1题目分析可参考：<https://blog.csdn.net/qq_39591838/article/details/103207138>

拟将计算两个中文段落的相似度，将中文段落根据标点符号分成句子，将句子根据分词划分成词语，根据句子中的词汇相似度计算出句子相似度，然后根据句子相似度计算段落的相似度。

过程：1将句子分词，去掉停用词，获得词语集合S1，S2。后得到句子的相似度矩阵，其中的元素是两个句子中对应词汇的相似度。

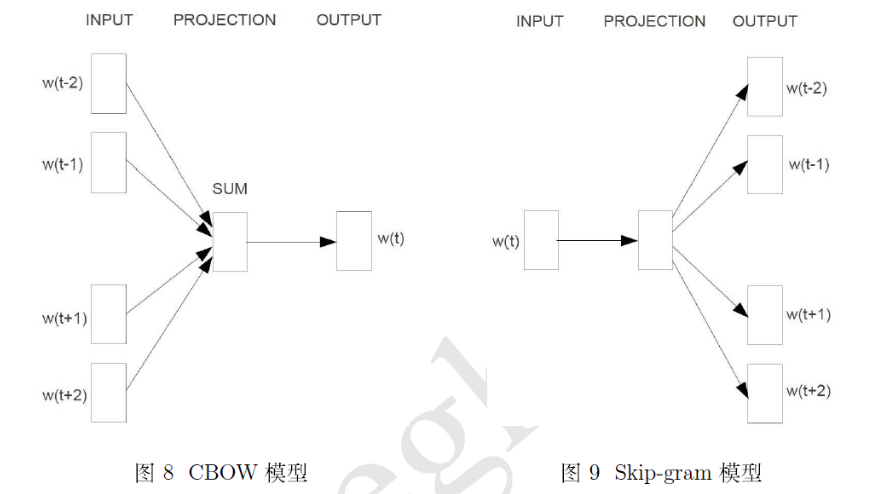
2计算句子相似度：取矩阵最大值放入序列，后删除该值所在的行列，重复取最大值，得到最大值序列maxMS，对maxMS求和并除以S1，S2的词语长度的较大值得到句子相似度。

3计算词语相似度：gensim包（效果不理想）

问题3：代码再细看

2较详细的了解word2vec原理：<https://blog.csdn.net/itplus/article/details/37969519>

词向量定义：所谓词向量就是通过神经网络来训练语言模型，并在训练过程钟生成一组向量，这组向量将每个词表示为一个n维向量。举个例子，假如我们要将"北京"表示为一个2维向量，可能的一种结果如 北京=（1.1,2.2）,在这里北京这个词就被表示为一个2维的向量。但是除了将词表示为向量以外，词向量还要保证语义相近的词在词向量表示方法中的空间距离应当是相近的。比如 '中国' - '北京' ≈ '英国' - '伦敦' 。上述条件可在下列词向量分布时满足，'北京'=（1.1,2.2），'中国'=（1.2,2.3），'伦敦'=（1.5,2.4），'英国'=(1.6,2.5)。

问题1：统计语言模型：计算一个句子的概率模型，通常基于一个语料库来构建CBOW：

Skip-gram：

3：CNN在NLP的实践：

1词向量解释。2词向量与神经网络结合的概念理解

3神经网络结构设计：

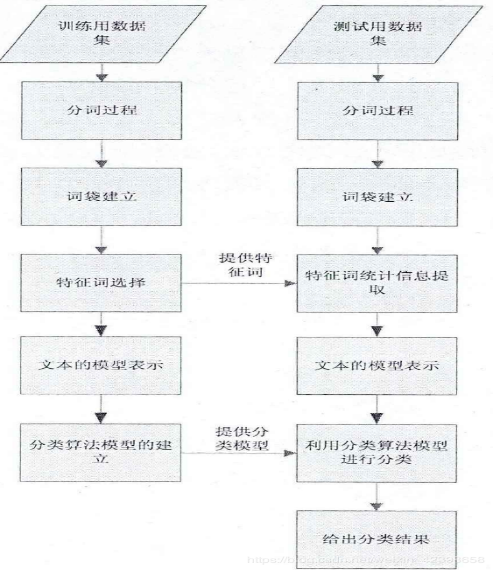
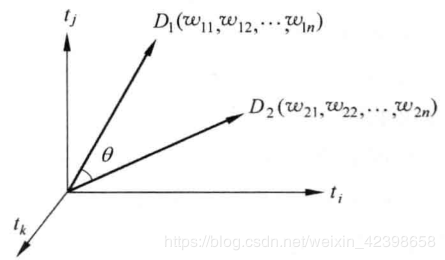
步骤：（1）、文本预处理

