东北大学自然语言处理实验室
重复翻译

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

机器翻译后处理 –重复翻译问题

目标：对机器翻译译文进行判断，确定含有重复翻译问题的行号。初步了解 n-gram。

机器翻译在处理长句子时偶尔会产生重复翻译问题，即将一个单词多次翻译，在最终译文中**某些单词或短语连续重复出现了很多次**。现给定一个文件，要求编写程序，确定给定文件中的含有重复翻译的语句的行号，并将行号以及错误句子输出到文件中。

内容：

- **示例错误**

1. Take agricultural products products as an example , with the implementation of the self-trade agreement , consumers can taste all year round from different places of origin , quality and cheap agricultural products .

2. Since we have been here before , and have sworn here , and have been particularly heartened by the arrival of the President of this habit , as if it were very encouraging , the two of us agreed that we must go again today , and revisit the process of our oath , so we are very excited to be here today after the arrival of the President of the United States of America and the United States of America and the United States of America , the United States of America , the United States of America , and the United States of America , the United States of America , and the United States of America , and the United States of the United States of America .

- **基础要求**

单词重复：只需考虑重复单词。如示例 1 中的 products 重复。

重复次数：2 次。 如果句子中某个单词重复出现两次，即认为是重复翻译。

输入：re_trans.easy （100 行，已分词，示例 1 为文件中第五句）

输出：文件，保存重复翻译数量，重复翻译的行号，以及错误句子。

- **高级要求**

短语重复：考虑不同长度的短语，短语长度范围手动输入。短语长度若为 1，即单词重复问题。示例 2 中短语“the United States of ”长度为 4。

最小重复次数：手动输入。

输入：re_trans（2481 行，已分词）

输出：文件，保存不同长度重复翻译的数量，重复翻译的行号，短语长度以及重复翻译句子（如 2-gram：15 个 3-gram：10 个。示例 1 为 1-gram，示例 2 为 4-gram）

- **数据**

当前文件夹 sample-data\目录下

祝好运！