NLP 文档

2018-03-22

百度云

cloud.baidu.com

目录

1 API参考 1

1.1 简介 1

1.1.1 接口能力 1

1.1.2 请求格式 2

1.1.3 返回格式 2

1.2 调用方式 2

1.2.1 调用方式一 2

1.2.2 调用方式二 2

1.3 词法分析接口 3

1.3.1 接口描述 3

1.3.2 请求说明 3

1.3.3 返回说明 4

1.4 依存句法分析接口 8

1.4.1 接口描述 8

1.4.2 请求说明 8

1.4.3 返回说明 9

1.5 词向量表示接口 15

1.5.1 接口描述 15

1.5.2 请求说明 16

1.5.3 返回说明 16

1.6 DNN语言模型接口 17

1.6.1 接口描述 . . . 17
1.6.2 请求说明 . . . 17
1.6.3 返回说明 . . . 18
1.7 词义相似度接口 . . . 19
1.7.1 接口描述 . . . 19
1.7.2 请求说明 . . . 19
1.7.3 返回说明 . . . 20
1.8 短文本相似度接口 . . . 21
1.8.1 接口描述 . . . 21
1.8.2 请求说明 . . . 21
1.8.3 返回说明 . . . 22
1.9 评论观点抽取接口 . . . 23
1.9.1 接口描述 . . . 23
1.9.2 请求说明 . . . 23
1.9.3 返回说明 . . . 25
1.10 情感倾向分析接口 . . . 26
1.10.1 接口描述 . . . 26
1.10.2 请求说明 . . . 26
1.10.3 返回说明 . . . 27
1.11 文章标签接口 . . . 28
1.11.1 接口描述 . . . 28
1.11.2 请求说明 . . . 28
1.11.3 返回说明 . . . 29
1.12 文章分类接口 . . . 30
1.12.1 接口描述 . . . 30
1.12.2 请求说明 . . . 30
1.12.3 返回说明 . . . 31
1.13 分词接口（旧版） . . . 32

1.13.1 接口描述 . . . 32
1.13.2 请求说明 . . . 32
1.13.3 返回说明 . . . 33
1.14 词性标注接口（旧版） . . . 37
1.14.1 接口描述 . . . 37
1.14.2 请求说明 . . . 37
1.14.3 返回说明 . . . 38
1.15 中文词向量表示接口（旧版） . . . 42
1.15.1 接口描述 . . . 42
1.15.2 请求说明 . . . 42
1.15.3 返回说明 . . . 43
1.16 中文DNN语言模型接口（旧版） . . . 44
1.16.1 接口描述 . . . 44
1.16.2 请求说明 . . . 44
1.16.3 返回说明 . . . 45
1.17 短文本相似度接口（旧版） . . . 45
1.17.1 接口描述 . . . 45
1.17.2 请求说明 . . . 45
1.17.3 返回说明 . . . 46
1.18 评论观点抽取接口（旧版） . . . 48
1.18.1 接口描述 . . . 48
1.18.2 请求说明 . . . 48
1.18.3 返回说明 . . . 49
1.19 错误信息 . . . 50
1.19.1 错误码 . . . 50
2 C# SDK文档 53
2.1 简介 . . . 53

2.1.1 接口能力 . . . 53
2.1.2 版本更新记录 . . . 53
2.2 快速入门 . . . 54
2.2.1 安装自然语言处理 C# SDK . . . 54
方法一：使用Nuget管理 . . . 54
方法二：下载安装 . . . 54
2.2.2 新建交互类 . . . 55
2.3 接口说明 . . . 55
2.3.1 词法分析 . . . 55
2.3.2 词法分析（定制版） . . . 59
2.3.3 依存句法分析 . . . 61
2.3.4 词向量表示 . . . 63
2.3.5 DNN语言模型 . . . 64
2.3.6 词义相似度 . . . 65
2.3.7 短文本相似度 . . . 67
2.3.8 评论观点抽取 . . . 68
2.3.9 情感倾向分析 . . . 70
2.3.10 文章标签 . . . 71
2.3.11 文章分类 . . . 72
2.3.12 FAQ . . . 74
1. throw exception “fail to fetch token: 基础连接已关闭” . . 74
2. SSL报错 “The authentication or decryption has failed” . . 74
2.4 错误信息 . . . 74
2.4.1 错误返回格式 . . . 74
2.4.2 错误码 . . . 74
3 Java SDK文档 76
3.1 简介 . . . 76

3.1.1 接口能力 . . . 76
3.1.2 版本更新记录 . . . 76
3.2 快速入门 . . . 77
3.2.1 安装NLP Java SDK . . . 77
3.2.2 新建AipNlp . . . 78
3.2.3 配置AipNlp . . . 80
3.3 接口说明 . . . 81
3.3.1 词法分析 . . . 81
3.3.2 词法分析（定制版） . . . 85
3.3.3 依存句法分析 . . . 87
3.3.4 词向量表示 . . . 88
3.3.5 DNN语言模型 . . . 90
3.3.6 词义相似度 . . . 91
3.3.7 短文本相似度 . . . 92
3.3.8 评论观点抽取 . . . 94
3.3.9 情感倾向分析 . . . 95
3.3.10 文章标签 . . . 97
3.3.11 文章分类 . . . 98
3.4 错误信息 . . . 100
3.4.1 错误返回格式 . . . 100
3.4.2 错误码 . . . 100

4 Nodejs SDK文档 102
4.1 简介 . . . 102
4.1.1 接口能力 . . . 102
4.1.2 版本更新记录 . . . 102
4.2 快速入门 . . . 103
4.2.1 安装自然语言处理 Node SDK . . . 103

4.2.2 新建AipNlpClient ····· 104
4.3 接口说明 ····· 105
4.3.1 词法分析 ····· 105
4.3.2 词法分析（定制版） ····· 109
4.3.3 依存句法分析 ····· 111
4.3.4 词向量表示 ····· 113
4.3.5 DNN语言模型 ····· 114
4.3.6 词义相似度 ····· 116
4.3.7 短文本相似度 ····· 117
4.3.8 评论观点抽取 ····· 119
4.3.9 情感倾向分析 ····· 121
4.3.10 文章标签 ····· 122
4.3.11 文章分类 ····· 123
4.4 错误信息 ····· 125
4.4.1 错误返回格式 ····· 125
4.4.2 错误码 ····· 125
5 PHP SDK文档 127
5.1 简介 ····· 127
5.1.1 接口能力 ····· 127
5.1.2 版本更新记录 ····· 127
5.2 快速入门 ····· 128
5.2.1 安装自然语言处理 PHP SDK ····· 128
5.2.2 新建AipNlp ····· 128
5.2.3 配置AipNlp ····· 129
5.3 接口说明 ····· 129
5.3.1 词法分析 ····· 129
5.3.2 词法分析（定制版） ····· 133

5.3.3 依存句法分析 ····· 135
5.3.4 词向量表示 ····· 137
5.3.5 DNN语言模型 ····· 138
5.3.6 词义相似度 ····· 139
5.3.7 短文本相似度 ····· 141
5.3.8 评论观点抽取 ····· 142
5.3.9 情感倾向分析 ····· 144
5.3.10 文章标签 ····· 145
5.3.11 文章分类 ····· 146
5.4 错误信息 ····· 148
5.4.1 错误返回格式 ····· 148
5.4.2 错误码 ····· 148
6 Python SDK文档 150
6.1 简介 ····· 150
6.1.1 接口能力 ····· 150
6.1.2 版本更新记录 ····· 150
6.2 快速入门 ····· 151
6.2.1 安装自然语言处理 Python SDK ····· 151
6.2.2 新建AipNlp ····· 151
6.2.3 配置AipNlp ····· 152
6.3 接口说明 ····· 152
6.3.1 词法分析 ····· 152
6.3.2 词法分析（定制版） ····· 156
6.3.3 依存句法分析 ····· 158
6.3.4 词向量表示 ····· 160
6.3.5 DNN语言模型 ····· 161
6.3.6 词义相似度 ····· 162

6.3.7 短文本相似度 163
6.3.8 评论观点抽取 165
6.3.9 情感倾向分析 166
6.3.10 文章标签 167
6.3.11 文章分类 169
6.4 错误信息 170
6.4.1 错误返回格式 170
6.4.2 错误码 171
7 C++ SDK文档 173
7.1 简介 173
7.1.1 接口能力 173
7.1.2 版本更新记录 173
7.2 快速入门 174
7.2.1 安装自然语言处理 C++ SDK 174
7.2.2 新建client 175
7.3 接口说明 175
7.3.1 词法分析 175
7.3.2 词法分析（定制版） 179
7.3.3 依存句法分析 181
7.3.4 词向量表示 183
7.3.5 DNN语言模型 184
7.3.6 词义相似度 185
7.3.7 短文本相似度 187
7.3.8 评论观点抽取 188
7.3.9 情感倾向分析 190
7.3.10 文章标签 191

7.3.11 文章分类 192
7.4 错误信息 194
7.4.1 错误返回格式 194
7.4.2 错误码 194
8 常见问题 196

第1章　API参考

1.1　简介

Hi，您好，欢迎使用百度自然语言处理API服务。

本文档主要针对API开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流

1.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

1.1.2 请求格式

POST方式调用

注意：要求使用JSON格式的结构体来描述一个请求的具体内容。发送时默认需要对body整体进行GBK编码。若使用UTF-8编码，请在url参数中添加charset=UTF-8（大小写敏感）例如：https://aip.baidubce.com/rpc/2.0/nlp/v1/lexer?charset=UTF-8&access_token=24.f9ba9c5241b67688bb4adb2592000.1485570332.282335-8574074

1.1.3 返回格式

JSON格式，返回内容为GBK编码

1.2 调用方式

调用AI服务相关的API接口有两种调用方式，两种不同的调用方式采用相同的接口URL。

区别在于请求方式和鉴权方法不一样，请求参数和返回结果一致。

1.2.1 调用方式一

向API服务地址使用POST发送请求，必须在URL中带上参数：

access_token：必须参数，参考“Access Token获取。

POST中参数按照API接口说明调用即可。

例如自然语言处理API，使用HTTPS POST发送：

```
https://aip.baidubce.com/rpc/2.0/nlp/v1/lexer?access_token=24.f9ba9c5241b67688bb4adbed8bc91dec.2592000.1485570332.282335-8574074
```

获取access_token示例代码

{% AccessToken %}

> 说明： 方式一鉴权使用的Access_token必须通过API Key和Secret Key获取。

1.2.2 调用方式二

请求头域内容

NLP的API服务需要在请求的HTTP头域中包含以下信息：

- host（必填）
- x-bce-date（必填）
- x-bce-request-id（选填）
- authorization（必填）
- content-type（选填）
- content-length（选填）

作为示例，以下是一个标准的请求头域内容:

```
POST rpc/2.0/nlp/v1/wordseg? HTTP/1.1
accept-encoding: gzip, deflate
x-bce-date: 2015-03-24T13:02:00Z
connection: keep-alive
accept: */*
host: aip.baidubce.com
x-bce-request-id: 73c4e74c-3101-4a00-bf44-fe246959c05e
content-type: application/x-www-form-urlencoded;
authorization: bce-auth-v1/46bd9968a6194b4bbdf0341f2286ccce/2015-03-24T13:02:00Z/
1800/host;x-bce-date/994014d96b0eb26578e039fa053a4f9003425da4bfedf33f4790882fb4c54903
```

说明： 方式二鉴权使用的API认证机制authorization必须通过百度云的AK/SK生成。

1.3 词法分析接口

1.3.1 接口描述

（通用版）词法分析接口：向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。（定制版）词法分析接口：向用户提供分词、词性标注、专名识别三大功能；用户在控制台中进行个性化配置，支持自定义词表与规则，通过定制版可有效识别应用场景中的小众词汇与类别。

1.3.2 请求说明

请求示例一

- HTTP方法: POST

- （通用版）请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/lexer
- （定制版）请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/lexer_custom
- URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

- Header如下：

参数	值
Content-Type	application/json

- body请求示例:

```
{
  "text": "百度是一家高科技公司"
}
```

请求参数

参数名称	**类型**	**详细说明**
text	string	待分析文本（目前仅支持GBK编码），长度不超过20000字节

1.3.3 返回说明

返回参数

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array of objects	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串

参数名称	类型	**必需**	详细说明
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array of strings	是	基本词成分
+loc_details	array of objects	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

返回示例

```
{
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["是"]
         },
         {
           "byte_length":4,
           "byte_offset":6,
           "formal":"",
           "item":"一家",
           "ne":"",
           "pos":"m",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["一","家"]
         },
         {
           "byte_length":6,
           "byte_offset":10,
           "formal":"",
           "item":"高科技",
           "ne":"",
           "pos":"n",
           "uri":"",
```

```
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
      ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **
PER	人名	LOC	地名	ORG	机构名	TIME	时间

1.4 依存句法分析接口

1.4.1 接口描述

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

1.4.2 请求说明

请求示例一

- HTTP方法: POST
- 请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/depparser
- URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

- Header如下：

参数	值
Content-Type	application/json

- body请求示例:

```
{
"text": "今天天气怎么样",
"mode": 1
}
```

请求参数

参数名称	**类型**	**是否必须**	**描述**
text	string	是	待分析文本（目前仅支持GBK编码），长度不超过256字节
mode	int	否	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

关于模型选择

依存句法分析接口，可由用户自主选择合适的模型：

Query模型：该模型的训练数据来源于用户在百度的日常搜索数据，适用于处理信息需求类的搜索或口语query。

例如：

“手机缝隙灰尘怎么清除”

“百度云登陆首页”

“给我订一张明天上海到北京的飞机票”

Web模型：该模型的训练数据来源于全网网页数据，适用于处理网页文本等书面表达句子。

例如：

“后台任务定义为某过程在用户当时未登录机器期间运行”

“一般而言,股份的表现形式可以是股票、股权份额等等”

“有两条途径可以让本土的商业智慧在西方发扬光大”

1.4.3 返回说明

返回参数

参数名称	类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	int	词的ID
word	string	词

参数名称	类型	详细说明
postag	string	词性，请参照下方**词性（postag)取值范围**
head	int	词的父节点ID
deprel	string	词与父节点的依存关系，请参照下方**依存关系标识**

返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
}
```

词性取值范围

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
Ag	形语素	g	语素	ns	地名	u	助词
a	形容词	h	前接成分	nt	机构团体	vg	动语素
ad	副形词	i	成语	nz	其他专名	v	动词
an	名形词	j	简称略语	o	拟声词	vd	副动词
b	区别词	k	后接成分	p	介词	vn	名动词
c	连词	l	习用语	q	量词	w	标点符号
dg	副语素	m	数词	r	代词	x	非语素字
d	副词	Ng	名语素	s	处所词	y	语气词
e	叹词	n	名词	tg	时语素	z	状态词
f	方位词	nr	人名	t	时间词	un	未知词

依存关系标识

句法依存关系接口可以解析出的依存关系标识如下：

1.定中关系ATT

定中关系就是定语和中心词之间的关系，定语对中心词起修饰或限制作用。

如：工人/n师傅/n（工人/n ← 师傅/n）。

2. 数量关系QUN（quantity）

数量关系是指量词或名词同前面的数词之间的关系，该关系中，数词作修饰成分，依存于量词或名词。

如：三/m天/q（三/m ← 天/q）。

3.并列关系COO（coordinate）

并列关系是指两个相同类型的词并列在一起。

如：奔腾/v咆哮/v的怒江激流（奔腾/v → 咆哮/v）。

4.同位关系APP（appositive）

同位语是指所指相同、句法功能也相同的两个并列的词或词组。

如：我们大家（我们 → 大家）。

5.附加关系ADJ（adjunct）

附加关系是一些附属词语对名词等成分的一种补充说明，使意思更加完整，有时候去掉也不影响意思。

如：约/d 二十/m 多/m 米/q 远/a 处/n（二十/m → 多/m，米/q → 远/a）。

6.动宾关系VOB（verb-object）

对于动词和宾语之间的关系我们定义了两个层次，一是句子的谓语动词及其宾语之间的关系，我们定为OBJ，在下面的单句依存关系中说明；二是非谓语动词及其宾语的关系，即VOB。这两种关系在结构上没有区别，只是在语法功能上，OBJ中的两个词充当句子的谓语动词和宾语，VOB中的两个词构成动宾短语，作为句子的其他修饰成分。

