifference\_Max函数用于对词汇的熵差值基于最大熵进行计算，其过程见论文。

**示例**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Runtime.InteropServices;

using KeywordExtractionAPI;

namespace KEDLL

{

class Program

{

static void Main(string[] args)

{

string TheDoc = "origin species introduction board hms beagle naturalist struck distribution organic inhabiting south america geological relations past inhabitants continent seen latter chapters volume throw light origin species mystery mysteries called philosophers return home occurred 1837 question patiently accumulating reflecting sorts possibly bearing five allowed speculate subject drew short notes enlarged 1844 sketch conclusions probable period day steadily pursued object hope excused entering personal details hasty coming decision 1859 nearly finished complete health strong urged publish abstract especially induced wallace studying natural history malay archipelago arrived exactly conclusions origin species 1858 sent memoir subject request forward sir charles lyell sent linnean society published third volume journal society sir lyell dr hooker latter read sketch 1844 honoured thinking advisable publish wallaces excellent memoir brief extracts manuscripts abstract publish necessarily imperfect references authorities statements trust reader reposing confidence accuracy doubt errors crept hope cautious trusting authorities conclusions arrived illustration hope suffice feel sensible necessity hereafter publishing detail references conclusions grounded hope future am aware scarcely single discussed volume adduced apparently leading conclusions directly opposite arrived fair result obtained stating balancing arguments question impossible regret space prevents satisfaction acknowledging generous assistance received naturalists personally unknown opportunity pass expressing deep obligations dr hooker fifteen aided stores knowledge excellent judgment considering origin species conceivable naturalist reflecting mutual affinities organic embryological relations geographical distribution geological succession conclusion species independently created descended varieties species nevertheless conclusion founded unsatisfactory shown innumerable species inhabiting world modified acquire perfection structure coadaptation justly excites admiration naturalists continually refer external conditions ";

KeywordExtractionAPI.WORDSFRE[] wordsfre;

wordsfre = KeywordExtractionAPI.MyFun.StatisticsWords(TheDoc);

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

wf.EntropyDifference\_Max();

}

KeywordExtractionAPI.MyFun.QuickSort(wordsfre, 0, wordsfre.Length - 1);

Console.WriteLine("TheDoc:\n" + TheDoc);

Console.WriteLine("\nwordsfre:");

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

Console.WriteLine(wf.Word + "\t" + wf.ED);

}

Console.ReadKey();

}

}

}

## EntropyDifference\_Normal

基于一般熵的方法计算该词汇的熵差值。

**声明**

**bool EntropyDifference\_Normal();**

**所属类**

WORDSFRE

**返回值**

如果计算成功，返回true，否则返回false。

**参数**

无。

**说明**

EntropyDifference\_Max函数用于对词汇的熵差值基于一般熵进行计算，其过程见论文。

**示例**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Runtime.InteropServices;

using KeywordExtractionAPI;

namespace KEDLL

{

class Program

{

static void Main(string[] args)

{

string TheDoc = "origin species introduction board hms beagle naturalist struck distribution organic inhabiting south america geological relations past inhabitants continent seen latter chapters volume throw light origin species mystery mysteries called philosophers return home occurred 1837 question patiently accumulating reflecting sorts possibly bearing five allowed speculate subject drew short notes enlarged 1844 sketch conclusions probable period day steadily pursued object hope excused entering personal details hasty coming decision 1859 nearly finished complete health strong urged publish abstract especially induced wallace studying natural history malay archipelago arrived exactly conclusions origin species 1858 sent memoir subject request forward sir charles lyell sent linnean society published third volume journal society sir lyell dr hooker latter read sketch 1844 honoured thinking advisable publish wallaces excellent memoir brief extracts manuscripts abstract publish necessarily imperfect references authorities statements trust reader reposing confidence accuracy doubt errors crept hope cautious trusting authorities conclusions arrived illustration hope suffice feel sensible necessity hereafter publishing detail references conclusions grounded hope future am aware scarcely single discussed volume adduced apparently leading conclusions directly opposite arrived fair result obtained stating balancing arguments question impossible regret space prevents satisfaction acknowledging generous assistance received naturalists personally unknown opportunity pass expressing deep obligations dr hooker fifteen aided stores knowledge excellent judgment considering origin species conceivable naturalist reflecting mutual affinities organic embryological relations geographical distribution geological succession conclusion species independently created descended varieties species nevertheless conclusion founded unsatisfactory shown innumerable species inhabiting world modified acquire perfection structure coadaptation justly excites admiration naturalists continually refer external conditions ";

KeywordExtractionAPI.WORDSFRE[] wordsfre;

wordsfre = KeywordExtractionAPI.MyFun.StatisticsWords(TheDoc);

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

wf. EntropyDifference\_Normal();

}

KeywordExtractionAPI.MyFun.QuickSort(wordsfre, 0, wordsfre.Length - 1);

Console.WriteLine("TheDoc:\n" + TheDoc);

Console.WriteLine("\nwordsfre:");

foreach (KeywordExtractionAPI.WORDSFRE wf in wordsfre)

{

Console.WriteLine(wf.Word + "\t" + wf.ED);

}

Console.ReadKey();

}

}

}