近似连接

计63 黄松皓 2016011312

主要数据结构和算法

哈希

在ED里判字符串相等，在Jacc里建立字符串token到倒排列表的映射。

joinED

将file1中的字符串分为长度大于threshold的，和其他的。

对于大于的字符串，我们将字符串分为threshold+1个子串，然后对所有长度为len的字符串位置为i的子串通过哈希建立倒排列表。

接下来对于file2中的每个字符串，依次处理，每次在长度差不超过threshold中的字符串进行搜索，对于当前串，查找一定范围内开头的子串，找到对应的倒排列表，然后对候选的字符串进行ED值的计算，同时中途进行判断，如果已经超过threshold则退出。

joinJaccard

对每个字符串的空格隔开的子串进行计数，得到token的倒排列表。

之后对于某个查找的字符串，按token的出现次数进行排序，然后找到前缀长度为$[\frac{1}{1+\delta}|x|+1-\frac{\delta}{1+\delta}|y|]$所有token对应的倒排列表，然后对这些候选字符串，进行一些判断，例如$|x| \leq |y| * \delta, |y| \leq |x| * \delta$，同时在计算Jaccard值时进行判断，若不符合停止计算。