如：历时/v 三/m 天/q 三/m夜/q（历时/v → 天/q）。

7.介宾关系POB（preposition-object）

介词和宾语之间的关系，介词的属性同动词相似。

如：距/p球门/n（距/p → 球门/n）。

8.主谓关系SBV（subject-verb）

主谓关系是指名词和动作之间的关系。

如：父亲/n 逝世/v 1 0/m 周年/q 之际/nd（父亲/n ← 逝世/v）。

9.比拟关系SIM（similarity）

比拟关系是汉语中用于表达比喻的一种修辞结构。

如：炮筒/n 似的/u 望远镜/n（炮筒/n ← 似的/u）。

10.时间关系TMP（temporal）

时间关系定义的是时间状语和其所修饰的中心动词之间的关系。

如：十点以前到公司（以前 ← 到）。

11.处所关系LOC（locative）

处所关系定义的是处所状语和其所修饰的中心动词之间的关系，如：在公园里玩耍（在 ← 玩耍）。

12.“的”字结构DE

“的”字结构是指结构助词“的”和其前面的修饰语以及后面的中心词之间的关系。

如：上海/ns 的/u 工人/n（上海/ns ← 的/u，的/u ← 工人/n）。

13.“地”字结构DI

“地”字结构在构成上同DE类似，只是在功能上不同，DI通常作状语修饰动词。

如： 方便/a 地/u 告诉/v 计算机/n（方便/a ← 地/u，地/u ← 告诉/v）。

14.“得”字结构DEI

助词“得”同其后的形容词或动词短语等构成“得”字结构，对前面的动词进行补充说明。

如：讲/v 得/u 很/d 对/a（讲/v → 得/u，得/u → 对/a）。

15.“所”字结构SUO

“所”字为一结构助词，后接一宾语悬空的动词做“的”字结构的修饰语，“的”字经常被省略，使结构更加简洁。

如：机电/b 产品/n 所/u 占/v 比重/n 稳步/d 上升/v（所/u ← 占/v）。

16.“把”字结构BA

把字句是主谓句的一种，句中谓语一般都是及物动词。

如：我们把豹子打死了（把/p → 豹子/n）。

17.“被”字结构BEI

被字句是被动句，是主语接受动作的句子。

如：豹子被我们打死了（豹子/n ← 被/p）。

18.状中结构ADV（adverbial）

状中结构是谓词性的中心词和其前面的修饰语之间的关系，中心词做谓语时，前面的修饰成分即为句子的状语，中心词多为动词、形容词，修饰语多为副词，介词短语等。

如：连夜/d 安排/v 就位/v（连夜/d ← 安排/v）。

19.动补结构CMP（complement）

补语用于对核心动词的补充说明。

如：做完了作业（做/v → 完）。

20.兼语结构DBL（double）

兼语句一般有两个动词，第二个动词是第一个动作所要表达的目的或产生的结果。

如：[7]曾经/d [8]使/v [9]多少/r [10]旅游/n [11]人/n [12]隔/v [13]岸/n [14]惊叹/v [15]！/wp（使 → 人/n，/v使/v → 惊叹/v）。

21.关联词CNJ（conjunction）

关联词语是复句的有机部分。

如：只要他请客，我就来。（只要 ← 请，就 ← 来）。

22.关联结构 CS(conjunctive structure)

当句子中存在关联结构时，关联词所在的两个句子（或者两个部分）之间通过各部分的核心词发生依存关系CS。

如：只要他请客，我就来。（请 ← 来）。

23.语态结构MT（mood-tense）

汉语中，经常用一些助词表达句子的时态和语气，这些助词分语气助词，如：吧，啊，呢等；还有时态助词，如：着，了，过。

如： [12]答应/v [13]孩子/n [14]们/k [15]的/u [16]要求/n [17]吧/u [18]，/wp [19]他们/r [20]这/r [21]是/v [22]干/v [23]事业/n [24]啊/u [25]！/wp（[12]答应/v ← [17]吧/u，[21]是/v ← [24]啊/u）。

24.连谓结构VV（verb-verb）

连谓结构是同多项谓词性成分连用、这些成分间没有语音停顿、书面标点，也没有关联词语，没有分句间的逻辑关系，且共用一个主语。

如：美国总统来华访问。（来华/v → 访问/v）。

25.核心HED（head）

该核心是指整个句子的核心，一般是句子的核心词和虚拟词（<EOS>或ROOT）的依存关系。

如：这/r 就是/v恩施/ns最/d]便宜/a的/u出租车/n，/wp相当于/v北京/ns的/u “/wp 面的/n”/wp 。/wp <EOS>/<EOS>（就是/v ← <EOS>/<EOS>）

26.前置宾语FOB（fronting object）

在汉语中，有时将句子的宾语前置，或移置句首，或移置主语和谓语之间，以起强调作用，我认识这个人 ← 这个人我认识。

如：他什么书都读（书/n ← 读/v）。

27.双宾语DOB（double object）

动词后出现两个宾语的句子叫双宾语句，分别是直接宾语和间接宾语。

如：我送她一束花。（送/v → 她/r，送/v → 花/n）。

28.主题TOP（topic）

在表达中，我们经常会先提出一个主题性的内容，然后对其进行阐述说明；而主题部分与后面的说明部分并没有直接的语法关系，主题部分依存于后面的核心成分，且依存关系为TOP。

如：西直门，怎么走？（西直门 ← 走）。

29.独立结构IS（independent structure）

独立成分在句子中不与其他成分产生结构关系，但意义上又是全句所必需的，具有相对独立性的一种成分。

如：事情明摆着，我们能不管吗？

30.独立分句IC（independent clause）

两个单句在结构上彼此独立，都有各自的主语和谓语。

如：我是中国人，我们爱自己的祖国。（是 → 爱）

31.依存分句DC（dependent clause）

两个单句在结构上不是各自独立的，后一个分句的主语在形式上被省略，但不是前一个分句的主语，而是存在于前一个分句的其他成分中，如宾语、主题等成分。规定后一个分句的核心词依存于前一个分句的核心词。该关系同连谓结构的区别是两个谓词是否为同一主语，如为同一主语，则为VV，否则为DC。

如：大家/r叫/v 它/r “/wp 麻木/a 车/n”/wp ，/wp 听/v起来/v 怪怪的/a 。/wp（叫/v → 听/v）。

32.叠词关系VNV （verb-no-verb or verb-one-verb）

如果叠词被分开了，如“是 不 是”、“看一看”，那么这几个词先合并在一起，然后预存到其他词上，叠词的内部关系定义为：(是1→不；不→是2）。

33.一个词YGC

当专名或者联绵词等切散后，他们之间本身没有语法关系，应该合起来才是一个词。如：百度。

34.标点 WP

大部分标点依存于其前面句子的核心词上，依存关系WP。

1.5 词向量表示接口

1.5.1 接口描述

> 本接口已于2017年5月25日升级，仅支持词向量查询。如果希望查询两个词的相似度，可使用词义相似度。

> 如果您需要查阅旧版接口文档，请查看中文词向量表示接口（旧版），但建议您尽快升级到新版接口。

词向量表示接口提供中文词向量的查询功能。

1.5.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v2/word_emb_vec

URL参数:

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下:

参数	值
Content-Type	application/json

Body请求示例:

```
{
    "word":"张飞"
}
```

请求参数

参数	是否必选	类型	描述
word	是	string	文本内容（GBK编码），最大64字节
dem	否	int	词向量维度。默认值为0(对应1024维)，目前仅支持dem=0

1.5.3 返回说明

返回参数

参数	类型	描述
log_id	uint64	请求唯一标识码

参数	类型	描述
word	string	查询词
vec	float	词向量结果表示

返回示例


```
{
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
    0.283711,
    -0.233656,
    0.555556
  ]
}
```


1.6 DNN语言模型接口

1.6.1 接口描述

> 本接口已于2017年5月25日升级，如果您需要查阅旧版接口文档，请查看中文DNN语言模型（旧版），但建议您尽快升级到新版接口。

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

1.6.2 请求说明

请求示例

HTTP方法：POST

请求URL： https://aip.baidubce.com/rpc/2.0/nlp/v2/dnnlm_cn

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
    "text":"床前明月光"
}
```

请求参数

参数	类型	描述
text	string	文本内容（GBK编码），最大256字节，不需要切词

1.6.3 返回说明

返回参数

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺

返回示例

```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```


1.7 词义相似度接口

1.7.1 接口描述

输入两个词，得到两个词的相似度结果。

1.7.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v2/word_emb_sim

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
    "word_1":"北京",
    "word_2":"上海"
}
```

请求参数

参数	是否必选	类型	描述
word_1	是	string	词1（GBK编码），最大64字节
word_2	是	string	词2（GBK编码），最大64字节
mode	否	int	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

1.7.3 返回说明

返回参数

参数	类型	描述
log_id	uint64	请求唯一标识码,随机数
score	float	相似度结果，(0,1]，分数越高说明相似度越高

返回示例

```
{
    "score": 0.456862,
    "words": {
```

```
    "word_1": "北京",
    "word_2": "上海"
  }
}
```

1.8 短文本相似度接口

1.8.1 接口描述

> 本接口已于2017年6月15日升级，如果您需要查阅旧版接口文档，请查看短文本相似度接口（旧版），但建议您尽快升级到新版接口。

短文本相似度接口用来判断两个文本的相似度得分。

1.8.2 请求说明

请求示例

HTTP方法：POST

请求URL：https://aip.baidubce.com/rpc/2.0/nlp/v2/simnet

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
"text_1": "浙富股份", // 待比较文本1
"text_2": "万事通自考网" // 待比较文本2
}
```

请求参数

参数	类型	是否必须	描述
text_1	string	是	待比较文本1（GBK编码），最大512字节
text_2	string	是	待比较文本2（GBK编码），最大512字节
model	string	否	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

关于模型选择

短文本相似度接口，可由用户自主选择合适的模型：

BOW（词包）模型

基于bag of words的BOW模型，特点是泛化性强，效率高，比较轻量级，适合任务：输入序列的 term “确切匹配”、不关心序列的词序关系，对计算效率有很高要求；

GRNN（循环神经网络）模型

基于recurrent，擅长捕捉短文本“跨片段”的序列片段关系，适合任务：对语义泛化要求很高，对输入语序比较敏感的任务；

CNN（卷积神经网络）模型

模型语义泛化能力介于 BOW/RNN 之间，对序列输入敏感，相较于 GRNN 模型的一个显著优点是计算效率会更高些。

1.8.3 返回说明

返回参数

参数	描述	取值
log_id	uint64	随机数，请求唯一标识码
score	float	相似度结果取值(0,1]，分数越高说明相似度越高

返回示例

```
{
"log_id": 12345,
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
    "score":0.3300237655639648 //相似度结果
},
```

1.9 评论观点抽取接口

1.9.1 接口描述

本接口已于2017年5月25日升级，如果您需要查阅旧版接口文档，请查看评论观点抽取接口（旧版），但建议您尽快升级到新版接口。

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

1.9.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v2/comment_tag

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
    "text":"三星电脑电池不给力",
    "type":13
}
```

请求参数

参数	是否必选	类型	描述
text	必选	string	评论内容（GBK编码），最大10240字节
type	可选	int	评论行业类型，默认为4（餐饮美食）

其中type包含13个类别，具体取值说明如下：

type参数	说明	实例
1	酒店	『酒店设备齐全、干净卫生』->『酒店设备齐全』、『干净卫生』
2	KTV	『环境一般般把，音响设备也一般，隔音太差』->『环境一般』、『音响设备一般』、『隔音差』
3	丽人	『手法专业，重要的是效果很棒』->『手法专业』、『效果不错』
4	美食餐饮	『但是味道太好啦，舍不得剩下』->『味道不错』
5	旅游	『景区交通方便，是不错的旅游景点』->『交通方便』、『旅游景点不错』
6	健康	『环境很棒，技师服务热情』->『环境不错』、『服务热情』
7	教育	『教学质量不错，老师很有经验』->『教学质量不错』、『老师有经验』
8	商业	『该公司服务好，收费低，效率高』->『服务好』、『收费低』、『效率高』

type参数	说明	实例
9	房产	『该房周围设施齐全、出行十分方便』->『设施齐全』、『出行方便』
10	汽车	『路宝的优点就是安全性能高、空间大』->『安全性能高』、『空间大』
11	生活	『速度挺快、服务态度也不错』->『速度快』、『服务好』
12	购物	『他家的东西还是挺贵的』->『消费贵』
13	3C	『手机待机时间长』->『待机时间长』

1.9.3 返回说明

返回参数

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

返回示例

```
{
"items": [
{
```

```
    "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
        "end_pos": 18,
"abstract":"三星电脑<span>电池不给力</span>"
    }
    ]
}
```


1.10 情感倾向分析接口

1.10.1 接口描述

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

1.10.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/sentiment_classify

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
    "text": "苹果是一家伟大的公司"
}
```

请求参数

参数	类型	描述
text	string	文本内容（GBK编码），最大2048字节

1.10.3 返回说明

返回参数

参数	说明	描述
log_id	uint64	请求唯一标识码
sentiment	string	表示情感极性分类结果
confidence	float	表示分类的置信度，取值范围[0,1]
positive_prob	float	表示属于积极类别的概率，取值范围[0,1]
negative_prob	float	表示属于消极类别的概率，取值范围[0,1]

返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
        }
    ]
}
```

1.11 文章标签接口

1.11.1 接口描述

文本标签服务对文章的标题和内容进行深度分析，输出能够反映文章关键信息的主题、话题、实体等多维度标签以及对应的置信度，该技术在个性化推荐、文章聚合、内容检索等场景具有广泛的应用价值。

1.11.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/keyword

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
"title":"iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下",
"content": "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。在通电的情况下掉进清水，这种情况一不需要拆机处理。尽快断电。用力甩干，但别把机器甩掉，主意要把屏幕内的水甩出来。如果屏幕残留有水滴，干后会有痕迹。^H3 放在台灯，射灯等轻微热源下让水分慢慢散去。"
}
```

请求参数

参数	类型	描述	是否必填
title	string	文章标题（GBK编码），最大80字节	必填
content	string	文章内容（GBK编码），最大65535字节	必填

1.11.3 返回说明

返回参数

参数	说明	描述
items	array of objects	分析结果数组
+tag	string	内容标签
+score	float	权重值，取值范围[0,1]

返回示例


```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
```

```
        }
    ]
}
```

1.12 文章分类接口

1.12.1 接口描述

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。 目前支持的分类类目如下：1、国际 2、体育 3、娱乐 4、社会 5、财经 6、时事 7、科技 8、情感 9、汽车 10、教育 11、时尚 12、游戏 13、军事 14、旅游 15、美食 16、文化 17、健康养生 18、搞笑 19、家居 20、动漫 21、宠物 22、母婴育儿 23、星座运势 24、历史 25、音乐 26、综合

1.12.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/topic

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
"title":"欧洲冠军联赛",
"content": "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。"
}
```

请求参数

参数	类型	描述	是否必填
title	string	文章标题（GBK编码），最大80字节	必填
content	string	文章内容（GBK编码），最大65535字节	必填

1.12.3 返回说明

返回参数

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

返回示例


```
{
    "log_id": 3591049593939822907,
    "item": {
        "lv2_tag_list": [
            {
                "score": 0.877436,
                "tag": "足球"
            },
            {
                "score": 0.793682,
                "tag": "国际足球"
            },
            {
                "score": 0.775911,
                "tag": "英超"
            }
```

```
        ],
        "lv1_tag_list": [
            {
                "score": 0.824329,
                "tag": "体育"
            }
        ]
    }
}
```

1.13 分词接口（旧版）

1.13.1 接口描述

本接口已合并到词法分析接口，即将逐步下线，建议直接使用词法分析接口。

分词接口提供基本词和混排两种粒度的分词结果，基本词粒度较小，适用于搜索引擎等需要更多召回的任务，而混排粒度倾向于保留更多的短语。

1.13.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/wordseg

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
  "query":"百度是一家高科技公司",
  "lang_id":1
}
```

请求参数

参数名称 | 类型 | 是否必选 | 详细说明 |
——— | —— | —— |————————————— |
query | String | 是 | 待分词的文本（GBK编码的URL编码形式） |
lang_id | Int | 否 | 默认为1，输入字符串的语言对应的id，简体中文设置为1（目前不支持其他语言） |

