# 基于词典和规则的汉语自动分词系统

顾志浩 MF1933024

## 任务描述

1. 根据词典和规则实现一个汉语自动分词系统，输入一个句子给出分词的结果。

## 算法实现

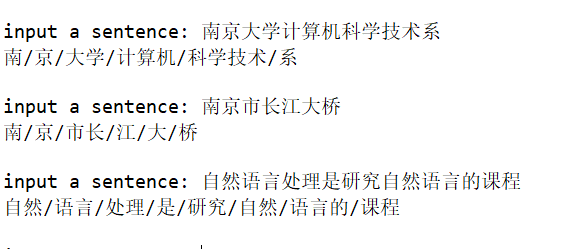
1. 正向最大匹配（FMM）
2. 开始时当前位置为0
3. 从当前位置开始找到最长的能匹配词典中词，并设置当前位置为匹配到的词的下一个词的位置
4. 重复b，直到当前位置为词的长度
5. 逆向最大匹配（RMM）
6. 开始时当前位置为词的长度
7. 从当前位置开始向前找到最长的能匹配词典中的词，并设置当前位置为匹配到的词的前一个词的位置
8. 重复b，知道当前位置为0
9. 双向最大匹配，根据正向最大匹配和逆向最大匹配的结果，选择分词数较少的结果。

## 数据

1. 词典：ce(ms-word).txt，词、英文，中间以逗号分隔。

## 测试

1. 截图，输入一个句子，程序给出分词的结果，每个词之间以斜杠隔开。



## 遇到的问题

在实验过程中，我发现产生歧义的一部分结果通常是分词数目较多的，所以我设计了一个联合正向最大匹配和逆向最大匹配的算法，通过比较两种匹配结果中分词数的大小，可以消除一些歧义。