**粗粒度分词文档**

**分词细则：**

根据词性来对应切分规则，词性可查询《汉语信息词典》。

《汉语信息词典》中的词都不切分，词典内没有包含的词，根据文档切分规则来判断是否切分。

1. **普通名词**

①亲属名

把亲属名看做一个词，均不切分。

②数词+名词

如果一个度量词 (如：米、件、名等) 可以插入到 数词 – 名词 之间，并且不

改变原来表达的意思，那么我们把 数词 和 名词 切分，否则不切分。

切分例子：一 学生

不切分例子：半边

③区别词+名词

不切分。

不切分举例：女朋友 慢性胃炎 古钱币 总产值

职位不切分。

④代词+名词

代词为单字时不切分，双字切分。

切分举例：什么 意思

不切分举例：我单位 本地区 各部门

特例：“**多少/那**“+名词，“**多少/那**” 与后面的名词切开。

⑤前缀/修饰成分（形容词）+ 名词

a. 前缀+名词

不切分。

不切分举例：阿哥 老鹰 非金属 超声波

b. 修饰成分+名词

修饰成分和名词都是单音节，不切分，否则切分。

切分举例：前 民主德国 前 女友 云 播放器 微 相册

不切分举例：女人

若修饰成分是形容词：

1. 如果切分开不影响原有组合的意义，则切分。

切分举例：绿 叶 小 床

1. 如果切分开影响原有组合的意义，则不切分。

不切分举例：小媳妇 藏高帽儿 小个子 老中医

c. 名词+后缀

名词2为“家”“手”“性””员””子“”化“”长“”头“”者““版”等，不切分。

不切分举例：科学家 拖拉机手 革命性 理发员 标准化 国行版

⑥方位词+名词

如果都是单音节的，那么就不切分。

⑦名词+方位词

如果以下三点都满足，则不切分：

（1）名词和方位词都是单音节

（2）在语境中名词是粘着的，或者无指的（non-referential）

（3）在语境中名词不能被 限定词-度量词 修饰。

否则切分。

⑧名词1+名词2（名词1 修饰 名词2）

（1）如果名词1 有一或两个音节，名词2只有一个音节，则不切分。

（2）二字名词或结合紧密的二字名词词组，不切分。

不切分举例：火车 牛肉 钢铁

（3）结合紧密，分开后违背原有组合意义的名词性词组，不切分。

不切分举例：有功功率 被子植物 （可认为是专有名词）

（4）各类专业的基本术语不切分。

不切分举例：加速度 中央处理器 （可认为是专有名词）

⑨动词+名词

首先查询《汉语信息词典》，已登陆词则不切分。

未登陆词参考9.7节《动宾结构》来判断是否切分。

⑩其他语言的汉字音译外来词，不切分。

不切分举例：巧克力 吉普

⑪ 名词+单语素“儿”，不切分。

不切分举例：花儿 悄悄儿 玩儿 小个儿

⑫群体名词属于普通名词，切分。

切分举例：牙科 医生 英语 老师

⑬ not+名词：根据not的词性来切分：

not是前缀时，参考1.5。

not是动词时，参考9.7。

1. **专有名词**

①人名

人名看做一个词，但有空格时要切分。

英文名字切分。

②带前后缀的人名

不切分。

③人名+称谓

切分。

④地名/团体/机构/组织/事件/运动等

1. 国家名称不论长短，不切分。

不切分举例：中华人民共和国 大不列颠及北爱尔兰联合国

1. 地名，包括行政区划等，不切分。【城市公共设施及地标性建筑物的标注也标注为地名，如广场、公园、动(植)物园、火车站、机场、体育场(馆)、游泳馆(池)、赛车场、天文馆、观测站、气象厅(堂)、大会堂、教堂、寺庙、展览馆、图书馆、大厦(楼)等。】