请求示例代码

提示一：使用示例代码前，请记得替换其中的示例Token、图片地址或Base64信息。

提示二：部分语言依赖的类或库，请在代码注释中查看下载地址。

{% NLP-API-Wordseg %}

1.13.3 返回说明

返回参数

参数名称	类型	详细说明
wordsepbuf	String	基本词粒度结果，以\t分割
wsbtermcount	int	基本词粒度输出的词个数
wsbtermoffsets	List	该参数为列表，元素个数为切分出来的词个数，每个元素值表示对应的基本词在被切分文本的起始位置（字节偏移）
wsbtermpos	List	参数值为列表，元素值为对应切分出来的基本词在wordsepbuf的字节偏移以及长度，整数的低24bit为偏移，高8bit为长度
wpcompbuf	String	混排粒度结果，以\t分割
wpbtermcount	Int	混排粒度输出的词个数

参数名称	类型	详细说明
wpbtermoffsets	List	该参数为列表，元素个数为切分出来的词个数，每个元素值表示对应的词是从第几个基本词开始的（基本词偏移）
wpbtermpos	List	参数值为列表，元素值为对应切分出来的词在 wp-compbuf的字节偏移以及长度，整数的低24bit为偏移，高8bit为长度
subphrbuf	String	所有识别出来的短语，以\t分割
spbtermcount	Int	识别出来的短语个数
spbtermoffsets	List	该参数为列表，元素个数为识别出来的短语个数，每个元素值表示对应短语是从第几个基本词开始的（基本词偏移）
spbtermpos	List	参数值为列表，元素值为对应切分出来的短语在subphrbuf的字节偏移以及长度，整数的低24bit为偏移，高8bit为长度

返回示例


```
{
    "scw_out": {
        "phrase_merged": 0,
        "pdisambword": {
            "newwordbuf": "",
            "newwordb_curpos": 0,
            "newwordbmaxcount": 0,
            "newwordbsize": 0,
            "newwordbtermcount": 0,
            "newwordneprop": [],
            "newwordbtermoffsets": [],
            "newwordbtermpos": []
        },
        "pnewword": {
```

```
        "newwordbuf": "",
        "newwordb_curpos": 0,
        "newwordbmaxcount": 0,
        "newwordbsize": 0,
        "newwordbtermcount": 0,
        "newwordneprop": [],
        "newwordbtermoffsets": [],
        "newwordbtermpos": []
    },
    "booknamebuf": "",
    "mergebuf": "",
    "namebuf": "",
    "subphrbuf": "\\t\\u4f60\\u597d\\t",
    "wordsepbuf": "\\t\\u4f60\\t\\u597d\\t\\u767e\\u5ea6\\t",
    "wpcompbuf": "\\t\\u4f60\\u597d\\t\\u767e\\u5ea6\\t",
    "bnb_curpos": 0,
    "bnbsize": 0,
    "bnbtermcount": 0,
    "mb_curpos": 0,
    "mbsize": 0,
    "mbtermcount": 0,
    "nameb_curpos": 0,
    "namebsize": 0,
    "namebtermcount": 0,
    "spb_curpos": 6,
    "spbsize": 1024000,
    "spbtermcount": 1,
    "wordtotallen": 8,
    "wpb_curpos": 11,
    "wpbsize": 1024000,
    "wpbtermcount": 2,
    "wsb_curpos": 12,
    "wsbsize": 1024000,
    "wsbtermcount": 3,
    "bnbtermprop": [],
    "namebtermprop": [],
    "spbtermprop": [
        {
            "m_hprop": 1,
            "m_lprop": 32
        }
    ],
    "wpbtermprop": [
        {
            "m_hprop": 1,
```

```
            "m_lprop": 32
        },
        {
            "m_hprop": 0,
            "m_lprop": 32
        }
    ],
    "wsbtermprop": [
        {
            "m_hprop": 0,
            "m_lprop": 32
        },
        {
            "m_hprop": 0,
            "m_lprop": 32
        },
        {
            "m_hprop": 0,
            "m_lprop": 32
        }
    ],
    "bnbtermoffsets": [],
    "bnbtermpos": [],
    "mbtermoffsets": [],
    "mbtermpos": [],
    "namebtermoffsets": [],
    "namebtermpos": [],
    "spbtermoffsets": [
        0
    ],
    "spbtermpos": [
        67108865
    ],
    "wpbtermoffsets": [
        0,
        2
    ],
    "wpbtermpos": [
        67108865,
        67108870
    ],
    "wsbtermoffsets": [
        0,
        2,
        4
```

```
        ],
        "wsbtermpos": [
            33554433,
            33554436,
            67108871
        ]
    }
}
```

1.14 词性标注接口（旧版）

1.14.1 接口描述

本接口已合并到词法分析接口，即将下线，建议直接使用词法分析接口

词性标注接口为分词结果中的每个单词标注一个正确的词性的程序，也标注每个词是名词、动词、形容词或其他词性。

1.14.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/wordpos

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
```

```
    "query": "你好百度"
  }
```

请求参数

Key	类型	含义及备注
query	string	带标注的文本串。算法内部使用GBK编码，外部如果要求UTF8编码，则需进行编码转换。

请求示例代码

提示一：使用示例代码前，请记得替换其中的示例Token、图片地址或Base64信息。

提示二：部分语言依赖的类或库，请在代码注释中查看下载地址。

{% NLP-API-Wordpos %}

1.14.3 返回说明

返回参数

Key	类型	含义及备注
result_out	array	词性标注结果数组，数组中每个元素对应一个词汇。每个词汇是一个dict
+word	string	词汇的字面
+offset	int	偏移量，以基本粒度词汇为单位
+length	int	长度，以基本粒度词汇为单位
+type	string	词性
+confidence	float	置信度分值，0\~1

返回示例

```
{
  "result_out" :
     [
```

```
        {"word" : "你好", "offset" : 0, "length" : 1, "type" : "v", "confidence" : 1.0},
        {"word" : "百度", "offset" : 1, "length" : 1, "type" : "nz", "confidence" : 1.0}
      ]
}
```


词性缩略词说明

type	代码	名称	帮助记忆的诠释
1	Ag	形语素	形容词性语素。形容词代码为a，语素代码g前面置以A
2	Dg	副语素	副词性语素。副词代码为d，语素代码g前面置以D。
3	Ng	名语素	名词性语素。名词代码为n，语素代码g前面置以N。
4	Tg	时语素	时间词性语素。时间词代码为t,在语素的代码g前面置以T。
5	Vg	动语素	动词性语素。动词代码为v。在语素的代码g前面置以V
6	a	形容词	取英语形容词adjective的第1个字母
7	ad	副形词	直接作状语的形容词。形容词代码a和副词代码d并在一起。
8	an	名形词	具有名词功能的形容词。形容词代码a和名词代码n并在一起。
9	b	区别词	取汉字“别”的声母。
10	c	连词	取英语连词conjunction的第1个字母。

type	代码	名称	帮助记忆的诠释
11	d	副词	取adverb的第2个字母，因其第1个字母已用于形容词。
12	e	叹词	取英语叹词exclamation的第1个字母
13	f	方位词	取汉字“方”
14	g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母
15	h	前接成分	取英语head的第1个字母
16	i	成语	取英语成语idiom的第1个字母
17	j	简称略语	取汉字“简”的声母。
18	k	后接成分	
19	l	习用语	习用语尚未成为成语，有点“临时性”， 取“临”的声母。
20	m	数词	取英语numeral的第3个字母，n，u已有他用。
21	n	名词	取英语名词noun的第1个字母。
22	nr	人名	名词代码n和“人(ren)”的声母并在一起。
23	ns	地名	名词代码n和处所词代码s并在一起。
24	nt	机构团体	“团”的声母为t，名词代码n和t并在一起。
25	nx	外文专名	一般是全角英文专名，如：ＺＢＴ

type	代码	名称	帮助记忆的诠释
26	nz	其他专名	“专”的声母的第1个字母为z，名词代码n和z并在一起
27	o	拟声词	取英语拟声词onomatopoeia的第1个字母。
28	p	介词	取英语介词prepositional的第1个字母。
29	q	量词	取英语quantity的第1个字母。
30	r	代词	取英语代词pronoun的第2个字母,因p已用于介词。
31	s	处所词	取英语space的第1个字母。
32	t	时间词	取英语time的第1个字母。
33	u	助词	取英语助词auxiliary
34	v	动词	取英语动词verb的第一个字母。
35	vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
36	vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起
37	w	标点符号	
38	y	语气词	取汉字“语”的声母。
39	z	状态词	取汉字“状”的声母的前一个字母。

1.15 中文词向量表示接口（旧版）

1.15.1 接口描述

新版中文词向量表示接口已上线，建议您及时升级，旧版本即将逐步下线。

中文词向量表示接口提供两种功能：输入两个词tid=1得到两个词的相似度结果，输入1个词tid=2得到词的词向量。

1.15.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/wordembedding

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

输入两个词

```
{
  "query1":"百度",
  "query2":"谷歌",
  "tid":1
}
```

输入一个词

```
{
```

```
    "query1":"百度",
    "tid":2
  }
```

请求参数

参数	是否必选	说明
query1	是	输入的第一个词
query2	是	输入的第二个词
tid	是	指定算法类型，tid=1，返回两个词的相似度；tid=2，返回词向量

请求示例代码

提示一：使用示例代码前，请记得替换其中的示例Token、图片地址或Base64信息。

提示二：部分语言依赖的类或库，请在代码注释中查看下载地址。

{% NLP-API-Wordembedding %}

1.15.3 返回说明

返回参数

参数	描述
ret	属性对象的集合
message	词汇的字面
data	返回数据
+vec	词向量结果
+sim	相似度对象
++sim	相似度

返回示例

```
{
  "ret":0,
  "message":"",
  "data":
    {  "sim":
```

```
        {
          "sim":0.180343},
          "vec":null
        }
}
```

1.16 中文DNN语言模型接口（旧版）

1.16.1 接口描述

新版中文DNN语言模型已上线，建议您及时升级，旧版本即将逐步下线。

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值。

1.16.2 请求说明

请求示例

HTTP方法：POST

请求URL：https://aip.baidubce.com/rpc/2.0/nlp/v1/dnnlm_cn

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
  "input_sequence":"百度是个搜索公司"
}
```

请求参数

参数 | 说明 |
——————— | ———- |
input_sequence | 输入的句子，不需要切词 |

请求示例代码

提示一：使用示例代码前，请记得替换其中的示例Token。

提示二：部分语言依赖的类或库，请在代码注释中查看下载地址。

{% NLP-API-Dnnlm-cn %}

1.16.3 返回说明

返回参数

参数	说明
seq\seg	句子的切词结果		seq\prob	切词后每个词在句子中的概率值

返回示例

```
{
  "seq_seg":"百度 是 个 搜索 公司",
  "seq_prob":" 0.00059052   0.00373688  0.0372463   0.00137015  0.000118814 "
}
```

1.17 短文本相似度接口（旧版）

1.17.1 接口描述

> 新版短文本相似度接口已上线，建议您及时升级，旧版本即将逐步下线。

短文本相似度接口用来判断两个文本的相似度得分。

1.17.2 请求说明

请求示例

HTTP方法：POST

请求URL：https://aip.baidubce.com/rpc/2.0/nlp/v1/simnet

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
  "input":
    {
      "qslots":[{"terms_sequence":"你好百度” ", "type":0, "items":[]}],
      "tslots":[{"terms_sequence":"你好世界” ", "type":0, "items":[]}],
      "type":0
    }
}
```

请求参数

参数	说明
qslots中的terms\sequence | 短文本1 | | tslots中的terms\sequence	短文本2
items	均设置为空列表
type	类别，均设置为0

请求示例代码

提示一：使用示例代码前，请记得替换其中的示例Token。

提示二：部分语言依赖的类或库，请在代码注释中查看下载地址。

{% NLP-API-Simnet %}

1.17.3 返回说明

返回参数

参数 | 说明 |
———– | ———– |
score | 两个文本相似度得分 |
error | error code |
type | 默认为0 |
error_note | error对应文字说明 |
items | 默认为空 |

返回示例

```
{
  "output":
    {
      "score":0.758419,
      "error":0,
      "type":0,
      "error_note":"",
      "items":[]
    }
}
```

此接口业务错误码说明

Code	Message	返回说明
0	NO\ERROR | 正确返回 | | 1 | BEYOND\SLOT\LENGTH | 输入长度过长 | | 2 | OOV \ERROR	输入文本不在词表中
3	LEGO\LIB\RET\ERROR | 内部库错误 | | 4 | OTHER \SERVER \ERROR | 其它服务错误 | | 5 | INPUT\HAS \EMPTY | 输入为空 | | 6 | INPUT\FORMAT\ERROR | 输入格式错误 | | 7 | OTHER \CLIENT _ERROR | 客服端错误 |	

1.18 评论观点抽取接口（旧版）

1.18.1 接口描述

新版评论观点抽取接口已上线，建议您及时升级，旧版本即将逐步下线。

评论观点抽取接口用来提取一个句子观点评论的情感属性。

1.18.2 请求说明

请求示例

HTTP方法: POST

请求URL: https://aip.baidubce.com/rpc/2.0/nlp/v1/comment_tag

URL参数：

参数	值
access_token	通过API Key和Secret Key获取的access_token,参考“Access Token获取”

Header如下：

参数	值
Content-Type	application/json

Body请求示例:

```
{
  "comment":"个人觉得福克斯好，外观漂亮年轻，动力和操控性都不错",
  "entity":"NULL",
  "type":"10"
}
```

请求参数

参数	类型	说明
comment	string	评论内容
entity	string	实体名，当前取值为NULL，暂时不生效

参数	类型	说明
type	string	类别,默认类别为4（餐厅）

其中type包含12个类别，具体取值说明如下：

参数	说明
1	酒店
2	KTV
3	丽人
4	美食（默认）
5	旅游
6	健康
7	教育
8	商业
9	房产
10	汽车
11	生活
12	购物

请求示例代码

提示一：使用示例代码前，请记得替换其中的示例Token。

提示二：部分语言依赖的类或库，请在代码注释中查看下载地址。

{% NLP-API-Comment_tag %}

1.18.3 返回说明

返回参数

参数	说明
abstract	表示评论观点在评论文本中的位置。
adj	表示抽取结果中的评价词
comment	表示待抽取观点的评论文本。
entity	实体名，当前取值为NULL，暂时不生效
fea	抽取结果中的属性词

参数	说明
type	表示情感极性（其中2表示积极、1表示中性、0表示消极）。
其他参数	暂不生效

返回示例


```
{
  "abstract":"<span>动力和操控性都不错</span>",
  "adj":"不错",
  "comment":"个人觉得福克斯好，外观漂亮年轻，动力和操控性都不错",
  "entity":"NULL",
  "fea":"动力",
  "type":"2"
}
```


1.19 错误信息

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。
- error_msg：错误描述信息，帮助理解和解决发生的错误。

例如Access Token失效返回：


```
{
  "error_code": 110,
  "error_msg": "Access token invalid or no longer valid"
}
```


需要重新获取新的Access Token再次请求即可。

1.19.1 错误码

错误码	错误信息	描述
1	Unknown error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（224994340）或工单联系技术支持团队
2	Service temporarily unavailable	服务暂不可用，请再次请求，如果持续出现此类错误，请通过QQ群（224994340）或工单联系技术支持团队
3	Unsupported openapi method	调用的API不存在，请检查后重新尝试
4	Open api request limit reached	集群超限额
6	No permission to access data	无权限访问该用户数据
17	Open api daily request limit reached	每天请求量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额
100	Invalid parameter	包含了无效或错误参数，请检查代码
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试

错误码	错误信息	描述
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂 未 收录，无法比对相似度

第2章　C# SDK文档

2.1　简介

Hi，您好，欢迎使用百度自然语言处理服务。

本文档主要针对C#开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流：http://ai.baidu.com/forum/topic/list/169

2.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

2.1.2　版本更新记录

上线日期	版本号	更新内容
2018.1.25	3.3.2	新增文本分类接口
2018.1.11	3.3.1	新增文本标签接口
2017.12.21	3.3.0	接口升级
2017.9.12	3.0.0	更新SDK打包方式：所有AI服务集成一个SDK
2017.6.30	2.2.0	增加句法依存接口
2017.6.15	2.1.0	短文本相似度接口升级
2017.5.25	2.0.0	调用方式变更，新增多个接口
2017.5.4	1.0.0	第一版

2.2 快速入门

2.2.1 安装自然语言处理 C# SDK

C# SDK 现已开源! https://github.com/Baidu-AIP/dotnet-sdk

方法一：使用Nuget管理 在NuGet中搜索 `Baidu.AI`，安装最新版即可。

packet地址 https://www.nuget.org/packages/Baidu.AI/

方法二：下载安装 自然语言处理 C# SDK目录结构

