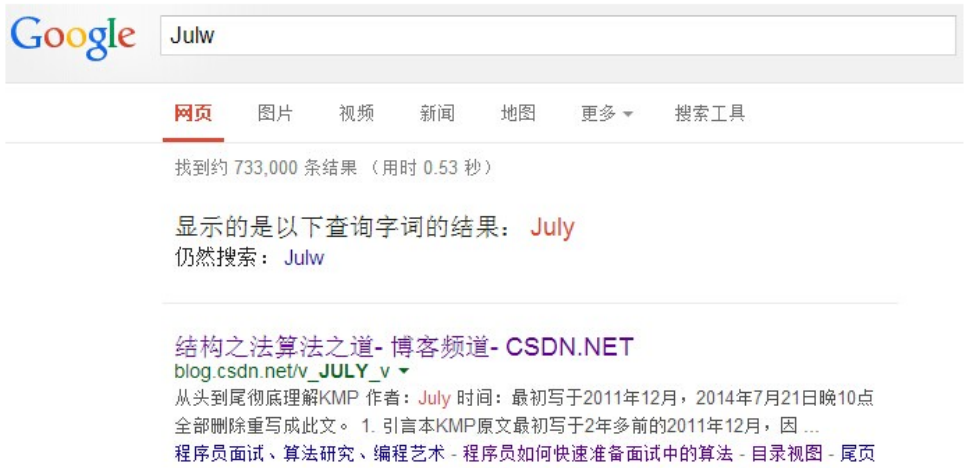
36、经常在网上搜索东西的朋友知道，当你不小心输入一个不存在的单词时，搜索引擎会提示你是不是要输入某一个正确的单词，比如当你在Google中输入“Julw”时，系统会猜测你的意图：是不是要搜索“July”，如下图所示：



这叫做拼写检查。

根据谷歌一员工写的文章显示，Google的拼写检查基于贝叶斯方法。请说说的你的理解，具体Google是怎么利用贝叶斯方法，实现"拼写检查"的功能。

用户输入一个单词时，可能拼写正确，也可能拼写错误。如果把拼写正确的情况记做c（代表correct），拼写错误的情况记做w（代表wrong），那么"拼写检查"要做的事情就是：在发生w的情况下，试图推断出c。换言之：已知w，然后在若干个备选方案中，找出可能性最大的那个c，也就是求

$P(c|w)$

的最大值。

而根据贝叶斯定理，有：

$$P(c|w) = P(w|c) * P(c)/P(w)$$

由于对于所有备选的c来说，对应的都是同一个w，所以它们的P(w)是相同的，因此我们只要最大化

$P(w|c) * P(c)$

即可。其中：

P(c)表示某个正确的词的出现"概率"，它可以用"频率"代替。如果我们有一个足够大的文本库，那么这个文本库中每个单词的出现频

率，就相当于它的发生概率。某个词的出现频率越高，P(c)就越大。比如在你输入一个错误的词“Julw”时，系统更倾向于去猜测你可能想输入的词是“July”，而不是“Jult”，因为“July”更常见。

P(w|c)表示在试图拼写c的情况下，出现拼写错误w的概率。为了简化问题，假定两个单词在字形上越接近，就有越可能拼错，P(w|c)就越大。举例来说，相差一个字母的拼法，就比相差两个字母的拼法，发生概率更高。你想拼写单词July，那么错误拼成Julw（相差一个字母）的可能性，就比拼成Jullw高（相差两个字母）。值得一提的是，一般把这种问题称为“编辑距离”，参见博客中的这篇文章。

所以，我们比较所有拼写相近的词在文本库中的出现频率，再从中挑出出现频率最高的一个，即是用户最想输入的那个词。具体的计算过程及此方法的缺陷请参见这里。