12月11日作业

1. 何谓多重对齐？给给出中星算法描述，并用中星算法求解S={ABCD，XBCD，XABC，YBCD}。
    1. **多重对齐:** 匹配两个以上的字符串或者树
    2. **中星算法:**
        1. 首先选出一个与其他字符串距离都最近的字符串 S_c
        2. 再建立一个字符串集合 M
        3. 然后对除了 S_c 之外的每个字符串 s:
            1. 将最佳对齐的 s'_c 和 s' 加入集合M
            2. 对集合 M 的所有除了 s'_c 和 s'，在 s_c 加空格的地方加上空格。
        4. 返回 M
    3. **问题求解**
        1. 首先计算每个字符串与其他字符串的差异之和，通过我的计算得到:
            1. ABCD, 4
            2. XBCD, 4
            3. XABC, 7
            4. YBCD, 5
        2. 于是选定中心串为 ABCD
        3. **第一次迭代**

           $c^{*\prime}$： ABCD

           s'： XBCD

           更新 M $\rightarrow$ ABCD

           XBCD
        4. **第二次迭代**

           $c^{*\prime}$： _ABCD

           s'： XABC_

           更新 M $\rightarrow$ _ABCD

           _XBCD

           XABC_
        5. **第三次迭代**

           $c^{*\prime}$： _ABCD

           s'： _YBCD

           更新 M $\rightarrow$ _ABCD

           _XBCD

           XABC_

           _YBCD
        6. 最后的结果就是第三次迭代达到的 M
2. 什么是信息集成？为什么需要信息集成？

1. **信息集成：**从大量的站点中提取数据，让提取出来的信息成为一个统一的数据库。
2. 信息基础可以用集成出来的信息提供增值服务。

3. 什么是模式匹配？主要的匹配形式有哪些？
    1. **模式匹配：**对于两个或者更多数据库的模式之间产生映射，把具有相同语义的元素映射到一起。
    2. **匹配形式：**
        1. **模式层的匹配：**只考虑模式信息（属性名称和数据类型），不考虑实例信息
        2. **域和实例层的匹配**：只考虑实例数据还有每个属性，没有模式信息。
        3. **模式、域和实例的综合匹配**

4. 模式匹配时的数据预处理包括哪些方面？
    1. **分词：**将可能是多个单词组成的长字符串中间加上空格
    2. **扩展：**将缩写扩展成原形，例：dept → departure
    3. **移除停用词和词干提取：**停用词指介词、连词、 冠词 、代词等没有实际意义的词
    4. **单词标准化：**将单词的不同拼写转换成相同的，例：`child` → `children`，`colour` → `color`