```
Baidu.Aip
   ├── AipSdk.dll                    // 百度AI服务 windows 动态库
   ├── AipSdk.XML                    // DLL注释
   └── thirdparty                   // 第三方依赖
```

支持平台：.Net Framework 3.5 及以上版本

> 如果需要在 Unity / .net core 等平台使用，可引用工程源码自行编译。

安装

1.在官方网站下载C# SDK压缩工具包。

2.解压后，将 `AipSdk.dll` 和 thirdparty 中 的dll文件添加为引用。

2.2.2 新建交互类

Baidu.Aip.Nlp.Nlp是自然语言处理的交互类，为使用自然语言处理的开发人员提供了一系列的交互方法。

用户可以参考如下代码新建一个交互类：

```
// 设置APPID/AK/SK
var APP_ID = "你的 App ID";
var API_KEY = "你的 Api Key";
var SECRET_KEY = "你的 Secret Key";

var client = new Baidu.Aip.Nlp.Nlp(API_KEY, SECRET_KEY);
```

在上面代码中，常量APP_ID在百度云控制台中创建，常量API_KEY与SECRET_KEY是在创建完毕应用后，系统分配给用户的，均为字符串，用于标识用户，为访问做签名验证，可在AI服务控制台中的应用列表中查看。

注意：如您以前是百度云的老用户，其中API_KEY对应百度云的“Access Key ID”，SECRET_KEY对应百度云的“Access Key Secret”。

2.3 接口说明

2.3.1 词法分析

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
public void LexerDemo() {
var text = "百度是一家高科技公司";

// 调用词法分析，可能会抛出网络等异常，请使用try/catch捕获
var result = client.Lexer(text);
Console.WriteLine(result);
}
```

词法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分

参数名称	类型	**必需**	详细说明
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析 返回示例


```
{
  "status":0,
  "version":"ver_1_0_1",
  "results":[
    {
      "retcode":0,
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
```

```
            "uri":"",
            "loc_details":[ ],
            "basic_words":["是"]
          },
          {
            "byte_length":4,
            "byte_offset":6,
            "formal":"",
            "item":"一家",
            "ne":"",
            "pos":"m",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["一","家"]
          },
          {
            "byte_length":6,
            "byte_offset":10,
            "formal":"",
            "item":"高科技",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
        ]
      }
    ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **
PER	人名	LOC	地名	ORG	机构名	TIME	时间

2.3.2 词法分析（定制版）

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
public void LexerCustomDemo() {
var text = "百度是一家高科技公司";

// 调用词法分析（定制版），可能会抛出网络等异常，请使用try/catch捕获
var result = client.LexerCustom(text);
Console.WriteLine(result);
}
```

词法分析（定制版） 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析（定制版） 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县

参数名称	类型	**必需**	详细说明
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析（定制版） 返回示例

参考词法分析接口

2.3.3 依存句法分析

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

```
public void DepParserDemo() {
var text = "张飞";

// 调用依存句法分析，可能会抛出网络等异常，请使用try/catch捕获
var result = client.DepParser(text);
Console.WriteLine(result);
// 如果有可选参数
var options = new Dictionary<string, object>{
    {"mode", 1}
};
// 带参数调用依存句法分析
result = client.DepParser(text, options);
Console.WriteLine(result);
}
```

依存句法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过256字节

参数名称	是否必选	类型	说明
mode	否	string	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

依存句法分析 返回数据参数详情

参数名称	+类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	number	词的ID
word	string	词
postag	string	词性，请参照API文档中的**词性（postag)取值范围**
head	int	词的父节点ID
+deprel	string	词与父节点的依存关系，请参照API文档的**依存关系标识**

依存句法分析 返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
```

```
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
}
```

2.3.4 词向量表示

词向量表示接口提供中文词向量的查询功能。

```
public void WordEmbeddingDemo() {
var word = "张飞";

// 调用词向量表示，可能会抛出网络等异常，请使用try/catch捕获
var result = client.WordEmbedding(word);
Console.WriteLine(result);
}
```

词向量表示 请求参数详情

参数名称	是否必选	类型	说明
word	是	string	文本内容（GBK编码），最大64字节

词向量表示 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
word	string	查询词
vec	float	词向量结果表示

词向量表示 返回示例

```
{
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
    0.283711,
    -0.233656,
    0.555556
  ]
}
```

2.3.5 DNN语言模型

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

```
public void DnnlmCnDemo() {
var text = "床前明月光";

// 调用DNN语言模型，可能会抛出网络等异常，请使用try/catch捕获
var result = client.DnnlmCn(text);
Console.WriteLine(result);
}
```

DNN语言模型 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大512字节，不需要切词

DNN语言模型 返回数据参数详情

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺


```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```


DNN语言模型 返回示例

2.3.6 词义相似度

输入两个词，得到两个词的相似度结果。

```
public void WordSimEmbeddingDemo() {
var word1 = "北京";
```

```
var word2 = "上海";

// 调用词义相似度，可能会抛出网络等异常，请使用try/catch捕获
var result = client.WordSimEmbedding(word1, word2);
Console.WriteLine(result);
// 如果有可选参数
var options = new Dictionary<string, object>{
    {"mode", 0}
};
// 带参数调用词义相似度
result = client.WordSimEmbedding(word1, word2, options);
Console.WriteLine(result);
}
```

词义相似度 请求参数详情

参数名称	是否必选	类型	说明
word_1	是	string	词1（GBK编码），最大64字节
word_2	是	string	词1（GBK编码），最大64字节
mode	否	string	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

词义相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识码,随机数
score	number	相似度分数
words	array	输入的词列表
+word_1	string	输入的word1参数
+word_2	string	输入的word2参数

词义相似度 返回示例

```
{
    "score": 0.456862,
    "words": {
```

```
        "word_1": "北京",
        "word_2": "上海"
    }
}
```

2.3.7 短文本相似度

短文本相似度接口用来判断两个文本的相似度得分。

```
public void SimnetDemo() {
var text1 = "浙富股份";

var text2 = "万事通自考网";

// 调用短文本相似度，可能会抛出网络等异常，请使用try/catch捕获
var result = client.Simnet(text1, text2);
Console.WriteLine(result);
// 如果有可选参数
var options = new Dictionary<string, object>{
    {"model", "CNN"}
};
// 带参数调用短文本相似度
result = client.Simnet(text1, text2, options);
Console.WriteLine(result);
}
```

短文本相似度 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text_1	是	string		待比较文本1（GBK编码），最大512字节
text_2	是	string		待比较文本2（GBK编码），最大512字节
model	否	string	BOWCNNGRNN	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

短文本相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识
score	number	两个文本相似度得分
texts	array	输入文本
+text_1	string	第一个短文本
+text_2	string	第二个短文本

短文本相似度 返回示例

```
{
    "log_id": 12345,
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
    "score":0.3300237655639648 //相似度结果
},
```

2.3.8 评论观点抽取

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

```
public void CommentTagDemo() {
var text = "三星电脑电池不给力";

// 调用评论观点抽取，可能会抛出网络等异常，请使用try/catch捕获
var result = client.CommentTag(text);
Console.WriteLine(result);
// 如果有可选参数
var options = new Dictionary<string, object>{
    {"type", 13}
};
// 带参数调用评论观点抽取
result = client.CommentTag(text, options);
Console.WriteLine(result);
}
```

评论观点抽取 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text	是	string		评论内容（GBK编码），最大10240字节
type	否	string	1 - 酒店2 - KTV3 - 丽人4 - 美食餐饮5 - 旅游6 - 健康7 - 教育8 - 商业9 - 房产10 - 汽车11 - 生活12 - 购物13 - 3C	评论行业类型，默认为4（餐饮美食）

评论观点抽取 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

评论观点抽取 返回示例


```
{
    "items": [
        {
        "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
```

```
        "end_pos": 18,
        "abstract":"三星电脑<span>电池不给力</span>"
        }
    ]
}
```

2.3.9 情感倾向分析

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

```
public void SentimentClassifyDemo() {
var text = "苹果是一家伟大的公司";

// 调用情感倾向分析，可能会抛出网络等异常，请使用try/catch捕获
var result = client.SentimentClassify(text);
Console.WriteLine(result);
}
```

情感倾向分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大102400字节

情感倾向分析 返回数据参数详情

参数	是否必须	类型	说明
text	是	string	输入的文本内容
items	是	array	输入的词列表
+sentiment	是	number	表示情感极性分类结果, 0:负向，1:中性，2:正向
+confidence	是	number	表示分类的置信度
+positive_prob	是	number	表示属于积极类别的概率
+negative_prob	是	number	表示属于消极类别的概率

情感倾向分析 返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
        }
    ]
}
```

2.3.10 文章标签

文章标签服务能够针对网络各类媒体文章进行快速的内容理解，根据输入含有标题的文章，输出多个内容标签以及对应的置信度，用于个性化推荐、相似文章聚合、文本内容分析等场景。

```
public void KeywordDemo() {
var title = "iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下";

var content = "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。";

// 调用文章标签，可能会抛出网络等异常，请使用try/catch捕获
var result = client.Keyword(title, content);
Console.WriteLine(result);
}
```

文章标签 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章标签 返回数据参数详情

参数	是否必须	类型	说明
items	是	array(object)	关键词结果数组，每个元素对应抽取到的一个关键词
+tag	是	string	关注点字符串
+score	是	number	权重(取值范围0~1)

文章标签 返回示例


```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
        }
    ]
}
```


2.3.11 文章分类

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。

```
public void TopicDemo() {
var title = "欧洲冠军杯足球赛";

var content = "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。";

// 调用文章分类，可能会抛出网络等异常，请使用try/catch捕获
var result = client.Topic(title, content);
Console.WriteLine(result);
}
```

文章分类 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章分类 返回数据参数详情

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

文章分类 返回示例

```
{
    "log_id": 5710764909216517248,
    "item": {
        "lv2_tag_list": [
            {
                "score": 0.895467,
                "tag": "足球"
            },
```

```
            {
                "score": 0.794878,
                "tag": "国际足球"
            }
        ],
        "lv1_tag_list": [
            {
                "score": 0.88808,
                "tag": "体育"
            }
        ]
    }
}
```

2.3.12 FAQ

1. throw exception “fail to fetch token: 基础连接已关闭”

- 检查网络连接
- 检查网络是否有代理

2. SSL报错 “The authentication or decryption has failed” 可能由于网络代理等原因导致证书不正确，属于常见的网络问题，可以参考这个答案

2.4 错误信息

2.4.1 错误返回格式

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。
- error_msg：错误描述信息，帮助理解和解决发生的错误。

2.4.2 错误码

错误码	错误信息	描述
4	Open api request limit reached	集群超限额

错误码	错误信息	描述
14	IAM Certification failed	IAM鉴权失败，建议用户参照文档自查生成sign的方式是否正确，或换用控制台中ak sk的方式调用
17	Open api daily request limit reached	每天流量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额
100	Invalid parameter	无效参数
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂未收录，无法比对相似度

第3章　Java SDK文档

3.1　简介

Hi，您好，欢迎使用百度自然语言处理服务。

本文档主要针对Java开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流：http://ai.baidu.com/forum/topic/list/169

3.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

3.1.2　版本更新记录

上线日期	版本号	更新内容
2018.1.26	4.1.1	新增文本分类接口
2018.1.11	4.1.0	新增文本标签接口
2017.12.22	4.0.0	接口统一升级
2017.10.18	3.2.1	使用proxy问题修复
2017.8.25	3.0.0	更新sdk打包方式：所有AI服务集成一个SDK
2017.7.14	1.5.2	更新sdk打包方式
2017.6.30	1.5.1	新增句法依存接口
2017.6.15	1.5.0	短文本相似度接口升级
2017.5.25	1.4.0	词向量、评论观战、中文DNN接口升级，新增词相似度和情感分析接口
2017.4.20	1.3.3	新增词法分析接口，规范分词接口错误码
2017.4.13	1.3.2	AI SDK同步版本更新
2017.3.23	1.3	对安卓环境兼容问题进行修复
2017.3.2	1.2	增加设置超时接口
2017.1.20	1.1	对部分云用户调用不成功的错误修复
2017.1.6	1.0	初始版本，上线中文分词、词性标注、词向量表示、中文DNN语言模型、短文本相似度和评论观点抽取接口

3.2 快速入门

3.2.1 安装NLP Java SDK

NLP Java SDK目录结构

```
com.baidu.aip
        ├── auth                    //签名相关类
        ├── http                    //Http通信相关类
        ├── client                  //公用类
        ├── exception               //exception类
        ├── nlp
```

```
        |       └── AipNlp            //AipNlp类
        └── util                                  //工具类
```

支持 JAVA版本：1.7+

查看源码 Java SDK代码现已公开，您可以查看代码、或者在License范围内修改和编译SDK以适配您的环境。 github链接：https://github.com/Baidu-AIP/java-sdk

使用maven依赖：

添加以下依赖即可。其中版本号可在maven官网查询

```
<dependency>
    <groupId>com.baidu.aip</groupId>
    <artifactId>java-sdk</artifactId>
    <version>\${version}</version>
</dependency>
```

直接使用JAR包步骤如下：

1.在官方网站下载Java SDK压缩工具包。

2.将下载的aip-java-sdk-version.zip解压后，复制到工程文件夹中。

3.在Eclipse右键“工程 -> Properties -> Java Build Path -> Add JARs”。

4.添加SDK工具包aip-java-sdk-version.jar和第三方依赖工具包json-20160810.jar log4j-1.2.17.jar。

其中，version为版本号，添加完成后，用户就可以在工程中使用NLP Java SDK。

3.2.2 新建AipNlp

AipNlp是自然语言处理的Java客户端，为使用自然语言处理的开发人员提供了一系列的交互方法。

用户可以参考如下代码新建一个AipNlp,初始化完成后建议单例使用,避免重复获取access_token：

```
public class Sample {
    //设置APPID/AK/SK
    public static final String APP_ID = ”你的␣App␣ID”;
    public static final String API_KEY = ”你的␣Api␣Key”;
    public static final String SECRET_KEY = ”你的␣Secret␣Key”;

    public static void main(String[] args) {
        // 初始化一个AipNlp
```

```
        AipNlp client  = new AipNlp(APP_ID, API_KEY, SECRET_KEY);

        // 可选：设置网络连接参数
        client.setConnectionTimeoutInMillis(2000);
        client.setSocketTimeoutInMillis(60000);

        // 可选：设置代理服务器地址，和二选一，或者均不设置httpsocket
        client.setHttpProxy("proxy_host", proxy_port); // 设置代理http
        client.setSocketProxy("proxy_host", proxy_port); // 设置代理socket

        // 可选：设置日志输出格式，若不设置，则使用默认配置log4j
        // 也可以直接通过启动参数设置此环境变量jvm
        System.setProperty("aip.log4j.conf", "path/to/your/log4j.properties");

        // 调用接口
        String text = "百度是一家高科技公司";
        JSONObject res = client.lexer(text, null);
        System.out.println(res.toString(2));

