# Data Mining and Optimization

## Lecture 3: Word Segmentation

Liu Yang

Nanjing University

Spring, 2025

# Table of Contents

# 中文分词简介

- 词(word)是最小的能够独立活动的有意义的语言成分。
- 与英文不同，中文中的词是以字为基本单位的，而且词与词之间没有自然分隔符。
- 中文分词是通过计算机自动识别出句子的词，在词间加入分隔符，分离出各个词汇。
- 分词歧义：“结婚的和尚未结婚的”：
  - “结婚/的/和/尚未/结婚/的”？
  - “结婚/的/和尚/未/结婚/的”？

# Table of Contents

# 规则分词

基于规则的分词是一种机械分词方法，主要通过维护词典，在切分语句时，将语句的每个字符串与词表中的词进行逐一匹配，找到则切分，否则不予切分。

- 正向最大匹配法(Maximum Match Method，MM法)；
- 逆向最大匹配法(Reverse Maximum Match Method，RMM法)；
- 双向最大匹配法(Bi-direction Match Method)。

# 正向最大匹配法

- 假定词典中的最长词有$m$个汉字，则从左向右取被处理文档的前$m$个字作为匹配字段，查找字典；
- 若字典中存在这样一个$m$字词，则匹配成功，匹配字段作为一个词切分出来；
- 若字典中找不到这样一个$m$字词，则匹配失败，将匹配字段中的最后一个字去掉，对剩下的字符串重新进行匹配；
- 如此进行下去，直到匹配成功，即切分出一个词或剩余字符串长度为零为止；
- 取下一个$m$字字符串进行匹配，直到文档被扫描完为止。

# 逆向最大匹配法

- 假定词典中的最长词有$m$个汉字，从被处理文档的末端开始扫描，每次取最末端的$m$个字符作为匹配字段，查找字典；
- 若字典中存在这样一个$m$字词，则匹配成功，匹配字段作为一个词切分出来；
- 若字典中找不到这样一个$m$字词，则匹配失败，将匹配字段最前面的一个字去掉，对剩下的字符串重新进行匹配；
- 如此进行下去，直到匹配成功，即切分出一个词或剩余字符串长度为零为止；
- 取下一个$m$字字符串进行匹配，直到文档被扫描完为止。

# 双向最大匹配法

- 将正向最大匹配法得到的分词结果和逆向最大匹配法得到的结果进行比较；
- 根据最大匹配原则，选取次数切分最少的作为结果；
- 中文中90%的句子正向最大匹配法和逆向最大匹配法完全重合且正确；
- 9%的句子两种切分方法得到的结果不一样，但其中必有一个正确；
- 1%的句子两种切分方法虽然重合却是错的，或者两种方法切分不同但两个都不对。

# 双向最大匹配法

- 如果正向最大匹配法和逆向最大匹配法得到的结果词数不同，则选择分词数量较少的那个；
- 如果正向最大匹配法和逆向最大匹配法得到的结果词数相同：
  - 分词结果相同，说明没有歧义，可返回任意一个；
  - 分词结果不同，返回其中单字较少的那个。

# Table of Contents

# 中文分词包——Jieba

- jieba是一款开源的中文分词工具，可访问官方网站；
- jieba并不是只有分词功能，还提供关键词提取、词性标注等功能;
- 通过pip install jieba进行安装。

# Jieba的三种分词模式

- 精确模式：试图将句子最精确地切开，适合文本分析，最常用；
- 全模式：把句子中所有可以成词的词语都扫描出来，速度非常快，但不能解决歧义问题；
- 搜索引擎模式：在精确模式基础上，对长词再次切分。