逆向最大匹配分词说明及要求

From：吉晋锋

这次的任务是让大家学习一下利用程序进行分词。它是我们对得到的数据或语料进行处理的第一步，所以平时使用频率较高。而逆向最大匹配分词需要在已有词典的基础上，从被处理文档的末端开始匹配扫描，每次取最末端的i个[字符](http://baike.baidu.com/item/%E5%AD%97%E7%AC%A6" \t "_blank)（分词所确定的阈值i）作为匹配字段，若匹配失败，则去掉匹配字段最前面的一个字，继续匹配。而且选择的阈值越大，分词越慢，但准确性越好。

下面以“我们在野生动物园玩”这句话为例进行说明：

词典里有‘玩’,‘我们’，‘野生动物园’，‘在’，‘今天’这么几个词，知道了词典和要匹配的句子，这里的阈值为5，下面就要开始分词啦

第1轮扫描：“生动物园玩”

第1次：“生动物园玩”，扫描，不在词典中，下一步

第2次：“动物园玩”，扫描，不在词典中，下一步

……

第7次：“玩”，扫描有

扫描中止，输出“玩”，加‘/’,开始第2轮扫描

第2轮扫描：“野生动物园”

第1次：“野生动物园”，扫描看其是否在词典中，有

扫描中止，输出“野生动物园”， 加‘/’,开始第3轮扫描

第3轮扫描：“我们在”

第1次：“我们在”， 扫描，不在词典中，下一步

第2次：“们在”， 扫描，不在词典中，下一步

第3次：“在”，扫描有

扫描中止，输出“在”， 加‘/’,开始第4轮扫描

第4轮扫描：“我们”

第1次：“我们”，扫描有

扫描中止，输出“我们”，整体扫描结束。

利用逆向最大匹配分词，这个句子最终切分结果为：“我们/在/野生动物园/玩”

逆向最大匹配分词是中文分词基本算法之一，因为是机械切分，所以它也有分词速度快的优点，且逆向最大匹配分词比起正向最大匹配分词更符合人们的语言习惯。具体原因可以参见该网站-- <http://blog.sina.com.cn/s/blog_5dd3e9f10100u3vr.html>

而且网上有很多介绍逆向最大匹配分词的资料，可以自行百度。

要求：

1. 根据停用词词表以及词表使用逆向最大算法进行分词，即stopwords.xlsx和word.xlsx两个词表

2.此次分词不可使用jieba分词（因为jieba分词的分词效果不是很好，所以在之后的任务或实际使用中使用频率较低）