    }
}
```

其中示例的log4j.properties文件内容如下：

```
## 可以设置级别：debug>info>error

## debug：显示debug、info、error

## info：显示info、error

## error：只error
log4j.rootLogger=debug,appender1

## log4j.rootLogger=info,appender1

## log4j.rootLogger=error,appender1


## 输出到控制台
log4j.appender.appender1=org.apache.log4j.ConsoleAppender

## 样式为TTCCLayout
log4j.appender.appender1.layout=org.apache.log4j.PatternLayout


## 自定义样式
```

```
## %r 时间 0

## %t 方法名 main

## %p 优先级 DEBUG/INFO/ERROR

## %c 所属类的全名(包括包名)

## %l 发生的位置，在某个类的某行

## %m 输出代码中指定的讯息，如log(message)中的message

## %n 输出一个换行

log4j.appender.appender1.layout.ConversionPattern=[%d{yy/MM/dd HH:mm:ss:SSS}][%t]
[%p] -%l %m%n
```

在上面代码中，常量APP_ID在百度云控制台中创建，常量API_KEY与SECRET_KEY是在创建完毕应用后，系统分配给用户的，均为字符串，用于标识用户，为访问做签名验证，可在AI服务控制台中的应用列表中查看。

注意：如您以前是百度云的老用户，其中API_KEY对应百度云的“Access Key ID”，SECRET_KEY对应百度云的“Access Key Secret”。

3.2.3 配置AipNlp

如果用户需要配置AipNlp的一些细节参数，可以在构造AipNlp之后调用接口设置参数，目前只支持以下参数：

接口	说明
setConnectionTi meoutInMillis	建立连接的超时时间（单位：毫秒）
setSocketTimeou tInMillis	通过打开的连接传输数据的超时时间（单位：毫秒）
setHttpProxy	设置http代理服务器
setSocketProxy	设置socket代理服务器（http和socket类型代理服务器只能二选一）

3.3 接口说明

3.3.1 词法分析

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
public void sample(AipNlp client) {
    String text = "百度是一家高科技公司";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();

    // 词法分析
    JSONObject res = client.lexer(text, options);
    System.out.println(res.toString(2));

}
```

词法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	String	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串

参数名称	类型	**必需**	详细说明
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析 返回示例

```
{
  "status":0,
  "version":"ver_1_0_1",
```

```
  "results":[
    {
      "retcode":0,
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["是"]
         },
         {
           "byte_length":4,
           "byte_offset":6,
           "formal":"",
           "item":"一家",
           "ne":"",
           "pos":"m",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["一","家"]
         },
         {
           "byte_length":6,
           "byte_offset":10,
           "formal":"",
           "item":"高科技",
           "ne":"",
           "pos":"n",
```

```
            "uri":"",
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
        ]
      }
    ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词 **	** 含义 **	** 缩略词 **	** 含义 **	** 缩略词 **	** 含义 **	** 缩略词 **	** 含义 **
PER	人名	LOC	地名	ORG	机构名	TIME	时间

3.3.2 词法分析（定制版）

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
public void sample(AipNlp client) {
    String text = "百度是一家高科技公司";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();

    // 词法分析（定制版）
    JSONObject res = client.lexerCustom(text, options);
    System.out.println(res.toString(2));

}
```

词法分析（定制版） 请求参数详情

参数名称	是否必选	类型	说明
text	是	String	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析（定制版） 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串

参数名称	类型	**必需**	详细说明
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析（定制版） 返回示例

参考词法分析接口

3.3.3 依存句法分析

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

```
public void sample(AipNlp client) {
    String text = "张飞";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();
    options.put("mode", 1);

    // 依存句法分析
    JSONObject res = client.depParser(text, options);
    System.out.println(res.toString(2));

}
```

依存句法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	String	待分析文本（目前仅支持GBK编码），长度不超过256字节
mode	否	String	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

依存句法分析 返回数据参数详情

参数名称	+类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	number	词的ID
word	string	词

参数名称	+类型	详细说明
postag	string	词性，请参照API文档中的**词性（postag)取值范围**
head	int	词的父节点ID
+deprel	string	词与父节点的依存关系，请参照API文档的**依存关系标识**

依存句法分析 返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
}
```

3.3.4 词向量表示

词向量表示接口提供中文词向量的查询功能。

```
public void sample(AipNlp client) {
    String word = "张飞";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();

    // 词向量表示
    JSONObject res = client.wordEmbedding(word, options);
    System.out.println(res.toString(2));

}
```

词向量表示 请求参数详情

参数名称	是否必选	类型	说明
word	是	String	文本内容（GBK编码），最大64字节

词向量表示 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
word	string	查询词
vec	float	词向量结果表示

词向量表示 返回示例

```
{
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
    0.283711,
```

```
    -0.233656,
    0.555556
  ]
}
```

3.3.5 DNN语言模型

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

```
public void sample(AipNlp client) {
    String text = "床前明月光";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();

    // DNN语言模型
    JSONObject res = client.dnnlmCn(text, options);
    System.out.println(res.toString(2));

}
```

DNN语言模型 请求参数详情

参数名称	是否必选	类型	说明
text	是	String	文本内容（GBK编码），最大512字节，不需要切词

DNN语言模型 返回数据参数详情

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺

```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```


DNN语言模型 返回示例

3.3.6 词义相似度

输入两个词，得到两个词的相似度结果。

```
public void sample(AipNlp client) {
    String word1 = "北京";
    String word2 = "上海";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();
    options.put("mode", 0);

    // 词义相似度
    JSONObject res = client.wordSimEmbedding(word1, word2, options);
    System.out.println(res.toString(2));

}
```

词义相似度 请求参数详情

参数名称	是否必选	类型	说明
word_1	是	String	词1（GBK编码），最大64字节
word_2	是	String	词1（GBK编码），最大64字节
mode	否	String	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

词义相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识码,随机数
score	number	相似度分数
words	array	输入的词列表
+word_1	string	输入的word1参数
+word_2	string	输入的word2参数

词义相似度 返回示例

```
{
    "score": 0.456862,
    "words": {
      "word_1": "北京",
      "word_2": "上海"
    }
}
```

3.3.7 短文本相似度

短文本相似度接口用来判断两个文本的相似度得分。

```
public void sample(AipNlp client) {
```

```
    String text1 = "浙富股份";
    String text2 = "万事通自考网";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();
    options.put("model", "CNN");

    // 短文本相似度
    JSONObject res = client.simnet(text1, text2, options);
    System.out.println(res.toString(2));

}
```

短文本相似度 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text_1	是	String		待比较文本1（GBK编码），最大512字节
text_2	是	String		待比较文本2（GBK编码），最大512字节
model	否	String	BOWCNNGRNN	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

短文本相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识
score	number	两个文本相似度得分
texts	array	输入文本
+text_1	string	第一个短文本
+text_2	string	第二个短文本

短文本相似度 返回示例

```
{
    "log_id": 12345,
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
    "score":0.3300237655639648 //相似度结果
},
```

3.3.8 评论观点抽取

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

```
public void sample(AipNlp client) {
    String text = "三星电脑电池不给力";

    // 获取美食评论情感属性
   JSONObject response = client.commentTag("这家餐馆味道不错", ESimnetType.FOOD, options);
    System.out.println(response.toString());

    // 获取酒店评论情感属性
    response = client.commentTag("喜来登酒店不错", ESimnetType.HOTEL, options);
    System.out.println(response.toString());

}
```

评论观点抽取 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text	是	String		评论内容（GBK编码），最大10240字节
type	否	String	1 - 酒店2 - KTV3 - 丽人4 - 美食餐饮5 - 旅游6 - 健康7 - 教育8 - 商业9 - 房产10 - 汽车11 - 生活12 - 购物13 - 3C	评论行业类型，默认为4（餐饮美食）

评论观点抽取 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

评论观点抽取 返回示例

```
{
    "items": [
        {
        "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
        "end_pos": 18,
        "abstract":"三星电脑<span>电池不给力</span>"
        }
    ]
}
```

3.3.9 情感倾向分析

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

```
public void sample(AipNlp client) {
    String text = "苹果是一家伟大的公司";
```

```
    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();

    // 情感倾向分析
    JSONObject res = client.sentimentClassify(text, options);
    System.out.println(res.toString(2));

}
```

情感倾向分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	String	文本内容（GBK编码），最大102400字节

情感倾向分析 返回数据参数详情

参数	是否必须	类型	说明
text	是	string	输入的文本内容
items	是	array	输入的词列表
+sentiment	是	number	表示情感极性分类结果, 0:负向，1:中性，2:正向
+confidence	是	number	表示分类的置信度
+positive_prob	是	number	表示属于积极类别的概率
+negative_prob	是	number	表示属于消极类别的概率

情感倾向分析 返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
```

```
        }
    ]
}
```

3.3.10 文章标签

文章标签服务能够针对网络各类媒体文章进行快速的内容理解，根据输入含有标题的文章，输出多个内容标签以及对应的置信度，用于个性化推荐、相似文章聚合、文本内容分析等场景。

```
public void sample(AipNlp client) {
    String title = "iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下";
    String content = "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();

    // 文章标签
    JSONObject res = client.keyword(title, content, options);
    System.out.println(res.toString(2));

}
```

文章标签 请求参数详情

参数名称	是否必选	类型	说明
title	是	String	篇章的标题，最大80字节
content	是	String	篇章的正文，最大65535字节

文章标签 返回数据参数详情

参数	是否必须	类型	说明
items	是	array(object)	关键词结果数组，每个元素对应抽取到的一个关键词
+tag	是	string	关注点字符串
+score	是	number	权重(取值范围0~1)

文章标签 返回示例

```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
        }
    ]
}
```

3.3.11 文章分类

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。

```
public void sample(AipNlp client) {
    String title = "欧洲冠军杯足球赛";
    String content = "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。";

    // 传入可选参数调用接口
    HashMap<String, Object> options = new HashMap<String, Object>();
```

```
    // 文章分类
    JSONObject res = client.topic(title, content, options);
    System.out.println(res.toString(2));

}
```

文章分类 请求参数详情

参数名称	是否必选	类型	说明
title	是	String	篇章的标题，最大80字节
content	是	String	篇章的正文，最大65535字节

文章分类 返回数据参数详情

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

文章分类 返回示例

```
{
    "log_id": 5710764909216517248,
    "item": {
        "lv2_tag_list": [
            {
                "score": 0.895467,
                "tag": "足球"
            },
            {
                "score": 0.794878,
                "tag": "国际足球"
            }
        ],
```

```
        "lv1_tag_list": [
            {
                "score": 0.88808,
                "tag": "体育"
            }
        ]
    }
}
```

3.4 错误信息

3.4.1 错误返回格式

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。
- error_msg：错误描述信息，帮助理解和解决发生的错误。

3.4.2 错误码

SDK本地检测参数返回的错误码：

error_code	error_msg	备注
SDK100	image size error	图片大小超限
SDK101	image length error	图片边长不符合要求
SDK102	read image file error	读取图片文件错误
SDK108	connection or read data time out	连接超时或读取数据超时
SDK109	unsupported image format	不支持的图片格式

服务端返回的错误码

错误码	错误信息	描述
4	Open api request limit reached	集群超限额
14	IAM Certification failed	IAM鉴权失败，建议用户参照文档自查生成sign的方式是否正确，或换用控制台中ak sk的方式调用

错误码	错误信息	描述
17	Open api daily request limit reached	每天流量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额
100	Invalid parameter	无效参数
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂未收录，无法比对相似度

第4章　Nodejs SDK文档

4.1　简介

Hi，您好，欢迎使用百度自然语言处理服务。

本文档主要针对Nodejs开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流：http://ai.baidu.com/forum/topic/list/169

4.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

4.1.2　版本更新记录

上线日期	版本号	更新内容
2018.1.26	2.1.1	新增文章分类接口
2018.1.12	2.1.0	新增文本标签接口
2017.12.21	2.0.0	实现代码重构，新增自定义词法分析接口，接口返回标准promise对象
2017.6.30	1.2.0	增加句法依存接口
2017.4.13	1.0.0	初版

4.2 快速入门

4.2.1 安装自然语言处理 Node SDK

自然语言处理 Node SDK目录结构

```
├── src
│   ├── auth                    //授权相关类
│   ├── http                    //Http通信相关类
│   ├── client                  //公用类
│   ├── util                    //工具类
│   └── const                   //常量类
├── AipNlp.js                   //自然语言处理交互类
├── index.js                    //入口文件
└── package.json                //npm包描述文件
```

支持 node 版本 4.0+

查看源码 Nodejs SDK代码已开源，您可以查看代码、或者在License范围内修改和编译SDK以适配您的环境。 github链接：https://github.com/Baidu-AIP/nodejs-sdk

直接使用node开发包步骤如下：

1.在官方网站下载node SDK压缩包。

2.将下载的aip-node-sdk-version.zip解压后，复制到工程文件夹中。

3.进入目录，运行npm install安装sdk依赖库

4.把目录当做模块依赖

其中，version为版本号，添加完成后，用户就可以在工程中使用自然语言处理 Node SDK。

直接使用npm安装依赖：

{\color{emcolor}\textbf{npm install baidu-aip-sdk}}

4.2.2 新建AipNlpClient

AipNlpClient是自然语言处理的node客户端，为使用自然语言处理的开发人员提供了一系列的交互方法。

用户可以参考如下代码新建一个AipNlpClient：

```
var AipNlpClient = require("baidu-aip-sdk").nlp;

// 设置APPID/AK/SK
var APP_ID = "你的 App ID";
var API_KEY = "你的 Api Key";
var SECRET_KEY = "你的 Secret Key";

// 新建一个对象，建议只保存一个对象调用服务接口
var client = new AipNlpClient(APP_ID, API_KEY, SECRET_KEY);
```

为了使开发者更灵活的控制请求，模块提供了设置全局参数和全局请求拦截器的方法；本库发送网络请求依赖的是request模块，因此参数格式与request模块的参数相同 更多参数细节您可以参考request官方参数文档。

```
var HttpClient = require("baidu-aip-sdk").HttpClient;

// 设置request库的一些参数，例如代理服务地址，超时时间等
// request参数请参考 https://github.com/request/request#requestoptions-callback
HttpClient.setRequestOptions({timeout: 5000});

// 也可以设置拦截每次请求（设置拦截后，调用的setRequestOptions设置的参数将不生效），
// 可以按需修改request参数（无论是否修改，必须返回函数调用参数）
// request参数请参考 https://github.com/request/request#requestoptions-callback
HttpClient.setRequestInterceptor(function(requestOptions) {
    // 查看参数
    console.log(requestOptions)
    // 修改参数
    requestOptions.timeout = 5000;
    // 返回参数
    return requestOptions;
});
```

在上面代码中，常量APP_ID在百度云控制台中创建，常量API_KEY与SECRET_KEY是在创建完毕应用后，系统分配给用户的，均为字符串，用于标识用户，为访问做签名验证，可在AI服务控制台中的应用列表中查看。

注意：如您以前是百度云的老用户，其中API_KEY对应百度云的“Access Key ID”，SECRET_KEY对应百度云的“Access Key Secret”。

4.3 接口说明

4.3.1 词法分析

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
var text = "百度是一家高科技公司";

// 调用词法分析
client.lexer(text).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

词法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词

参数名称	类型	**必需**	详细说明
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析 返回示例


```
{
  "status":0,
  "version":"ver_1_0_1",
  "results":[
    {
      "retcode":0,
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["是"]
         },
         {
           "byte_length":4,
           "byte_offset":6,
           "formal":"",
           "item":"一家",
           "ne":"",
           "pos":"m",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["一","家"]
         },
         {
```

```
            "byte_length":6,
            "byte_offset":10,
            "formal":"",
            "item":"高科技",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
        ]
      }
    ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词**	** 含义**	** 缩略词**	** 含义**	** 缩略词**	** 含义**	** 缩略词**	** 含义**
PER	人名	LOC	地名	ORG	机构名	TIME	时间

4.3.2 词法分析（定制版）

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
var text = "百度是一家高科技公司";

// 调用词法分析（定制版）
client.lexerCustom(text).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

词法分析（定制版） 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析（定制版） 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串

参数名称	类型	**必需**	详细说明
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析（定制版） 返回示例

参考词法分析接口

4.3.3 依存句法分析

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

```
var text = "张飞";

// 调用依存句法分析
client.depparser(text).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});

// 如果有可选参数
var options = {};
options["mode"] = "1";

// 带参数调用依存句法分析
client.depparser(text, options).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});;
```

依存句法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过256字节

参数名称	是否必选	类型	说明
mode	否	string	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

依存句法分析 返回数据参数详情

参数名称	+类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	number	词的ID
word	string	词
postag	string	词性，请参照API文档中的**词性（postag)取值范围**
head	int	词的父节点ID
+deprel	string	词与父节点的依存关系，请参照API文档的**依存关系标识**

依存句法分析 返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
```

```
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
}
```

4.3.4 词向量表示

词向量表示接口提供中文词向量的查询功能。