不切分举例：四川省 宁夏回族自治区 亚马逊河 台湾海峡

蒋家胡同 陶然亭公园 哈尔滨火车站 北京图书馆

**长地名的切分逻辑：从大地点到小地点**

**切分举例：**

**在 中国 河南省 郑州市 中原路 93 号 举行**

1. 组织/机构/单位/公司的全名，不切分。

不切分举例：联合国儿童基金会 中国计算机学会 富士通株式会社

1. 事件运动/大学名称等，不切分。

不切分举例：卢沟桥事件 明治维新 北京大学

⑤专有名词+专有名词（缩写）

不切分。

不切分举例：中美 东新澳 德意日 港澳台

⑥以序号命名的名称

切分。

⑦书/报纸/杂志/文档/电影名/游戏/APP/红头文件/品牌等

不切分。

不切分举例：《人民日报》 红楼梦 杜甫诗选

全民K歌 苹果手机

类似《底特律：成为人类》，名字中带有标点符号的，将标点符号单独切分：

底特律 ： 成为人类

像 手游版/电脑版 这种不在原本游戏名字里的，要和游戏名字切分开

切分举例: 方舟 手游版

注意：专有名词如果后接英文字符格式，例如“txt“”dvd“等，切分。

切分举例：加油妈妈 dvd 银狐 txt

品牌和产品型号切分。

切分举例：苹果 11 pro

品牌+产品，是否切分取决于是否登录，登录情况可通过

百度百科，维基百科等查询。

不切分举例：苹果手机 华为手机

切分举例：恒驰 电动车 自然旋律 六胜肽小安瓶

⑧无法辨别的人名/地名/机构名

切分。

⑨大型活动/商标

活动名称不切分。

不切分举例：迎香港回归京九植绿护绿活动 首都健康中国发展大会

商标名称切分。

切分举例：康师傅 方便面

⑩食谱

如果已经拥有词条，则作为登录菜品，不切分。

否则切分。

不切分举例：鸡蛋汤 炖鸡

切分举例：竹笋 虫草花 炖鸡

⑪专有名词+单语素

专有名称后接单音节的语素，如表示民族的“族”、表示语言的“语”，表示文字

的“文”， 专有名称后接单音节的名词，如表示人种的“人”、表示奖项的“奖”，

不切分。“吧”是贴吧的缩写词，五字以内不切分，五字以上再进行处理。

⑫专有名词+多音节名词

如“文学”“精神”“主义”等，不切分。

不切分举例：法国文学 马克思主义 雷锋精神 社会主义

1. **其余修饰成分+名词**

【切分规则不变】

1. **时间词**

①一年的十二个月，每周的七天，年、月、日、小时、分钟 等等，均不切分。

不切分举例：五月 星期日 礼拜三

② “前、后、上、下、大前、大后“等+时间名词/量词，不切分。

不切分举例：前天 后年 上星期 大前天 大后年

**注意**

不切分：一会儿

【其余规则不变】

1. **方位词**

【切分规则不变】

不切分：以上 之上

1. **代词**

①代词+方位词

不切分。

不切分举例：这以后

②代词+代词，不切分

不切分举例：他们自己

1. **数词**

①纯数字

基数、序数、小数、分数、百分数一律不切分。

不切分举例：一百二十三 123.45 第一 20% 三分之二

②约数

前有副词、形容词，或者后面有“来”“多”“左右”等词表约数的，切分。

切分举例：约 一百 多 万 仅 一百 个 四十 来 个

二十 余 只 近 20 年 来 很多 人

③带“几”或“零”这类的词素

“几”和“零”属于基本的系数词，因此不切分。

不切分举例：几十 人 十几万 元 第一百零一 个

④两个数字相连的，以及“成百”“上千”均不切分

不切分举例：五六 年 十七八 岁 成百 学生 上千 人

⑤数字+量词单位

度量单位是汉字，切分。度量单位是外文字符，切分。

切分举例：-42 ℃ 2 m 一 架

特殊：“一个“是数量词，已经登陆进词典里，所以不切分。

【其余规则不变】

1. **量词**

①重复量词不切分。

不切分举例：张张 个个

②复合量词或词组

不切分。

不切分举例：人次 架次 吨公里

1. **动词**

①重复动词1

不切分。

②重复动词2

不切分。

不切分举例：说说看 想一想 想了想 想了一想 看一眼

③A-NOT-A 或者 NOT-V

不切分。

不切分举例：看不看 不写 未完成 没研究 相不相信

【补充】：

**非 未 不 无** 在粗粒度切分时：

在该词未登陆的情况下：

+名词：作为前缀，均不切分；（不 和 无 没有前缀词性）

作为动词，参考9.7

+动词：均不切分（非 未 不 此时作副词，无 没有副词词性）

+形容词：切分，此时作副词

④偏正结构动词

切分规则不变。

⑤其他小品词+动词

切分规则不变。

⑥名词+动词

