东北大学自然语言处理实验室

东北大学自然语言处理实验室
http://www.nlplab.com

NiuTrans 团队
niutrans@mail.neu.edu.cn

©2018 NiuTrans

字-词还原

- **问题描述**

自然语言中句子由词组成，借助分词工具可以将句子按词切分开来。分词后词与词之间用空格区分。

表 1 句子分词样例

句子序号	原始句子	分词后的句子
1	我热爱自然语言实验室。	我 热爱 自然 语言 实验室 。
2	我喜欢物理实验。	我 喜欢 物理实验 。

而词并不是语义的最小单位，例如：上图中“实验室”可以进一步切分“实验 室”，但一般不能分成“实 验室”，我们称其中的“实验”为子词。“自然”也可能切分成“自 然”，而“我”单字词不能切分，如何将词有效切分成子词，就用到子词统计的方法。

在某个词中，每组*相邻两个字符（字、字母或标点等）*都有可能组成子词，我们只能通过每个可能的子词在整个语料中出现的频次最高来判断哪些组合属于子词。如上面的例子，在上述整个语料集合中“实验”出现的频次高，才能被确定为子词。本次任务是，统计语料中前 N 个最可能成为子词的组合。

- **示例**

N = 2

语料：

我 热爱 自然 语言 实验室 。

我 喜欢 物理实验 。

即：

实验

热爱

（注：结果为前 N 组字符。并列排名时取其中一个即可。）

详解：

表 2 字符两两组合结果

热爱	1
自然	1
语言	1
实验	2
验室	1
喜欢	1
物理	1
理实	1

- **数据**

当前文件夹 sample-data\目录下

祝好运！