```
var word = "张飞";

// 调用词向量表示
client.wordembedding(word).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

词向量表示 请求参数详情

参数名称	是否必选	类型	说明
word	是	string	文本内容（GBK编码），最大64字节

词向量表示 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
word	string	查询词
vec	float	词向量结果表示

词向量表示 返回示例

```
{
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
    0.283711,
    -0.233656,
    0.555556
  ]
}
```

4.3.5 DNN语言模型

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

```
var text = "床前明月光";

// 调用DNN语言模型
client.dnnlmCn(text).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

DNN语言模型 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大512字节，不需要切词

DNN语言模型 返回数据参数详情

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺

```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```

DNN语言模型 返回示例

4.3.6 词义相似度

输入两个词，得到两个词的相似度结果。

```
var word1 = "北京";

var word2 = "上海";

// 调用词义相似度
client.wordSimEmbedding(word1, word2).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});

// 如果有可选参数
var options = {};
options["mode"] = "0";

// 带参数调用词义相似度
client.wordSimEmbedding(word1, word2, options).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});;
```

词义相似度 请求参数详情

参数名称	是否必选	类型	说明
word_1	是	string	词1（GBK编码），最大64字节
word_2	是	string	词1（GBK编码），最大64字节
mode	否	string	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

词义相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识码,随机数
score	number	相似度分数
words	array	输入的词列表
+word_1	string	输入的word1参数
+word_2	string	输入的word2参数

词义相似度 返回示例

```
{
    "score": 0.456862,
    "words": {
      "word_1": "北京",
      "word_2": "上海"
    }
}
```

4.3.7 短文本相似度

短文本相似度接口用来判断两个文本的相似度得分。

```
var text1 = "浙富股份";

var text2 = "万事通自考网";

// 调用短文本相似度
client.simnet(text1, text2).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});

// 如果有可选参数
var options = {};
options["model"] = "CNN";

// 带参数调用短文本相似度
client.simnet(text1, text2, options).then(function(result) {
```

```
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});;
```

短文本相似度 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text_1	是	string		待比较文本1（GBK编码），最大512字节
text_2	是	string		待比较文本2（GBK编码），最大512字节
model	否	string	BOWCNNGRNN	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

短文本相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识
score	number	两个文本相似度得分
texts	array	输入文本
+text_1	string	第一个短文本
+text_2	string	第二个短文本

短文本相似度 返回示例

```
{
    "log_id": 12345,
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
```

```
    "score":0.3300237655639648 //相似度结果
},
```

4.3.8 评论观点抽取

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

```
var text = "三星电脑电池不给力";

// 调用评论观点抽取
client.commentTag(text).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});

// 如果有可选参数
var options = {};
options["type"] = "13";

// 带参数调用评论观点抽取
client.commentTag(text, options).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});;
```

评论观点抽取 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text	是	string		评论内容（GBK编码），最大10240字节

参数名称	是否必选	类型	可选值范围	说明
type	否	string	1 - 酒店2 - KTV3 - 丽人4 - 美食餐饮5 - 旅游6 - 健康7 - 教育8 - 商业9 - 房产10 - 汽车11 - 生活12 - 购物13 - 3C	评论行业类型，默认为4（餐饮美食）

评论观点抽取 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

评论观点抽取 返回示例


```
{
    "items": [
        {
        "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
        "end_pos": 18,
        "abstract":"三星电脑<span>电池不给力</span>"
        }
    ]
}
```

4.3.9 情感倾向分析

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

```
var text = "苹果是一家伟大的公司";

// 调用情感倾向分析
client.sentimentClassify(text).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

情感倾向分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大102400字节

情感倾向分析 返回数据参数详情

参数	是否必须	类型	说明
text	是	string	输入的文本内容
items	是	array	输入的词列表
+sentiment	是	number	表示情感极性分类结果, 0:负向，1:中性，2:正向
+confidence	是	number	表示分类的置信度
+positive_prob	是	number	表示属于积极类别的概率
+negative_prob	是	number	表示属于消极类别的概率

情感倾向分析 返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
        }
    ]
}
```

4.3.10 文章标签

文章标签服务能够针对网络各类媒体文章进行快速的内容理解，根据输入含有标题的文章，输出多个内容标签以及对应的置信度，用于个性化推荐、相似文章聚合、文本内容分析等场景。

```
var title = "iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下";

var content = "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。";

// 调用文章标签
client.keyword(title, content).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

文章标签 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章标签 返回数据参数详情

参数	是否必须	类型	说明
items	是	array(object)	关键词结果数组，每个元素对应抽取到的一个关键词
+tag	是	string	关注点字符串
+score	是	number	权重(取值范围0~1)

文章标签 返回示例

```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
        }
    ]
}
```

4.3.11 文章分类

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。

```
var title = "欧洲冠军杯足球赛";

var content = "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。";

// 调用文章分类
client.topic(title, content).then(function(result) {
    console.log(JSON.stringify(result));
}).catch(function(err) {
    // 如果发生网络错误
    console.log(err);
});
```

文章分类 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章分类 返回数据参数详情

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

文章分类 返回示例

```
{
    "log_id": 5710764909216517248,
    "item": {
        "lv2_tag_list": [
            {
```

```
                "score": 0.895467,
                "tag": "足球"
            },
            {
                "score": 0.794878,
                "tag": "国际足球"
            }
        ],
        "lv1_tag_list": [
            {
                "score": 0.88808,
                "tag": "体育"
            }
        ]
    }
}
```


4.4 错误信息

4.4.1 错误返回格式

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。
- error_msg：错误描述信息，帮助理解和解决发生的错误。

4.4.2 错误码

错误码	错误信息	描述
4	Open api request limit reached	集群超限额
14	IAM Certification failed	IAM鉴权失败，建议用户参照文档自查生成sign的方式是否正确，或换用控制台中ak sk的方式调用
17	Open api daily request limit reached	每天流量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额

错误码	错误信息	描述
100	Invalid parameter	无效参数
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂未收录，无法比对相似度

第5章　PHP SDK文档

5.1　简介

Hi，您好，欢迎使用百度自然语言处理服务。

本文档主要针对PHP开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流：http://ai.baidu.com/forum/topic/list/169

5.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

5.1.2　版本更新记录

上线日期	版本号	更新内容
2018.01.25	2.2.0	新增文本标签API
2017.12.22	2.0.0	SDK代码重构
2017.5.11	1.0.0	自然语言处理服务上线

5.2 快速入门

5.2.1 安装自然语言处理 PHP SDK

自然语言处理 PHP SDK目录结构

```
├── AipNlp.php              //自然语言处理
└── lib
    ├── AipHttpClient.php       //内部http请求类
    ├── AipBCEUtil.php          //内部工具类
    └── AipBase                 //Aip基类
```

支持PHP版本：5.3+

使用PHP SDK开发骤如下：

1.在官方网站下载php SDK压缩包。

2.将下载的aip-php-sdk-version.zip解压后，复制AipNlp.php以及lib/*到工程文件夹中。

3.引入AipNlp.php

5.2.2 新建AipNlp

AipNlp是自然语言处理的PHP SDK客户端，为使用自然语言处理的开发人员提供了一系列的交互方法。

参考如下代码新建一个AipNlp：

```
require_once 'AipNlp.php';

// 你的 APPID AK SK
const APP_ID = '你的 App ID';
const API_KEY = '你的 Api Key';
const SECRET_KEY = '你的 Secret Key';
```

```
\$client = new AipNlp(APP_ID, API_KEY, SECRET_KEY);
```

在上面代码中，常量APP_ID在百度云控制台中创建，常量API_KEY与SECRET_KEY是在创建完毕应用后，系统分配给用户的，均为字符串，用于标识用户，为访问做签名验证，可在AI服务控制台中的应用列表中查看。

注意：如您以前是百度云的老用户，其中API_KEY对应百度云的“Access Key ID”，SECRET_KEY对应百度云的“Access Key Secret”。

5.2.3 配置AipNlp

如果用户需要配置AipNlp的网络请求参数(一般不需要配置)，可以在构造AipNlp之后调用接口设置参数，目前只支持以下参数：

接口	说明
setConnectionTi meoutInMillis	建立连接的超时时间（单位：毫秒)
setSocketTimeou tInMillis	通过打开的连接传输数据的超时时间（单位：毫秒）

5.3 接口说明

5.3.1 词法分析

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
\$text = "百度是一家高科技公司";

// 调用词法分析
\$client->lexer(\$text);
```

词法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分

参数名称	类型	**必需**	详细说明
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析 返回示例


```
{
  "status":0,
  "version":"ver_1_0_1",
  "results":[
    {
      "retcode":0,
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
```

```
            "uri":"",
            "loc_details":[ ],
            "basic_words":["是"]
          },
          {
            "byte_length":4,
            "byte_offset":6,
            "formal":"",
            "item":"一家",
            "ne":"",
            "pos":"m",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["一","家"]
          },
          {
            "byte_length":6,
            "byte_offset":10,
            "formal":"",
            "item":"高科技",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
        ]
      }
    ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **
PER	人名	LOC	地名	ORG	机构名	TIME	时间

5.3.2 词法分析（定制版）

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
\$text = "百度是一家高科技公司";

// 调用词法分析（定制版）
\$client->lexerCustom(\$text);
```

词法分析（定制版） 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析（定制版） 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县

参数名称	类型	**必需**	详细说明
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析（定制版） 返回示例

参考词法分析接口

5.3.3 依存句法分析

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

```
\$text = "张飞";

// 调用依存句法分析
\$client->depParser(\$text);

// 如果有可选参数
\$options = array();
\$options["mode"] = 1;

// 带参数调用依存句法分析
\$client->depParser(\$text, \$options);
```

依存句法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过256字节

参数名称	是否必选	类型	说明
mode	否	string	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

依存句法分析 返回数据参数详情

参数名称	+类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	number	词的ID
word	string	词
postag	string	词性，请参照API文档中的**词性（postag)取值范围**
head	int	词的父节点ID
+deprel	string	词与父节点的依存关系，请参照API文档的**依存关系标识**

依存句法分析 返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
```

```
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
}
```

5.3.4 词向量表示

词向量表示接口提供中文词向量的查询功能。

```
\$word = "张飞";

// 调用词向量表示
\$client->wordEmbedding(\$word);
```

词向量表示 请求参数详情

参数名称	是否必选	类型	说明
word	是	string	文本内容（GBK编码），最大64字节

词向量表示 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
word	string	查询词
vec	float	词向量结果表示

词向量表示 返回示例

```
{
```

```
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
    0.283711,
    -0.233656,
    0.555556
  ]
}
```

5.3.5 DNN语言模型

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

```
\$text = "床前明月光";

// 调用DNN语言模型
\$client->dnnlm(\$text);
```

DNN语言模型 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大512字节，不需要切词

DNN语言模型 返回数据参数详情

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果

参数	类型	说明
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺


```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```


DNN语言模型 返回示例

5.3.6 词义相似度

输入两个词，得到两个词的相似度结果。

```
\$word1 = "北京";

\$word2 = "上海";

// 调用词义相似度
\$client->wordSimEmbedding(\$word1, \$word2);
```

```
// 如果有可选参数
\$options = array();
\$options["mode"] = 0;

// 带参数调用词义相似度
\$client->wordSimEmbedding(\$word1, \$word2, \$options);
```

词义相似度 请求参数详情

参数名称	是否必选	类型	说明
word_1	是	string	词1（GBK编码），最大64字节
word_2	是	string	词1（GBK编码），最大64字节
mode	否	string	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

词义相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识码,随机数
score	number	相似度分数
words	array	输入的词列表
+word_1	string	输入的word1参数
+word_2	string	输入的word2参数

词义相似度 返回示例

```
{
    "score": 0.456862,
    "words": {
      "word_1": "北京",
      "word_2": "上海"
    }
}
```

5.3.7 短文本相似度

短文本相似度接口用来判断两个文本的相似度得分。

```
\$text1 = "浙富股份";

\$text2 = "万事通自考网";

// 调用短文本相似度
\$client->simnet(\$text1, \$text2);

// 如果有可选参数
\$options = array();
\$options["model"] = "CNN";

// 带参数调用短文本相似度
\$client->simnet(\$text1, \$text2, \$options);
```

短文本相似度 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text_1	是	string		待比较文本1（GBK编码），最大512字节
text_2	是	string		待比较文本2（GBK编码），最大512字节
model	否	string	BOWCNNGRNN	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

短文本相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识
score	number	两个文本相似度得分
texts	array	输入文本

参数	类型	描述
+text_1	string	第一个短文本
+text_2	string	第二个短文本

短文本相似度 返回示例

```
{
    "log_id": 12345,
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
    "score":0.3300237655639648 //相似度结果
},
```

5.3.8 评论观点抽取

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

```
\$text = "三星电脑电池不给力";

// 调用评论观点抽取
\$client->commentTag(\$text);

// 如果有可选参数
\$options = array();
\$options["type"] = 13;

// 带参数调用评论观点抽取
\$client->commentTag(\$text, \$options);
```

评论观点抽取 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text	是	string		评论内容（GBK编码），最大10240字节
type	否	string	1 - 酒店2 - KTV3 - 丽人4 - 美食餐饮5 - 旅游6 - 健康7 - 教育8 - 商业9 - 房产10 - 汽车11 - 生活12 - 购物13 - 3C	评论行业类型，默认为4（餐饮美食）

评论观点抽取 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

评论观点抽取 返回示例


```
{
    "items": [
        {
        "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
        "end_pos": 18,
        "abstract":"三星电脑<span>电池不给力</span>"
```

```
        }
    ]
}
```

5.3.9 情感倾向分析

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

```
\$text = "苹果是一家伟大的公司";

// 调用情感倾向分析
\$client->sentimentClassify(\$text);
```

情感倾向分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大102400字节

情感倾向分析 返回数据参数详情

参数	是否必须	类型	说明
text	是	string	输入的文本内容
items	是	array	输入的词列表
+sentiment	是	number	表示情感极性分类结果, 0:负向，1:中性，2:正向
+confidence	是	number	表示分类的置信度
+positive_prob	是	number	表示属于积极类别的概率
+negative_prob	是	number	表示属于消极类别的概率

情感倾向分析 返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
        }
    ]
}
```

5.3.10 文章标签

文章标签服务能够针对网络各类媒体文章进行快速的内容理解，根据输入含有标题的文章，输出多个内容标签以及对应的置信度，用于个性化推荐、相似文章聚合、文本内容分析等场景。

```
\$title = "iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下";

\$content = "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。";

// 调用文章标签
\$client->keyword(\$title, \$content);
```

文章标签 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章标签 返回数据参数详情

参数	是否必须	类型	说明
items	是	array(object)	关键词结果数组，每个元素对应抽取到的一个关键词

参数	是否必须	类型	说明
+tag	是	string	关注点字符串
+score	是	number	权重(取值范围0~1)

文章标签 返回示例

```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
        }
    ]
}
```

5.3.11 文章分类

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。

```
\$title = "欧洲冠军杯足球赛";

\$content = "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和
```

水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。";

```
// 调用文章分类
\$client->topic(\$title, \$content);
```

文章分类 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章分类 返回数据参数详情

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

文章分类 返回示例


```
{
    "log_id": 5710764909216517248,
    "item": {
        "lv2_tag_list": [
            {
                "score": 0.895467,
                "tag": "足球"
            },
            {
                "score": 0.794878,
                "tag": "国际足球"
            }
        ],
```

```
        "lv1_tag_list": [
            {
                "score": 0.88808,
                "tag": "体育"
            }
        ]
    }
}
```

5.4 错误信息

5.4.1 错误返回格式

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。
- error_msg：错误描述信息，帮助理解和解决发生的错误。

5.4.2 错误码

错误码	错误信息	描述
4	Open api request limit reached	集群超限额
14	IAM Certification failed	IAM鉴权失败，建议用户参照文档自查生成sign的方式是否正确，或换用控制台中ak sk的方式调用
17	Open api daily request limit reached	每天流量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额
100	Invalid parameter	无效参数
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期

错误码	错误信息	描述
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂未收录，无法比对相似度

第6章　Python SDK文档

6.1　简介

Hi，您好，欢迎使用百度自然语言处理服务。

本文档主要针对Python开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流：http://ai.baidu.com/forum/topic/list/169

6.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

6.1.2　版本更新记录

上线日期	版本号	更新内容
2018.01.25	2.2.0	新增文本标签API
2017.12.22	2.0.0	SDK代码重构
2017.5.11	1.0.0	自然语言处理服务上线

6.2 快速入门

6.2.1 安装自然语言处理 Python SDK

自然语言处理 Python SDK目录结构

