自然语言处理 API

接口调用规范

版本号 V1.0

北京新奥特云视科技有限公司

2018.07

目　录

第1章 自然语言处理 API 概述

欢迎使用云视 ONAIR 自然语言处理服务。本文档主要针对 API 开发者，描述自然语言处理接口服务的相关技术内容。

对服务接口调用统一采用标准的 RESTFull 风格的 API 接口，支持 HTTP POST 发送请求， 这种方式下请求参数根据示例要求业务数据需要以 JSON 格式放到 HTTP BODY 里传输。调用服务端统一为 HTTP(S) 协议的请求，为了获得更高的安全性，推荐您使用 HTTPS 通道发送请求。根据请求的处理情况，系统响应数据格式统一为 JSON 数据格式。

第2章 调用方式

2.1 请求结构说明

2.1.1 请求地址

服务调用接口地址规则：

http(s)://[域名]/[接口地址]?[参数]

访问地址由三部分组成：

域　　名：接口服务部署后提供访问域名。

接口地址：接口服务根据实现功能的不同提供相应的地址。

参　　数：【注】所有参数区分大小写，命名遵循驼峰命名规则。请求及返回结果都使用 UTF-8 字符集进行编码。

2.1.2 地址示例

https://【域名地址】/nlp/keyWords?appName=123456

【第一部分】【第二部分　　】【第三部分　　】

第一部分是具体接口服务的域名地址；

第二部分是接口名称，是每个接口调用的必填项；

第三部分是接口参数。

2.2 返回结果

```
{
    "code": 0,   #状态码,参照附录
    "message": "success",#描述信息
    "data": {
}#结果数据
}
```

第3章 中文分词服务

中文分词指的是将汉字序列切分成词序列。 因为在汉语中，词是承载语义的最基本的单元。分词是信息检索、文本分类、情感分析等多项中文自然语言处理任务的基础。

3.1 分词任务

3.1.1 功能描述

把用户输入的文本内容拆分成词序列。

3.1.2 请求参数

属性	含义	类型	说明	必填
content	待分词文本	String	待分词文本，统一输入为 utf-8	
type	分词算法	String	默认即可，暂不支持自定义	

3.1.3 返回参数

参数	含义	格式	说明
word	词语	String	
nature	词性	String	见词性标注集相关说明
offset	起始位置	Integer	在文本中的起始位置
frequency	词频	Integer	参考词频

3.1.4 请求示例

```
POST http(s)://【域名】/nlp/splitWords
Content-Type: application/json
{"content":"国务院总理李克强调研上海外高桥时提出，支持上海积极探索新机制。"}
```

3.1.5 返回示例

```
{
    "code": 0,
    "message": "success",
    "data": {
        "results": [
            {
                "word": "国务院",
```

```
            "nature": "nt",
            "offset": 0,
            "frequency": 2721
        },
        {
            "word": "总理",
            "nature": "nnt",
            "offset": 3,
            "frequency": 1000
        },
        {
            "word": "李克强",
            "nature": "nr",
            "offset": 5,
            "frequency": 1
        },
        {
            "word": "调研上海外高桥",
            "nature": "nt",
            "offset": 8,
            "frequency": 0
        },
        {
            "word": "时",
            "nature": "qt",
            "offset": 15,
            "frequency": 32431
        },
        {
            "word": "提出",
            "nature": "v",
            "offset": 16,
            "frequency": 8405
        },
        {
            "word": "，",
            "nature": "w",
            "offset": 18,
            "frequency": 1
        },
        {
            "word": "支持",
            "nature": "v",
```

```
                "offset": 19,
                "frequency": 5979
            },
            {
                "word": "上海",
                "nature": "ns",
                "offset": 21,
                "frequency": 7127
            },
            {
                "word": "积极探索",
                "nature": "v",
                "offset": 23,
                "frequency": 272
            },
            {
                "word": "新机制",
                "nature": "n",
                "offset": 27,
                "frequency": 89
            },
            {
                "word": "。",
                "nature": "w",
                "offset": 30,
                "frequency": 1
            }
        ]
    }
}
```


