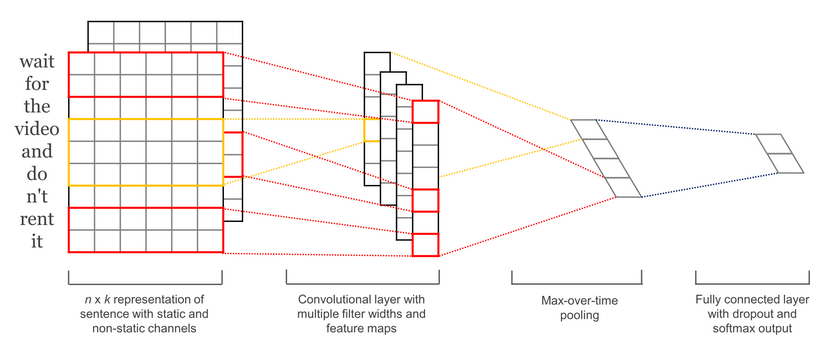
# 深度学习与NLP

## Auto-Encoder

Auto-Encoder作用：数据降噪，数据降维

## CNN4Text

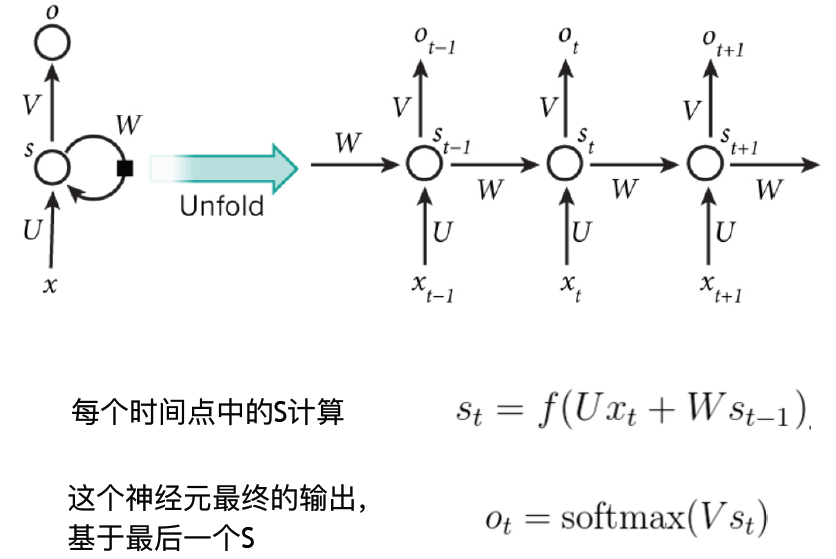


利用CNN处理文本流程如上图，首先不同的单词用vector表示后，形成类似于图片的形式，然后分别经过不同宽度的filter，然后是max pooling layer，最后送入full connected network。