```
├── README.md
├── aip                    //SDK目录
│   ├── __init__.py        //导出类
│   ├── base.py            //aip基类
│   ├── http.py            //http请求
│   └── nlp.py //自然语言处理
└── setup.py               //setuptools安装
```

支持Python版本：2.7.+ ,3.+

安装使用Python SDK有如下方式：

- 如果已安装pip，执行`pip install baidu-aip`即可。
- 如果已安装setuptools，执行`python setup.py install`即可。

6.2.2 新建AipNlp

AipNlp是自然语言处理的Python SDK客户端，为使用自然语言处理的开发人员提供了一系列的交互方法。

参考如下代码新建一个AipNlp：

```
from aip import AipNlp

""" 你的 APPID AK SK """
APP_ID = '你的 App ID'
API_KEY = '你的 Api Key'
SECRET_KEY = '你的 Secret Key'
```

```
client = AipNlp(APP_ID, API_KEY, SECRET_KEY)
```

在上面代码中，常量APP_ID在百度云控制台中创建，常量API_KEY与SECRET_KEY是在创建完毕应用后，系统分配给用户的，均为字符串，用于标识用户，为访问做签名验证，可在AI服务控制台中的应用列表中查看。

注意：如您以前是百度云的老用户，其中API_KEY对应百度云的“Access Key ID”，SECRET_KEY对应百度云的“Access Key Secret”。

6.2.3 配置AipNlp

如果用户需要配置AipNlp的网络请求参数(一般不需要配置)，可以在构造AipNlp之后调用接口设置参数，目前只支持以下参数：

接口	说明
setConnectionTi meoutInMillis	建立连接的超时时间（单位：毫秒
setSocketTimeou tInMillis	通过打开的连接传输数据的超时时间（单位：毫秒）

6.3 接口说明

6.3.1 词法分析

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
text = "百度是一家高科技公司"

""" 调用词法分析 """
client.lexer(text);
```

词法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县

参数名称	类型	**必需**	详细说明
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析 返回示例


```
{
  "status":0,
  "version":"ver_1_0_1",
  "results":[
    {
      "retcode":0,
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["是"]
         },
         {
           "byte_length":4,
           "byte_offset":6,
           "formal":"",
```

```
            "item":"一家",
            "ne":"",
            "pos":"m",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["一","家"]
          },
          {
            "byte_length":6,
            "byte_offset":10,
            "formal":"",
            "item":"高科技",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
        ]
      }
    ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **
PER	人名	LOC	地名	ORG	机构名	TIME	时间

6.3.2 词法分析（定制版）

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
text = "百度是一家高科技公司"

""" 调用词法分析（定制版） """
client.lexerCustom(text);
```

词法分析（定制版） 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过65536字节

词法分析（定制版） 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词

参数名称	类型	**必需**	详细说明
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用GBK编码）
+byte_length	int	是	字节级length（使用GBK编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用GBK编码）
++byte_length	int	是	字节级length（使用GBK编码）

词法分析（定制版） 返回示例

参考词法分析接口

6.3.3 依存句法分析

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

```
text = "张飞"

""" 调用依存句法分析 """
client.depParser(text);

""" 如果有可选参数 """
options = {}
options["mode"] = 1

""" 带参数调用依存句法分析 """
client.depParser(text, options)
```

依存句法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	待分析文本（目前仅支持GBK编码），长度不超过256字节
mode	否	string	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

依存句法分析 返回数据参数详情

参数名称	+类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	number	词的ID
word	string	词
postag	string	词性，请参照API文档中的**词性（postag)取值范围**
head	int	词的父节点ID
+deprel	string	词与父节点的依存关系，请参照API文档的**依存关系标识**

依存句法分析 返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
```

```
}
```

6.3.4 词向量表示

词向量表示接口提供中文词向量的查询功能。

```
word = "张飞"

""" 调用词向量表示 """
client.wordEmbedding(word);
```

词向量表示 请求参数详情

参数名称	是否必选	类型	说明
word	是	string	文本内容（GBK编码），最大64字节

词向量表示 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
word	string	查询词
vec	float	词向量结果表示

词向量表示 返回示例

```
{
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
```

```
    0.283711,
    -0.233656,
    0.555556
  ]
}
```

6.3.5 DNN语言模型

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

```
text = "床前明月光"

""" 调用DNN语言模型 """
client.dnnlm(text);
```

DNN语言模型 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大512字节，不需要切词

DNN语言模型 返回数据参数详情

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺

```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
```

```
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```


DNN语言模型 返回示例

6.3.6 词义相似度

输入两个词，得到两个词的相似度结果。

```
word1 = "北京"

word2 = "上海"

""" 调用词义相似度 """
client.wordSimEmbedding(word1, word2);

""" 如果有可选参数 """
options = {}
options["mode"] = 0

""" 带参数调用词义相似度 """
client.wordSimEmbedding(word1, word2, options)
```

词义相似度 请求参数详情

参数名称	是否必选	类型	说明
word_1	是	string	词1（GBK编码），最大64字节
word_2	是	string	词1（GBK编码），最大64字节
mode	否	string	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

词义相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识码,随机数
score	number	相似度分数
words	array	输入的词列表
+word_1	string	输入的word1参数
+word_2	string	输入的word2参数

词义相似度 返回示例

```
{
    "score": 0.456862,
    "words": {
      "word_1": "北京",
      "word_2": "上海"
    }
}
```

6.3.7 短文本相似度

短文本相似度接口用来判断两个文本的相似度得分。

```
text1 = "浙富股份"

text2 = "万事通自考网"
```

```
""" 调用短文本相似度 """
client.simnet(text1, text2);

""" 如果有可选参数 """
options = {}
options["model"] = "CNN"

""" 带参数调用短文本相似度 """
client.simnet(text1, text2, options)
```

短文本相似度 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text_1	是	string		待比较文本1（GBK编码），最大512字节
text_2	是	string		待比较文本2（GBK编码），最大512字节
model	否	string	BOWCNNGRNN	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

短文本相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识
score	number	两个文本相似度得分
texts	array	输入文本
+text_1	string	第一个短文本
+text_2	string	第二个短文本

短文本相似度 返回示例

```
{
    "log_id": 12345,
```

```
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
    "score":0.3300237655639648 //相似度结果
},
```

6.3.8 评论观点抽取

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

```
text = "三星电脑电池不给力"

""" 调用评论观点抽取 """
client.commentTag(text);

""" 如果有可选参数 """
options = {}
options["type"] = 13

""" 带参数调用评论观点抽取 """
client.commentTag(text, options)
```

评论观点抽取 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text	是	string		评论内容（GBK编码），最大10240字节
type	否	string	1 - 酒店2 - KTV3 - 丽人4 - 美食餐饮5 - 旅游6 - 健康7 - 教育8 - 商业9 - 房产10 - 汽车11 - 生活12 - 购物13 - 3C	评论行业类型，默认为4（餐饮美食）

评论观点抽取 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

评论观点抽取 返回示例

```
{
    "items": [
        {
        "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
        "end_pos": 18,
        "abstract":"三星电脑<span>电池不给力</span>"
        }
    ]
}
```

6.3.9 情感倾向分析

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

```
text = "苹果是一家伟大的公司"

""" 调用情感倾向分析 """
client.sentimentClassify(text);
```

情感倾向分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	string	文本内容（GBK编码），最大102400字节

情感倾向分析 返回数据参数详情

参数	是否必须	类型	说明
text	是	string	输入的文本内容
items	是	array	输入的词列表
+sentiment	是	number	表示情感极性分类结果, 0:负向，1:中性，2:正向
+confidence	是	number	表示分类的置信度
+positive_prob	是	number	表示属于积极类别的概率
+negative_prob	是	number	表示属于消极类别的概率

情感倾向分析 返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
        }
    ]
}
```

6.3.10 文章标签

文章标签服务能够针对网络各类媒体文章进行快速的内容理解，根据输入含有标题的文章，输出多个内容标签以及对应的置信度，用于个性化推荐、相似文章聚合、文本内容分析等场景。

```
title = "iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下"

content = "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。"

""" 调用文章标签 """
client.keyword(title, content);
```

文章标签 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章标签 返回数据参数详情

参数	是否必须	类型	说明
items	是	array(object)	关键词结果数组，每个元素对应抽取到的一个关键词
+tag	是	string	关注点字符串
+score	是	number	权重(取值范围0~1)

文章标签 返回示例

```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
```

```
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
        }
    ]
}
```

6.3.11 文章分类

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。

```
title = "欧洲冠军杯足球赛"

content = "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。"

""" 调用文章分类 """
client.topic(title, content);
```

文章分类 请求参数详情

参数名称	是否必选	类型	说明
title	是	string	篇章的标题，最大80字节
content	是	string	篇章的正文，最大65535字节

文章分类 返回数据参数详情

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

文章分类 返回示例

```
{
    "log_id": 5710764909216517248,
    "item": {
        "lv2_tag_list": [
            {
                "score": 0.895467,
                "tag": "足球"
            },
            {
                "score": 0.794878,
                "tag": "国际足球"
            }
        ],
        "lv1_tag_list": [
            {
                "score": 0.88808,
                "tag": "体育"
            }
        ]
    }
}
```

6.4 错误信息

6.4.1 错误返回格式

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。

- `error_msg`：错误描述信息，帮助理解和解决发生的错误。

6.4.2 错误码

错误码	错误信息	描述
4	Open api request limit reached	集群超限额
14	IAM Certification failed	IAM鉴权失败，建议用户参照文档自查生成sign的方式是否正确，或换用控制台中ak sk的方式调用
17	Open api daily request limit reached	每天流量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额
100	Invalid parameter	无效参数
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中

错误码	错误信息	描述
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂 未 收录，无法比对相似度

第7章　C++ SDK文档

7.1　简介

Hi，您好，欢迎使用百度自然语言处理服务。

本文档主要针对C++开发者，描述百度自然语言处理接口服务的相关技术内容。如果您对文档内容有任何疑问，可以通过以下几种方式联系我们：

- 在百度云控制台内提交工单，咨询问题类型请选择人工智能服务；
- 如有疑问，进入AI社区交流：http://ai.baidu.com/forum/topic/list/169

7.1.1　接口能力

接口名称	接口能力简要描述
词法分析	分词、词性标注、专名识别
依存句法分析	自动分析文本中的依存句法结构信息
词向量表示	查询词汇的词向量，实现文本的可计算
DNN语言模型	判断一句话是否符合语言表达习惯，输出分词结果并给出每个词在句子中的概率值
词义相似度	计算两个给定词语的语义相似度
短文本相似度	判断两个文本的相似度得分
评论观点抽取	提取一个句子观点评论的情感属性
情感倾向分析	对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度
中文分词	切分出连续文本中的基本词汇序列（已合并到词法分析接口）
词性标注	为自然语言文本中的每个词汇赋予词性（已合并到词法分析接口）

7.1.2　版本更新记录

上线日期	版本号	更新内容
2018.1.26	0.5.1	新增文章分类接口
2018.1.12	0.5.0	新增文本标签接口
2017.12.21	0.4.0	更改了词向量表示接口和依存句法分析接口名称,新增了自定义词法分析接口
2017.11.24	0.3.2	修复windows平台VC环境的编译错误
2017.11.9	0.3.0	初始化参数修改
2017.10.31	0.1.0	自然语言处理第一版

7.2 快速入门

7.2.1 安装自然语言处理 C++ SDK

自然语言处理 C++ SDK目录结构

```
├── base
│   ├── base.h                    // 请求客户端基类
│   ├── base64.h                  // base64加密相关类
│   ├── http.h                    // http请求封装类
│   └── utils.h                   // 工具类
└── nlp.h                       // 自然语言处理 交互类
```

最低支持 C++ 11+

直接使用开发包步骤如下：

1.在官方网站下载C++ SDK压缩包。

2.将下载的`aip-cpp-sdk-version.zip`解压, 其中文件为包含实现代码的头文件。

3.安装依赖库libcurl（需要支持https） openssl jsoncpp(>1.6.2版本，0.x版本将不被支持)。

4.编译工程时添加 C++11 支持 (gcc/clang 添加编译参数 -std=c++11), 添加第三方库链接参数 lcurl, lcrypto, ljsoncpp。

5.在源码中include nlp.h ，引入压缩包中的头文件以使用aip命名空间下的类和方法。

7.2.2 新建client

client是自然语言处理的C++客户端，为使用自然语言处理的开发人员提供了一系列的交互方法。当您引入了相应头文件后就可以新建一个client对象

用户可以参考如下代码新建一个client：

```
#include "nlp.h"

// 设置APPID/AK/SK
std::string app_id = "你的 App ID";
std::string api_key = "你的 Api key";
std::string secret_key = "你的 Secret Key";

aip::Nlp client(app_id, api_key, secret_key);
```

nlp 在上面代码中，常量APP_ID在百度云控制台中创建，常量API_KEY与SECRET_KEY是在创建完毕应用后，系统分配给用户的，均为字符串，用于标识用户，为访问做签名验证，可在AI服务控制台中的应用列表中查看。

注意：如您以前是百度云的老用户，其中API_KEY对应百度云的“Access Key ID”，SECRET_KEY对应百度云的“Access Key Secret”。

7.3 接口说明

7.3.1 词法分析

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
Json::Value result;

std::string text = "百度是一家高科技公司";

// 调用词法分析
result = client.lexer(text, aip::null);
```

词法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	std::string	待分析文本（目前仅支持UTF8编码），长度不超过65536字节

词法分析 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用UTF8编码）
+byte_length	int	是	字节级length（使用UTF8编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分

参数名称	类型	**必需**	详细说明
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县
++byte_offset	int	是	在item中的字节级offset（使用UTF8编码）
++byte_length	int	是	字节级length（使用UTF8编码）

词法分析 返回示例


```
{
  "status":0,
  "version":"ver_1_0_1",
  "results":[
    {
      "retcode":0,
      "text":"百度是一家高科技公司",
      "items":[
         {
           "byte_length":4,
           "byte_offset":0,
           "formal":"",
           "item":"百度",
           "ne":"ORG",
           "pos":"",
           "uri":"",
           "loc_details":[ ],
           "basic_words":["百度"]
         },
         {
           "byte_length":2,
           "byte_offset":4,
           "formal":"",
           "item":"是",
           "ne":"",
           "pos":"v",
```

```
            "uri":"",
            "loc_details":[ ],
            "basic_words":["是"]
          },
          {
            "byte_length":4,
            "byte_offset":6,
            "formal":"",
            "item":"一家",
            "ne":"",
            "pos":"m",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["一","家"]
          },
          {
            "byte_length":6,
            "byte_offset":10,
            "formal":"",
            "item":"高科技",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["高","科技"]
          },
          {
            "byte_length":4,
            "byte_offset":16,
            "formal":"",
            "item":"公司",
            "ne":"",
            "pos":"n",
            "uri":"",
            "loc_details":[ ],
            "basic_words":["公司"]
          }
        ]
      }
    ]
}
```


词性缩略说明

** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **	** 词性 **	** 含义 **
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

专名识别缩略词含义

** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **	** 缩略词**	** 含义 **
PER	人名	LOC	地名	ORG	机构名	TIME	时间

7.3.2 词法分析（定制版）

词法分析接口向用户提供分词、词性标注、专名识别三大功能；能够识别出文本串中的基本词汇（分词），对这些词汇进行重组、标注组合后词汇的词性，并进一步识别出命名实体。

```
Json::Value result;

std::string text = "百度是一家高科技公司";

// 调用词法分析（定制版）
result = client.lexer_custom(text, aip::null);
```

