N-gram简介及相关运用

参考文章：https://www.cnblogs.com/ljy2013/p/6425277.html

https://blog.csdn.net/songbinxu/article/details/80209197

N-Gram是基于一个假设：第n个词出现与前n-1个词相关（如果这个假设不成立，及第N个词出现的概率和前N-1个词都不相关，那么每次统计各个词的出现概率可能相差很大，概率的变化量可以用来考量适用场景吧），而与其他任何词不相关。整个句子出现的概率就等于各个词出现的概率乘积。各个词的概率可以通过语料中统计计算得到。假设句子T是有词序列w1,w2,w3...wn组成，用公式表示N-Gram语言模型如下：

P(T)=P(w1)*p(w2)*p(w3)***p(wn)=p(w1)*p(w2|w1)*p(w3|w1w2)***p(wn|w1w2w3...)　（公式参考：朴素贝叶斯）

→ 通俗解释，一个句子出现的概率是，以w1开头的句子出现概率 * 在w1开头的句子下后面是w2的概率 * 在w1w2开头的句子下，后面是w3的概率 * 。。。。

一般常用的N-Gram模型是Bi-Gram和Tri-Gram。分别用公式表示如下：

Bi-Gram:　P(T)=p(w1|begin)*p(w2|w1)*p(w3|w2)***p(wn|wn-1) → 只考虑当前词和前一个词的关系，而不是前 n-1个词的关系。

Tri-Gram:　P(T)=p(w1|begin1,begin2)*p(w2|w1,begin1)*p(w3|w2w1)***p(wn|wn-1,wn-2) → 只考虑当前词和前两个词的关系，而不是前 n-1个词的关系。

很明显，Tri-Gram 考虑的词比Bi-Gram 多一个，在计算概率的时候更准确。

运用场景

（场景运用介绍就不搬运了）

1.词语联想

2.垃圾短信分类

3.分词器

根据几种可能的分词情况，计算出一种概率最高的。

在搜索中的运用

在solr中使用Tri-Gram的方式进行分词。

在计算酒店特征的时候，按照实体名进行分词，假设了实体名之间有顺序关系，进行筛选，选取最优结果。