第4章 关键词提取服务

关键词提取是指，从文章中抽取代表性词作为该文章的关键词。

4.1 关键词任务

4.1.1 功能描述

从文章中抽取代表性词作为该文章的关键词。

4.1.2 请求参数

属性	含义	类型	说明	必填
content	待分词文本	String	待分词文本，统一输入为 utf-8	是
type	垂直领域	String	默认新闻行业，暂不支持自定义	否

4.1.3 返回参数

参数	含义	格式	说明
word	词语	String	
score	参考得分	Float	

4.1.4 请求示例

POST http(s)://【域名】/nlp/keyWords
Content-Type: application/json
{"content":"北京新奥特云视科技有限公司是中国领先的视频云技术服务商，公司致力于打造为广电、互联网等领域服务的专业化的全媒体云服务平台，推出的ONAIR PaaS+平台采用开放的设计思想和先进的云计算架构，已经为国内多家省级电视台提供了云服务产品，并获得了业内专家和用户的好评。ONAIR平台提供的全业务、全功能、全流程的一站式云服务产品，将全面助力传统业务的转型和融合媒体的发展。"}

4.1.5 返回示例


```
{
    "code": 0,
    "message": "success",
    "data": {
        "results": [
            {
                "keyword": "平台",
                "score": 2.4383237
            },
            {
                "keyword": "公司",
                "score": 1.9855571
            },
            {
                "keyword": "领域",
                "score": 1.7715139
            },
            {
                "keyword": "互联网",
                "score": 1.7642295
            },
            {
                "keyword": "专家",
                "score": 1.7357726
            },
            {
                "keyword": "技术",
                "score": 1.6918215
            }
        ]
    }
}
```

第5章 标签提取服务

5.1 提取标签任务

5.1.1 功能描述

根据文章内容给文章打标签。

5.1.2 请求参数

属性	含义	类型	说明	必填
content	待提取标签文本	String	待提取标签文本，统一输入为 utf-8	是

5.1.3 返回参数

参数	含义	格式	说明
word	词语	String	

5.1.4 请求示例

POST http(s)://【域名】/nlp/labelWords
Content-Type: application/json
{"content":"北京新奥特云视科技有限公司是中国领先的视频云技术服务商，公司致力于打造为广电、互联网等领域服务的专业化的全媒体云服务平台，推出的ONAIR PaaS+平台采用开放的设计思想和先进的云计算架构，已经为国内多家省级电视台提供了云服务产品，并获得了业内专家和用户的好评。ONAIR平台提供的全业务、全功能、全流程的一站式云服务产品，将全面助力传统业务的转型和融合媒体的发展。"}

5.1.5 返回示例


```
{
    "code": 0,
    "message": "success",
    "data": {
        "results": [
            {
                "word": "平台"
            },
            {
                "word": "公司"
            },
            {
                "word": "领域"
            },
            {
                "word": "互联网"
            },
            {
                "word": "专家"
            },
            {
                "word": "技术"
            },
            {
                "word": "ONAIR"
            },
            {
                "word": "流程"
            },
            {
                "word": "云服务产品"
            },
            {
                "word": "广电"
            }
        ]
    }
}
```

第6章 文本分类服务

6.1 文本分类任务

6.1.1 功能描述

根据文本内容获取文本属于各个分类的概率，如科技、娱乐、体育、时尚等。

6.1.2 请求参数

属性	含义	类型	说明	必填
content	待提取标签文本	String	待提取标签文本，统一输入为 utf-8	是
count	保留分类个数	int	保留分类数，默认为 3（保留三个分类）	否

6.1.3 返回参数

参数	含义	格式	说明
results	分类数组	category	文本可能的分类数组，包含分类名和概率

category 格式：

参数	含义	格式	说明
category	分类名	String	文本分类名
score	概率	Float	文本属于该分类的概率

6.1.4 请求示例

```
POST http(s)://【域名】/textClassify/textClassify
Content-Type: application/json
{"content":"想用衬衫打造女人味？可以！你只需要比别人多解开两颗扣子，打造深V领型，你的衬衫就变成了释放女性魅力的秘密武器！露出迷人的锁骨是另一种健康的性感！"}
```

