

中文分词

概念：由于中文以字为基础而非单词，词语之间没有明显的区分标记。因此需要分词将语句拆分。

基于词典分词

机械分词法：

① 最大正向匹配法 (MPM, Maximum Matching Method)

step 1: 取待切分汉语句的 m 个字符为匹配字符， m 为词典中最长词条个数。

step 2: 查找词典进行匹配。匹配成功，则将当前字符串作为一个词切分出来。若不成功，则将字符串最后一个字去掉，重新匹配。直到切分出所有词。

e.g.

我们都有光明的前途。 X

↓

我们都 有 光明的前途。 ✓

... 1 - 2 ...

↓
... 的 X

↓
... 的 X

...
我们都 ✓

Recursion = 2

有先明的前途 X

↓:

有先

Recursion ...

Result: 我们都 有先 明 的 前途

② 逆向最大匹配法 (Reverse Maximum Matching Method).

RMM.

分词切分方向与MM相反.

从文档末尾匹配扫描, 每次取最末 2 个字符.

(字符串).

匹配成功, 连接当前一位

... / ... / ...

分词词典为逆序词典，每个词在词典中逆序存放。

实际处理时，先将文档进行词频处理，生成逆序文档，并生成逆序词典进行MM。

(RMM准确性远大于MM)。

③ 双向匹配法。

MM与RMM的结合。

将句子同时用MM与RMM进行匹配切分。

如匹配结果相同，则分词正确。

否则按最小集处理。



n-grams. {
bi-grams
tri-grams
4-grams
...

文本预处理

~ , ~ ~ ~

分句



停用词



向量