随机单词生成器

一、题目描述

《统计自然语言处理基础》习题 1.4

使用 1.4.3 节中描述的随机字母生成器，构造一个类似表 1.3 的表（这个构造器产生从 a 到 z 的字母和空格，所有字符具有相同的生成概率 1/27）

二、程序代码

见 ./src/random_word_generator.cpp

三、结果分析

在 1000 万的词量上（我机器的极限），分布图像依然很陡峭，随机生成的单词中，%99 长度小于 3，导致 rank-频率图像几乎呈直线下降.rank 700 以后基本与 x 轴平行。
原因是：根据概率计算，长度为 n 的词产生的概率为(26/27)^n/27，即产生 n 个非空格字符后产生一个空格的概率。所以有如下结论：

- 长度 n+1 的词的数量应该比长度为 n 的词的数量（或者说种类）多 26 倍；
- 长度为 n 的词要比长度为 n+1 的词更加频繁的出现，并且他们出现的概率是的比值是一个常量。

从程序的输出 result.txt 可以验证。

四、程序运行方法

windows：
cmd 下，输入/bin/word.exe num1 num2
Linux :
./bin/word num1 num2
num1、num2 为可选参数，分别代表生成单词的数量和最大单词长度，默认为 100000 和 300 。程序运行结束后，会将结果报存在当前目录下的 result.txt 中

result.txt 为参数 10000000 300 情况下的输出

Joker Lee 李劼
Oct, 2009
jokerleee@gmail.com