6.1.5 返回示例

```
{
    "code": 0,
    "message": "success",
    "data": {
        "results": [
            {
                "category": "时尚",
                "score": 0.3261357691870842
            },
            {
                "category": "造型",
                "score": 0.31312350351235074
            },
            {
                "category": "街拍",
                "score": 0.12138302904619044
            }
        ]
    }
}
```

第7章 错误码

错误码	描述	状态码	说明
Success	处理成功	0	
Failure	处理失败	1	

第8章 词性标注集

ONAIR 词性标注集			
a	形容词	m	数词
ad	副形词	mg	数语素
ag	形容词性语素	Mg	甲乙丙丁之类的数词
al	形容词性惯用语	mq	数量词
an	名形词	n	名词
b	区别词	nb	生物名
begin	仅用于始##始	nba	动物名
bg	区别语素	nbc	动物纲目
bl	区别词性惯用语	nbp	植物名
c	连词	nf	食品，比如“薯片”
cc	并列连词	ng	名词性语素
d	副词	nh	医药疾病等健康相关名词
dg	辄,俱,复之类的副词	nhd	疾病
dl	连语	nhm	药品
e	叹词	ni	机构相关（不是独立机构名）
end	仅用于终##终	nic	下属机构
f	方位词	nis	机构后缀
g	学术词汇	nit	教育相关机构

gb	生物相关词汇	nl	名词性惯用语
gbc	生物类别	nm	物品名
gc	化学相关词汇	nmc	化学品名
gg	地理地质相关词汇	nn	工作相关名词
gi	计算机相关词汇	nnd	职业
gm	数学相关词汇	nnt	职务职称
gp	物理相关词汇	nr	人名
h	前缀	nr1	复姓
i	成语	nr2	蒙古姓名
j	简称略语	nrf	音译人名
k	后缀	nrj	日语人名
l	习用语	ns	地名
nsf	音译地名	rzt	时间指示代词
nt	机构团体名	rzv	谓词性指示代词
ntc	公司名	s	处所词
ntcb	银行	t	时间词
ntcf	工厂	tg	时间词性语素
ntch	酒店宾馆	u	助词
nth	医院	ud	助词
nto	政府机构	ude1	的 底
nts	中小学	ude2	地

ntu	大学	ude3	得
nx	字母专名	udeng	等 等等 云云
nz	其他专名	udh	的话
o	拟声词	ug	过
p	介词	uguo	过
pba	介词“把”	uj	助词
pbei	介词“被”	ul	连词
q	量词	ule	了 喽
qg	量词语素	ulian	连 （“连小学生都会”）
qt	时量词	uls	来讲 来说 而言 说来
qv	动量词	usuo	所
r	代词	uv	连词
rg	代词性语素	uyy	一样 一般 似的 般
Rg	古汉语代词性语素	uz	着
rr	人称代词	uzhe	着
ry	疑问代词	uzhi	之
rys	处所疑问代词	v	动词
ryt	时间疑问代词	vd	副动词
ryv	谓词性疑问代词	vf	趋向动词
rz	指示代词	vg	动词性语素
rzs	处所指示代词	vi	不及物动词（内动词）

vl	动词性惯用语	wm	冒号，全角：： 半角：：
vn	名动词	wn	顿号，全角：、
vshi	动词“是”	wp	破折号，全角：——　- - ——- 半角：— —-
vx	形式动词	ws	省略号，全角：…… …
vyou	动词“有”	wt	叹号，全角：！
w	标点符号	ww	问号，全角：？
wb	百分号千分号，全角：％ ‰ 半角：%	wyy	右引号，全角：” ’ 』
wd	逗号，全角：， 半角：,	wyz	左引号，全角：“ ‘ 『
wf	分号，全角：； 半角： ;	x	字符串
wh	单位符号，全角：￥ $ £ ° ℃ 半角：$	xu	网址 URL
wj	句号，全角：。	xx	非语素字
wky	右括号，全角：）〕 ］ ｝ 》 】 〗 〉 半角：)] { >	y	语气词(delete yg)
wkz	左括号，全角：（ 〔 ［ ｛ 《 【 〖 〈 半角：([{ <	yg	语气语素
zg	状态词	z	状态词