词法分析（定制版） 请求参数详情

参数名称	是否必选	类型	说明
text	是	std::string	待分析文本（目前仅支持UTF8编码），长度不超过65536字节

词法分析（定制版） 返回数据参数详情

参数名称	类型	**必需**	详细说明
text	string	是	原始单条请求文本
items	array(object)	是	词汇数组，每个元素对应结果中的一个词
+item	string	是	词汇的字符串
+ne	string	是	命名实体类型，命名实体识别算法使用。词性标注算法中，此项为空串
+pos	string	是	词性，词性标注算法使用。命名实体识别算法中，此项为空串
+byte_offset	int	是	在text中的字节级offset（使用UTF8编码）
+byte_length	int	是	字节级length（使用UTF8编码）
+uri	string	否	链指到知识库的URI，只对命名实体有效。对于非命名实体和链接不到知识库的命名实体，此项为空串
+formal	string	否	词汇的标准化表达，主要针对时间、数字单位，没有归一化表达的，此项为空串
+basic_words	array(string)	是	基本词成分
+loc_details	array(object)	否	地址成分，非必需，仅对地址型命名实体有效，没有地址成分的，此项为空数组。
++type	string	是	成分类型，如省、市、区、县

参数名称	类型	**必需**	详细说明
++byte_offset	int	是	在item中的字节级offset（使用UTF8编码）
++byte_length	int	是	字节级length（使用UTF8编码）

词法分析（定制版） 返回示例

参考词法分析接口

7.3.3 依存句法分析

依存句法分析接口可自动分析文本中的依存句法结构信息，利用句子中词与词之间的依存关系来表示词语的句法结构信息（如“主谓”、“动宾”、“定中”等结构关系），并用树状结构来表示整句的结构（如“主谓宾”、“定状补”等）。

```
Json::Value result;

std::string text = "张飞";

// 调用依存句法分析
result = client.dep_parser(text, aip::null);

// 如果有可选参数
std::map<std::string, std::string> options;
options["mode"] = "1";

// 带参数调用依存句法分析
result = client.dep_parser(text, options);
```

依存句法分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	std::string	待分析文本（目前仅支持UTF8编码），长度不超过256字节

参数名称	是否必选	类型	说明
mode	否	std::string	模型选择。默认值为0，可选值 mode=0（对应web模型）；mode=1（对应query模型）

依存句法分析 返回数据参数详情

参数名称	+类型	详细说明
log_id	uint64	随机数，本次请求的唯一标识码
id	number	词的ID
word	string	词
postag	string	词性，请参照API文档中的**词性（postag)取值范围**
head	int	词的父节点ID
+deprel	string	词与父节点的依存关系，请参照API文档的**依存关系标识**

依存句法分析 返回示例

```
{
    "log_id": 12345,
    "text":"今天天气怎么样",
    "items":[
        {
            "id":"1", //id
            "word":"今天", //word
            "postag":"t", //POS tag
            "head":"2", //id of current word's parent
            "deprel":"ATT"  //depend relations between current word and parent
        },
        {
            "id":"2",
            "word":"天气",
            "postag":"n",
```

```
            "head":"3",
            "deprel":"SBV",
        },
        {
            "id":"3",
            "word":"怎么样",
            "postag":"r",
            "head":"0",
            "deprel":"HED",
        }
    ]
}
```

7.3.4 词向量表示

词向量表示接口提供中文词向量的查询功能。

```
Json::Value result;

std::string word = "张飞";

// 调用词向量表示
result = client.word_embedding(word, aip::null);
```

词向量表示 请求参数详情

参数名称	是否必选	类型	说明
word	是	std::string	文本内容（UTF8编码），最大64字节

词向量表示 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
word	string	查询词
vec	float	词向量结果表示

词向量表示 返回示例

```
{
  "word": "张飞",
  "vec": [
    0.233962,
    0.336867,
    0.187044,
    0.565261,
    0.191568,
    0.450725,
    ...
    0.43869,
    -0.448038,
    0.283711,
    -0.233656,
    0.555556
  ]
}
```


7.3.5 DNN语言模型

中文DNN语言模型接口用于输出切词结果并给出每个词在句子中的概率值,判断一句话是否符合语言表达习惯。

```
Json::Value result;

std::string text = "床前明月光";

// 调用DNN语言模型
result = client.dnnlm_cn(text, aip::null);
```

DNN语言模型 请求参数详情

参数名称	是否必选	类型	说明
text	是	std::string	文本内容（UTF8编码），最大512字节，不需要切词

DNN语言模型 返回数据参数详情

参数	类型	说明
log_id	uint64	请求唯一标识码
word	string	句子的切词结果
prob	float	该词在句子中的概率值,取值范围[0,1]
ppl	float	描述句子通顺的值：数值越低，句子越通顺

```
{
  "text": "床前明月光",
  "items": [
    {
      "word": "床",
      "prob": 0.0000385273
    },
    {
      "word": "前",
      "prob": 0.0289018
    },
    {
      "word": "明月",
      "prob": 0.0284406
    },
    {
      "word": "光",
      "prob": 0.808029
    }
  ],
  "ppl": 79.0651
}
```

DNN语言模型 返回示例

7.3.6 词义相似度

输入两个词，得到两个词的相似度结果。

```
Json::Value result;

std::string word_1 = "北京";
```

```
std::string word_2 = "上海";

// 调用词义相似度
result = client.word_sim_embedding(word_1, word_2, aip::null);

// 如果有可选参数
std::map<std::string, std::string> options;
options["mode"] = "0";

// 带参数调用词义相似度
result = client.word_sim_embedding(word_1, word_2, options);
```

词义相似度 请求参数详情

参数名称	是否必选	类型	说明
word_1	是	std::string	词 1（UTF8 编码），最大64字节
word_2	是	std::string	词 1（UTF8 编码），最大64字节
mode	否	std::string	预留字段，可选择不同的词义相似度模型。默认值为0，目前仅支持mode=0

词义相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识码,随机数
score	number	相似度分数
words	array	输入的词列表
+word_1	string	输入的word1参数
+word_2	string	输入的word2参数

词义相似度 返回示例

```
{
    "score": 0.456862,
    "words": {
```

```
        "word_1": "北京",
        "word_2": "上海"
    }
}
```

7.3.7 短文本相似度

短文本相似度接口用来判断两个文本的相似度得分。

```
Json::Value result;

std::string text_1 = "浙富股份";

std::string text_2 = "万事通自考网";

// 调用短文本相似度
result = client.simnet(text_1, text_2, aip::null);

// 如果有可选参数
std::map<std::string, std::string> options;
options["model"] = "CNN";

// 带参数调用短文本相似度
result = client.simnet(text_1, text_2, options);
```

短文本相似度 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text_1	是	std::string		待比较文本1（UTF8编码），最大512字节
text_2	是	std::string		待比较文本2（UTF8编码），最大512字节
model	否	std::string	BOWCNNGRNN	默认为“BOW”，可选“BOW”、“CNN”与“GRNN”

短文本相似度 返回数据参数详情

参数	类型	描述
log_id	number	请求唯一标识
score	number	两个文本相似度得分
texts	array	输入文本
+text_1	string	第一个短文本
+text_2	string	第二个短文本

短文本相似度 返回示例

```
{
    "log_id": 12345,
    "texts":{
        "text_1":"浙富股份",
        "text_2":"万事通自考网"
    },
    "score":0.3300237655639648 //相似度结果
},
```

7.3.8 评论观点抽取

评论观点抽取接口用来提取一条评论句子的关注点和评论观点，并输出评论观点标签及评论观点极性。

```
Json::Value result;

std::string text = "三星电脑电池不给力";

// 调用评论观点抽取
result = client.comment_tag(text, aip::null);

// 如果有可选参数
std::map<std::string, std::string> options;
options["type"] = "13";

// 带参数调用评论观点抽取
result = client.comment_tag(text, options);
```

评论观点抽取 请求参数详情

参数名称	是否必选	类型	可选值范围	说明
text	是	std::string		评论内容（UTF8编码），最大10240字节
type	否	std::string	1 - 酒店2 - KTV3 - 丽人4 - 美食餐饮5 - 旅游6 - 健康7 - 教育8 - 商业9 - 房产10 - 汽车11 - 生活12 - 购物13 - 3C	评论行业类型，默认为4（餐饮美食）

评论观点抽取 返回数据参数详情

参数	类型	描述
log_id	uint64	请求唯一标识码
prop	string	匹配上的属性词
adj	string	匹配上的描述词
sentiment	int	该情感搭配的极性（0表示消极，1表示中性，2表示积极）
begin_pos	int	该情感搭配在句子中的开始位置
end_pos	int	该情感搭配在句子中的结束位置
abstract	string	对应于该情感搭配的短句摘要

评论观点抽取 返回示例

```
{
    "items": [
        {
        "prop":"电池",
        "adj": "不给力",
        "sentiment": 0,
        "begin_pos": 8,
```

```
        "end_pos": 18,
        "abstract":"三星电脑<span>电池不给力</span>"
        }
    ]
}
```

7.3.9 情感倾向分析

对包含主观观点信息的文本进行情感极性类别（积极、消极、中性）的判断，并给出相应的置信度。

```
Json::Value result;

std::string text = "苹果是一家伟大的公司";

// 调用情感倾向分析
result = client.sentiment_classify(text, aip::null);
```

情感倾向分析 请求参数详情

参数名称	是否必选	类型	说明
text	是	std::string	文本内容（UTF8编码），最大102400字节

情感倾向分析 返回数据参数详情

参数	是否必须	类型	说明
text	是	string	输入的文本内容
items	是	array	输入的词列表
+sentiment	是	number	表示情感极性分类结果, 0:负向，1:中性，2:正向
+confidence	是	number	表示分类的置信度
+positive_prob	是	number	表示属于积极类别的概率
+negative_prob	是	number	表示属于消极类别的概率

情感倾向分析 返回示例

```
{
    "text":"苹果是一家伟大的公司",
    "items":[
        {
            "sentiment":2,    //表示情感极性分类结果
            "confidence":0.40, //表示分类的置信度
            "positive_prob":0.73, //表示属于积极类别的概率
            "negative_prob":0.27  //表示属于消极类别的概率
        }
    ]
}
```

7.3.10 文章标签

文章标签服务能够针对网络各类媒体文章进行快速的内容理解，根据输入含有标题的文章，输出多个内容标签以及对应的置信度，用于个性化推荐、相似文章聚合、文本内容分析等场景。

```
Json::Value result;

std::string title = "iphone手机出现“白苹果”原因及解决办法，用苹果手机的可以看下";

std::string content = "如果下面的方法还是没有解决你的问题建议来我们门店看下成都市锦江区红星路三段99号银石广场24层01室。";

// 调用文章标签
result = client.keyword(title, content, aip::null);
```

文章标签 请求参数详情

参数名称	是否必选	类型	说明
title	是	std::string	篇章的标题，最大80字节
content	是	std::string	篇章的正文，最大65535字节

文章标签 返回数据参数详情

参数	是否必须	类型	说明
items	是	array(object)	关键词结果数组，每个元素对应抽取到的一个关键词
+tag	是	string	关注点字符串
+score	是	number	权重(取值范围0~1)

文章标签 返回示例


```
{
    "log_id": 4457308639853058292,
    "items": [
        {
            "score": 0.997762,
            "tag": "iphone"
        },
        {
            "score": 0.861775,
            "tag": "手机"
        },
        {
            "score": 0.845657,
            "tag": "苹果"
        },
        {
            "score": 0.83649,
            "tag": "苹果公司"
        },
        {
            "score": 0.797243,
            "tag": "数码"
        }
    ]
}
```


7.3.11 文章分类

对文章按照内容类型进行自动分类，首批支持娱乐、体育、科技等26个主流内容类型，为文章聚类、文本内容分析等应用提供基础技术支持。

```
Json::Value result;

std::string title = "欧洲冠军杯足球赛";

std::string content = "欧洲冠军联赛是欧洲足球协会联盟主办的年度足球比赛，代表欧洲俱乐部足球最高荣誉和水平，被认为是全世界最高素质、最具影响力以及最高水平的俱乐部赛事，亦是世界上奖金最高的足球赛事和体育赛事之一。";

// 调用文章分类
result = client.topic(title, content, aip::null);
```

文章分类 请求参数详情

参数名称	是否必选	类型	说明
title	是	std::string	篇章的标题，最大80字节
content	是	std::string	篇章的正文，最大65535字节

文章分类 返回数据参数详情

参数名称	类型	详细说明
item	object	分类结果，包含一级与二级分类
+lv1_tag_list	array of objects	一级分类结果
+lv2_tag_list	array of objects	二级分类结果
++score	float	类别标签对应得分，范围0-1
++tag	string	类别标签

文章分类 返回示例

```
{
    "log_id": 5710764909216517248,
    "item": {
        "lv2_tag_list": [
            {
                "score": 0.895467,
                "tag": "足球"
            },
```

```
            {
                "score": 0.794878,
                "tag": "国际足球"
            }
        ],
        "lv1_tag_list": [
            {
                "score": 0.88808,
                "tag": "体育"
            }
        ]
    }
}
```

7.4 错误信息

7.4.1 错误返回格式

若请求错误，服务器将返回的JSON文本包含以下参数：

- error_code：错误码。
- error_msg：错误描述信息，帮助理解和解决发生的错误。

7.4.2 错误码

错误码	错误信息	描述
4	Open api request limit reached	集群超限额
14	IAM Certification failed	IAM鉴权失败，建议用户参照文档自查生成sign的方式是否正确，或换用控制台中ak sk的方式调用
17	Open api daily request limit reached	每天流量超限额
18	Open api qps request limit reached	QPS超限额
19	Open api total request limit reached	请求总量超限额
100	Invalid parameter	无效参数

错误码	错误信息	描述
110	Access token invalid or no longer valid	Access Token失效
111	Access token expired	Access token过期
282000	internal error	服务器内部错误，请再次请求，如果持续出现此类错误，请通过QQ群（632426386）或工单联系技术支持团队。
282002	input encoding error	编码错误，请使用GBK编码
282004	invalid parameter(s)	请求中包含非法参数，请检查后重新尝试
282130	no result	当前查询无结果返回，出现此问题的原因一般为：参数配置存在问题，请检查后重新尝试
282131	input text too long	输入长度超限，请查看文档说明
282133	param {参数名} not exist	接口参数缺失
282300	word error	word不在算法词典中
282301	word\1 error | word\1提交的词汇暂未收录，无法比对相似度	
282302	word\2 error | word\2提交的词汇暂未收录，无法比对相似度	
282303	word\1&word\2 error	word \1 和 word \2 暂未收录，无法比对相似度

第8章　常见问题

Q：输入编码是什么？
A：统一用GBK编码。

Q：结果中的词性标注都是什么含义？
A：详见下表。

词性	含义	词性	含义	词性	含义	词性	含义
n	普通名词	f	方位名词	s	处所名词	t	时间名词
nr	人名	ns	地名	nt	机构团体名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副动词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号

Q：短文本相似度对文字字数有什么限制？
A：每个短文本不要超过30个汉字或60字节。

Q：短文本相似度计算，中英文混杂怎么办？
A：模型词表中包含常用高频英文单词，也可以进行匹配。

Q：为什么有时短文本相似度计算没有返回结果？
A：尽管模型词表很大（百万级），但仍然偶尔会出现不在词表的问题，当文本所有单词都不在词表中的时候，会得不到结果。

Q：评论观点抽取对输入的评论长度有限制么？
A：建议输入字符长度不超过150字，即保持在常用评论字符长度范围内。理论上评论长度不做限制，但是平台限制字符串长度为102400字符，超过即截断。

Q：评论观点抽取可以标记挖掘出观点的文本位置吗？
A：可以的，输出结果中包含观点标签在原始文本中的位置。例如可以标记出：这家旅店服务还是不错的。

Q：评论观点抽取支持自定义词典上传吗？

A：后续版本计划开放支持，敬请期待。

Q：评论观点抽取可以批量上传并总结好标签及个数吗？
A：可以利用接口实现该功能。接口可以实现对每个评论的评论观点标签抽取和极性分析，多次调用即可实现多评论的标签挖掘和分析。

Q：中文词向量表示为什么很多词的相似度都是1？
A：尽管词向量的词表在百万量级，但仍有可能出现不在词表中的词，不在词表中的词统一映射到OOV（out-of-vocabulary）中，所以当词对中的两个词都是OOV的时候，相似度为1。

Q：中文DNN语言模型对文本个数有什么限制？
A：单个句子不要超过150个汉字。

Q：中文DNN语言模型中英文混杂怎么办？
A：模型词表中包含常用高频英文单词，也可以进行匹配。