# 基于HMM的分词实验

## HMM训练与预测

训练过程就是求以下三个矩阵：

1. 初始概率矩阵
2. 转移概率矩阵
3. 发射矩阵

预测过程：

基于维特比算法找到路径最大的一条路径。

## 训练过程

输入：

**文本**

1. 今天 天气 真 不错 。
2. 麻辣肥牛 好吃 ！
3. 我 喜欢 吃 好吃 的 ！

**标签**

（a）B E B E S B E S

（b）B M M E B E S

（c）S B E S B E S S

### 初始矩阵

统计每篇文章的第一个字是什么

|  |  |  |  |
| --- | --- | --- | --- |
| B | M | S | E |
| 2 | 0 | 1 | 0 |

则有

|  |  |  |  |
| --- | --- | --- | --- |
| B | M | S | E |
| 0.667 | 0 | 0.333 | 0 |

### 转移矩阵

当前状态到下一状态的概率

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | B | M | S | E |
| B | 0 | 1 | 0 | 6 |
| M | 0 | 1 | 0 | 1 |
| S | 3 | 0 | 1 | 0 |
| E | 2 | 0 | 5 | 0 |

则有

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | B | M | S | E |
| B | 0 | 0.142 | 0 | 0.875 |
| M | 0 | 0.5 | 0 | 0.5 |
| S | 0.75 | 0 | 0.25 | 0 |
| E | 0.285 | 0 | 0.715 | 0 |

### 发射矩阵

统计某种状态下，所有字出现的次数（概率）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| B | 今：1 | 天：1 | 不：1 | 麻：1 | 好：2 | 喜：1 |
| M | 辣：1 | 肥：1 |  |  |  |  |
| S | 。：1 | ！：2 | 我：1 | 吃：1 | 的：1 | 真：1 |
| E | 天：1 | 错：1 | 牛：1 | 吃：1 | 欢：1 | 气：1 |

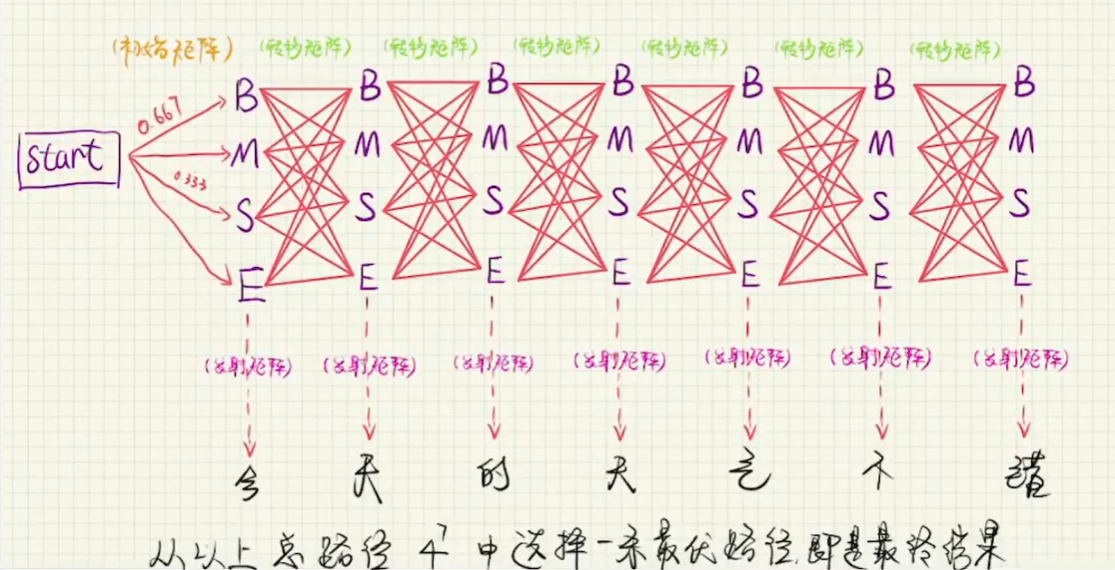
则有

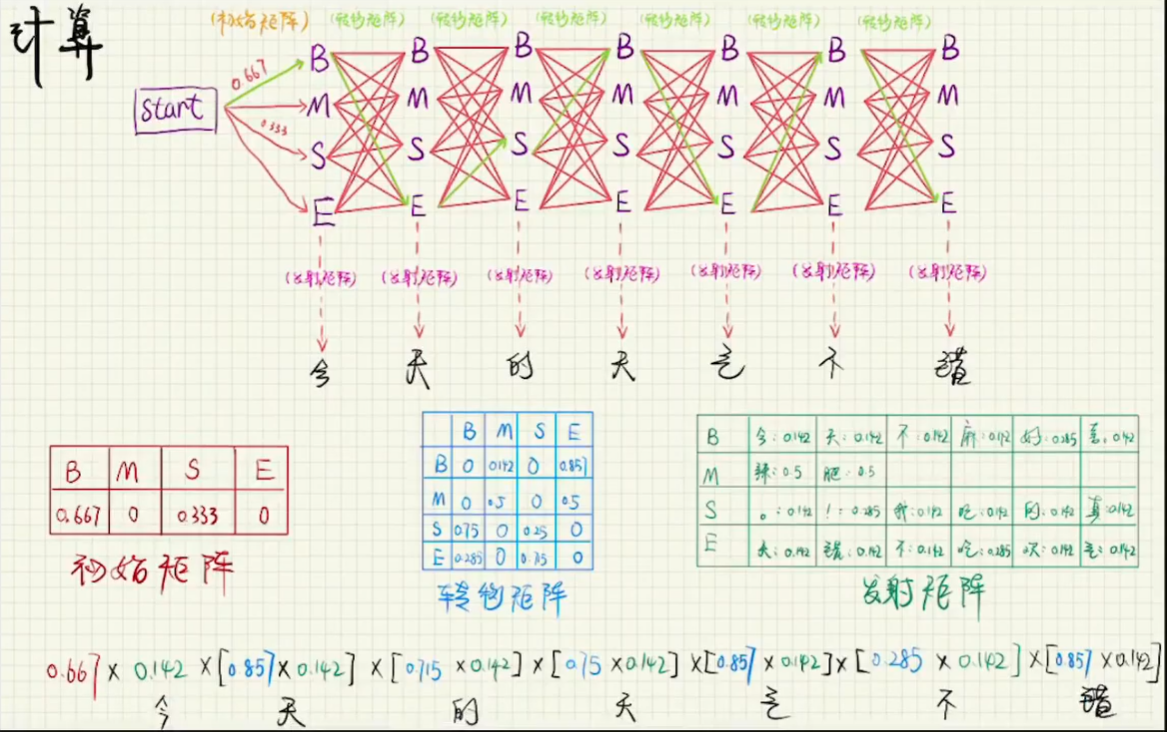
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| B | 今：0.142 | 天：0.142 | 不：0.142 | 麻：0.142 | 好：0.285 | 喜：0.142 |
| M | 辣：0.5 | 肥：0.5 |  |  |  |  |
| S | 。：0.142 | ！：0.285 | 我：0.142 | 吃：0.142 | 的：0.142 | 真：0.142 |
| E | 天：0.142 | 错：0.142 | 牛：0.142 | 吃：0.285 | 欢：0.142 | 气：0.142 |

## 预测过程

输入：

今天的天气不错 🡪 BESBEBE 🡪 今天 的 天气 不错





### 维特比算法