切分规则不变。

⑦动词+名词【动宾结构】

切分规则不变。

如果是已经登陆的动宾结构，则不切分；如果未登录，参考动宾结构的切分规则。

三字动词，形成 动词+名词结构，拆分后不破坏语义，则切分。

切分举例：**找 工作 玩 游戏 上 厕所**

不切分举例：**拉肚子**

【述补结构】

⑧动词+结果词

不切分。

不切分举例：做完 认识到 **梦见**

⑨动词+得/不+结果词

切分规则不变。

⑩动词+方向词

不切分。

不切分举例：出得去 进不来 寄来 跑出去

⑪ 动词+动态助词

切分。

⑫ 动词+得

切分。

已登陆动词不做切分：懂得 觉得

⑬ 并列动词（动词+动词），无连接词

切分规则不变。

⑭ 动词+动介词

切分规则不变。

⑮ X + 动词

切分规则不变。

属相切分：属 虎 属 牛

形如变黄，发烫，动词+形容词的结构，切分。

切分举例：变 黄；发 烫

⑯ 动词+代词

切分。

切分举例：怎么 做 去 哪里

⑰ 动词+方位词

单字不切分，双字切分。

切分举例：健身 前后

不切分举例：喝酒后

1. **副词**

①经常使用的副词词组不切分。

不切分举例：越来越 不得不 不能不 只是 只不过 只得

只顾 只管 只好 只怕 只有 一个劲 实际上

起关联作用的“越…越…”“又…又…”等切分。

切分举例：越 走 越 远 又 香 又 甜

②关于“无“

无+名词时，无作动词，意思是“没有”，所以要不要切分取决于是什么样的动宾结构。

无+动词时，无作副词，意思是“不”，所以无+动词可以不用切分。

在“无争议”中，争议是动词，所以无作副词，可不切分。

③副词与其修饰的形容词切开

切分举例：不 好

④“最“不切分的情况：

最好 最佳 最先 最早 最后 最多 最新

【其余规则不变】

1. **介词**

内部不作切分：每当 为此

1. **连词**

【切分规则不变】

不切分：哪怕

1. **助词**

【切分规则不变】

的：与前面成分切分。

1. **感叹词**

【切分规则不变】

1. **拟声词**

②双音节拟声词

不切分。

不切分举例：哗啦哗啦

1. **标点**

单独切分。

1. **外语**

专有名词内含有外语成分，参考2.7。

外文人名按照源语言停顿切分。切分举例：斯坦利 库布里克；不切分举例：小田和正

不切分举例：卡拉OK；APP

1. **其他**

**①成语**

成语一般不切分。

多字成语，例如 人心齐，泰山移，不合在一起。

人心齐，/ 泰山移

**②习用语**

（1）四字词或结合紧密、使用稳定的四字词组、习语，不切分。

不切分举例：春夏秋冬（标注为名词） 由此可见

（2）五字或五字以上的谚语、格言等，分开后如果不影响原本的语义，则切分；如影响原本的语义，则不切分。

切分举例：时间 就是 生命

（3）惯用语及有转义的词或词组，不切分。

不切分举例：铁公鸡 喝西北风 闲人免进 好家伙 对台戏

进一步 吃香 吃醋 批复 事实上 金手指

（4）口号/歌词

切分。

（5）简缩略语

简缩略语不切分。

不切分举例：科技 奥运会 工农业 离退休 零部件 石化 火电 农副业

亚运会 联大 教委 奥委会 环保

1. **形容词**

①AA、AABB、ABB、AAB、ABAB、A里AB、一A一B、半A半B、有A有B

均不切分。

不切分举例：雪白雪白 一心一意 一清二楚 半生不熟 有条有理

副词一般跟其修饰的形容词切开。注意下面这种情况：

②单字副词+单字形容词 作形容词/副词

与常见形容词搭配，如“多“”大“”小“等，不切分。

不切分举例：太多 更多 更大

与不常见形容词搭配，判断为是未登录词的，不切分，例如“丧“”绝“等。

不切分举例：好丧 太绝了

③“多”不切分的情况：

多少 多时 多数 多久

【其余规则不变】

1. **语素和常见搭配**

⑥有

有所 有点 有的 有些 只有 不切分。

双眼 不切分

【其余切分规则不变。】

1. **非汉字字符串**

非汉字符号，例如其他语言的字符串、数学符号、化学符号等，保留原有意义，根据其原有意义决定是否切分。

①单独的罗马数字，切分。

切分举例： II IV V

②其他英文字母或数字则单独切分。

举例：24 K 镀金 C 是 光速 Windows98 CAD 3.14

③ 数字和英文字母同汉字字符切开；如果有词条，算作专有名词。

切分举例：iOS 13

不切分举例：4S店

④ 错字

判断是否破坏原词的语义，如果不破坏语义，则不切分。

如果需要切分，则按照普通规则切分。