如何进行分词?

分词就是将句子、段落、文章这种长文本，分解为以字词为单位的数据结够，方便后续的建模处理。

词是表达完整含义的最小单位，字的粒度太小，无法表达完整的含义；句子的粒度太大，承载的信息多很难复用。

中文分词的方法大致分为三类;

1.基于词典匹配

基于思想是词典匹配，将待分词的文本根据一定规则切分和调整，然后跟词典中的词语进行匹配，匹配成功则按照词典的词分词，匹配失败通过调整或者重新选择，如此反复循环。（代表的方法有基于正向最大匹配和基于逆向最大匹配及双向匹配法）

优点：速度快、成本低。缺点：适应性不强，不同领域效果差异大。

2.基于统计

这类目前常用的算法是HMM、CRF、SVM、深度学习等算法。以CRF算法为例，其基本思路是对汉字进行标注训练，不仅考虑了词语出现的频率，还考虑了上下文，具备较好的学习能力，因此其对歧义词和未登录词的识别都具有更好的效果。

优点：准确率高、适应性强。缺点：成本高、速度慢

3.混合分词