东北大学自然语言处理实验室
语言模型

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

训练二元语言模型

- **问题描述**

给定经过分词的中文文本（词与词之间用空格隔开），计算：在出现一个词的条件下，出现另一个词的概率(也叫二元语言模型)。

- **示例**

假设“新华社”共出现 20 次，其后出现“乌鲁木齐”10 次 ， 出现“我”4 次，出现“报道”6 次，则 P(乌鲁木齐|新华社) = 10/20 = 0.5

- **输出**

将结果写到 result.out 文件中，输出格式：

当前词 \t {第一个下一个词，概率} \t {第二个下一个词，概率} … (概率由高到低排序）

如：新华社 \t {乌鲁木齐，0.6} \t {记者，0.3} \t {报道，0.1}

- **注意**

每句话的开始和结束处，自行添加两个自定义词汇， “<s>”和“</s>”，分别表示句首和句尾，作用是保证句子中的每一个词，都存在“前一个词”和“后一个词”

- **数据**

当前文件夹 sample-data\training.data

测试二元语言模型

- **问题描述**

  从 result.out 中读取二元语言模型，估计测试集中的句子 S 出现的概率。

- **提示**

  给定句子 S=$w_1 w_2 \ldots w_n$，其中w_i表示该句子中的第 i 个词，则句子 S 出现的概率为：

$$\mathrm{P(S)} = \prod_{i=2}^{n} P(w_i | w_{i-1})$$

  为了防止计算溢出，可以对上式两边同时取对数，将连乘变成连加

- **输出结果**

  将每句的概率值写到另一个文件中，每行是一个实数，表示该句子出现的概率。

- **注意**

  如果$P(w_i|w_{i-1})$不存在，怎么防止$\mathrm{P(S)}$为 0 ？

- **数据**

  当前文件夹 sample-data\test.data

祝好运！