就是利用CNN的filter获取word前后的信息。

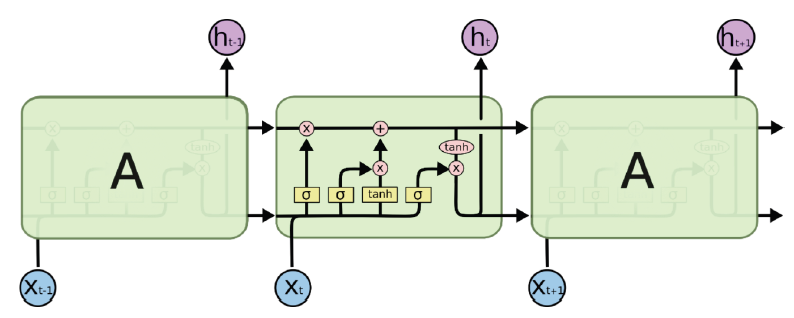
## RNN

RNN的目的就是信息的前后时间关系能够被捕捉到。



RNN的原理如上图，不再过多阐述。

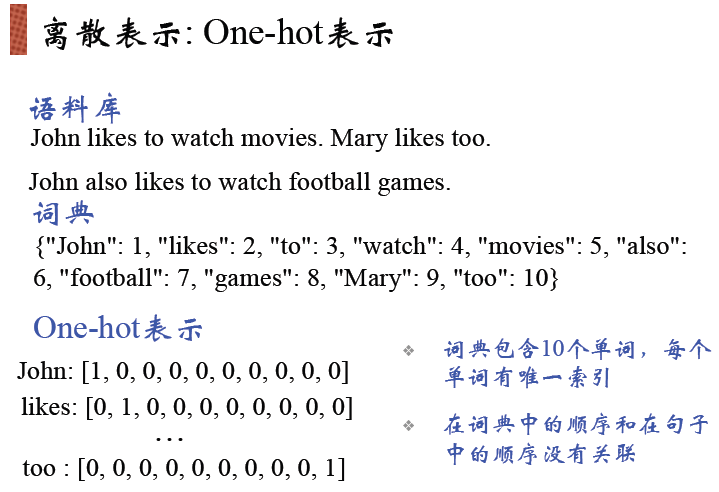
## LSTM



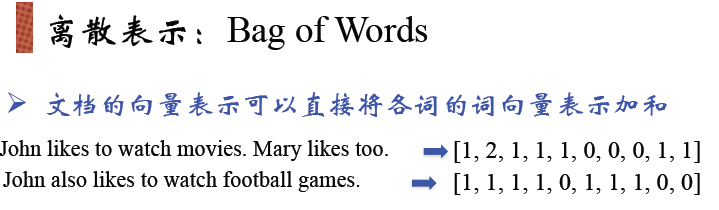
LSTM原理不再详细解释。

# 词向量与相关应用

词编码需要保证词的相似性

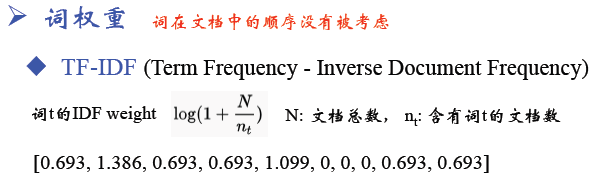


One-hot编码如上，不再多说。

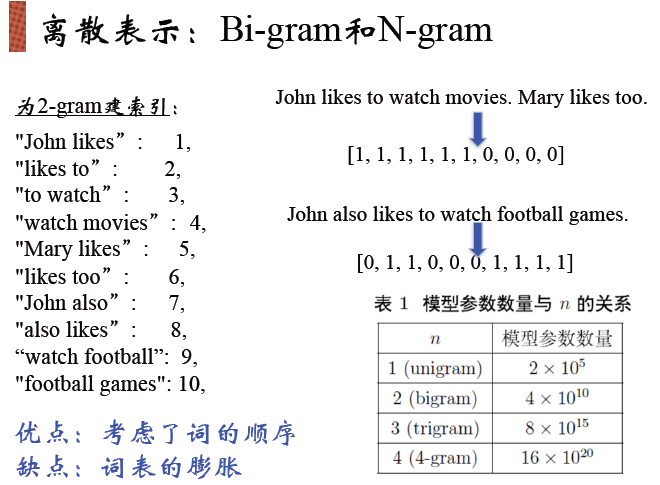


词袋模型如上，也很简单，在每个词one-hot编码产生词向量基础上，文档的向量表示直接将各词的词向量表示加和。也就是按照字典的顺序，依次统计字典中的词在文档出现的次数。

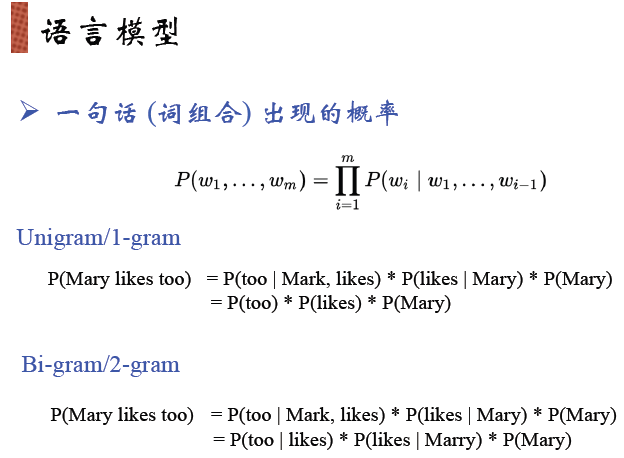
下面tf-idf表示，也比较简单，按照字典顺序，依次算字典中的词的tf-idf值。但是词的顺序信息没有捕捉到。

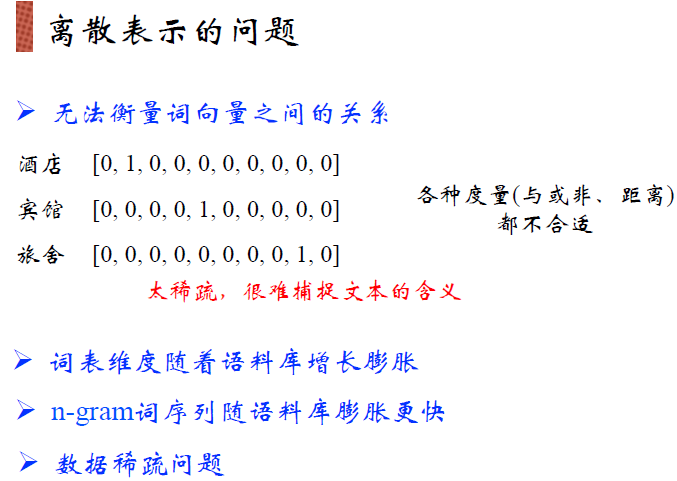


如上，因为词的顺序信息没有被捕捉到，所以引入N-gram如下，但是缺点是词表膨胀迅速。



下面是不同语言模型下概率的计算，也很简答，不多说。





**综上，离散表示的问题如上面总结的，主要是字典随着语料库膨胀的快，数据比较稀疏，还有最重要的一点就是词之间的关联信息没有被捕捉到。**