# 分词实验

一、实验内容

本次实验期望同学们使用train.csv文件训练自己的分词器，并在test.csv文件上进行测试，最终按照sample\_submission.csv文件的格式提交测试结果。

实验采用最大匹配分词方法。实验结果提交到Kaggle网站，并且通过排行榜进行排名。

二、核心算法

实验所采用的开发平台及语言工具开发平台：任意语言工具：任意

实验的核心思想和算法描述核心思想：

最大匹配算法算法描述： 正向最大匹配法算法如下所示：逆向匹配法思想与正向一样，只是从右向左切分，这里举一个例子：输入例句：S1="计算语言学课程有意思" ；定义：最大词长MaxLen = 5；S2= " "；分隔符 = “/”；假设存在词表：...，计算语言学，课程，意思，...；最大逆向匹配分词算法过程如下：

（1）S2=""；S1不为空，从S1右边取出候选子串W="课程有意思"；

（2）查词表，W不在词表中，将W最左边一个字去掉，得到W="程有意思"；

（3）查词表，W不在词表中，将W最左边一个字去掉，得到W="有意思"；（4）查词表，W不在词表中，将W最左边一个字去掉，得到W="意思"

（5）查词表，“意思”在词表中，将W加入到S2中，S2=" 意思/"，并将W从S1中去掉，此时S1="计算语言学课程有"；

（6）S1不为空，于是从S1左边取出候选子串W="言学课程有"；

（7）查词表，W不在词表中，将W最左边一个字去掉，得到W="学课程有"；

（8）查词表，W不在词表中，将W最左边一个字去掉，得到W="课程有"；（9）查词表，W不在词表中，将W最左边一个字去掉，得到W="程有"；（10）查词表，W不在词表中，将W最左边一个字去掉，得到W="有"，这W是单字，将W加入到S2中，S2=“ /有 /意思”，并将W从S1中去掉，此时S1="计算语言学课程"；

（11）S1不为空，于是从S1左边取出候选子串W="语言学课程"；

（12）查词表，W不在词表中，将W最左边一个字去掉，得到W="言学课程"；

（13）查词表，W不在词表中，将W最左边一个字去掉，得到W="学课程"；（14）查词表，W不在词表中，将W最左边一个字去掉，得到W="课程"；（15）查词表，“意思”在词表中，将W加入到S2中，S2=“ 课程/ 有/意思/”，并将W从S1中去掉，此时S1="计算语言学"；

（16）S1不为空，于是从S1左边取出候选子串W="计算语言学"；

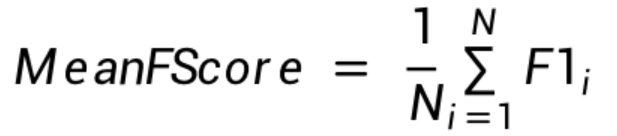
（17）查词表，“计算语言学”在词表中，将W加入到S2中，S2=“计算语

言学/ 课程/ 有/ 意思/”，并将W从S1中去掉，此时S1=""；

（18）S1为空，输出S2作为分词结果，分词过程结束。

三、评价标准：

本次实验的评价指标选用MeanFScore，具体计算方法如下：



即：分别计算每一行数据的F1，然后最终对这些F1值取平均。

四、数据下载与结果提交

数据存放在Kaggle网站上。结果也提交到Kaggle网站上。具体地址为：

<https://www.kaggle.com/competitions/csu-ai-inclass-nlp-2023>

操作教程：https://blog.csdn.net/weixin\_43722223/article/details/123121817