# 语言模型

## 知识点

- N-gram 语言模型
- HMM 隐含马尔可夫简介
- viterbi 解码与词性标注
- 分词原理与应用
- 文本读写
- python 字符串方法
- 正则表达式

## QA

1. 讲一下利用HMM分词原理；
2. 利用re，使用正则将字符串"罗志祥202004月真的很倒霉，替蒋凡d挡了36489点伤害"中，连续5个以上数字替换成*符号；

## 情感分类项目

进入数据页面阅读数据集介绍，下载数据集，运用jieba进行分词等工具完成:

- 统计词频， 输出词频最高50和最低50的词；
- 统计 句子的字符长度，句子分词的词长度， 用matplotlib或者seaborn 画出分布，观察特点。