**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](mailto:niutrans@mail.neu.edu.cn)

©2018 NiuTrans

长度比过滤

* **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。给定双语平行语料，即源语言句子与目标语言句子是互译的（共1W句）：

问题1：计算整个语料库中源语言与目标语言的平均长度比length\_ratio

问题2：根据平均长度比将语料分割成两个集合，集合1中保存长度比，集合2中保存其他句子。Ps:每个集合中使用两个文件来存储，源语言与目标语言。

* **示例**

**语料：**

I am Chinese . 我 是 中国人 。

The weather is very good today ! 今天 天气 非常 好 ！

**即：**

第一句的长度比 ：

第二句的长度比：

平均长度比： =1.2

* **数据**

当前文件夹sample-data\目录下，其中英文是源语言，中文是目标语言。

**祝好运!**