3.1：一、先分词（文本转换成词的序列），二、再将词序列换成以词编号（根据词表）为元素的序列。三、将词的编号序列中的每个元素展开为词向量的形式

4、VSM：<https://blog.csdn.net/weixin_42398658/article/details/85063004>

以文本分类任务为例分析向量空间模型VSM，其中主要有关VSM的内容是文本模型表示

定义：文档：文字片段；特征项：VSM最小单元，文档内容被看成它含有特征项组成的集合，表示为D(t1,t2,t3,……tn), tk是特征项。项的权重：每个项都有一个权重表示其在文档中的重要程度（计算方法不同），D可表示为,简记D(w1,w2,…,wn),其中wk是tk的特征项。

给定一个文档D(t1,w1;t2,w2;……tn,wn)，D符合：（1）各个特征项互异；（2）各个特征项无先后关系（不考虑文档结构）**在以上两个约定下，可以把特征项**t_1,t_2,...,t_n**看成一个n维坐标系，而权重**w_1,w_2....,w_n**为相应的坐标值，因此，一个 文本就表示为维空间中的一个向量，我们称**D=D(t_1,w_1;t_2,w_2;...;t_n,w_n)**为文本D的向量或向量空间模型，如下**

其中也提到用向量特征空间中用向量表示的文档的余弦相似度表示相似系数

下面两个是一些文本处理方法，可以看看

文本特征选择方法（可以看看）：DF（这个词在文档中的数量），IG（特征出现在文本后信息熵的差），CHI（和相似度无关，特征项和类别的关联程度）,互信息法（和CHI差不多）

特征权重算法（可以看看）：布尔权重（根据出现次数赋予权重），绝对词频权重（布尔进化版，减少无意义经常出现的词），IDF（特征出现次数越小，权重越大），TFIDF（一个特征在一篇文档中出现次数越多越重要，在所有文档中出现次数越多越不重要）

5、分词：<https://blog.csdn.net/weixin_42398658/article/details/85233753>

<https://blog.csdn.net/weixin_42398658/article/details/85037693>

1. 基于字符串匹配（机械式分词法，基于字典的）按照一定策略将汉字串与机器词典的词条进行匹配，分

（1）、逐词匹配：将词典的所有词按长到短，在文章中逐个搜索，效率低

（2）、正向最大匹配法：假定分词词典中的最长词有i个汉字字符，则用被处理文档的当前字串中的前i个字作为匹配字段，查找字典。若字典中存在这样的一个i字词，则匹配成功，匹配字段被作为一个词切分出来。如果词典中找不到这样的一个i字词，则匹配失败，将匹配字段中的最后一个字去掉，对剩下的字串重新进行匹配处理……如此进行下去，直到匹配成功，即切分出一个词或剩余字串的长度为零为止。这样就完成了一轮匹配，然后取下一个 i 字字串进行匹配处理，直到文档被扫描完为止。

（3）、逆向最大匹配法：与正向类似，但是是从后开始，词典与正向不同，每次匹配不对则切掉前面的子（在实际处理时，先将文档进行倒排处理，生成逆序文档。然后，根据逆序词典，对逆序文档用正向最大匹配法处理即可。）逆向最大匹配法比正向最大匹配法的误差要小。

（4）、双向匹配：结合正向和逆向，先切分成句子（根据标点），对（逐个）句子用正向和逆向分别切，取将句子切割成最小集的结果，简单易实现，缺点匹配速度慢，存在歧义。

二、基于统计的分成方法：通过对精心准备的中文语料中相邻共现的各个字的组合的频度进行统计计算不同字词的共现信息。根据两个字的统计信息计算两个汉字的相邻共现概率，体现中文环境下汉字之间结合紧密程度。当紧密程度高于某一个阈值时便可认为此字组可能构成一个词。（具体下面三个方法讲的不详）

（1）、N元文法模型：未提

（2）、HMM（隐马克夫）模型：

（3）、最大熵模型：

三、（不详）基于理解的分词（基于人工智能的分词方法）模拟人脑对语言和句子的理解，达到识别词汇单元的效果。其基本模式是基于句法、语法分析，并结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统和调度系统。在调度系统的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断：主要两种方法

（1）、专家系统分词法：

（2）、神经系统分词法

第一篇归纳：难点：未登录词（不常见的词）识别和歧义问题

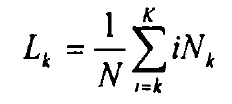
自动分词的评判标准：分词的正确率、切分速度、功能完备性、易扩充性、可维护性和可移植性。

四、由字构词法：

相关文献：http://cpfd.cnki.com.cn/Article/CPFDTOTAL-ZGZR200611002008.htm

（1）、把分词视为字的词位分类问题。即认为每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。其中“字”,也包括标点符号、外文字母、注音符号和阿拉伯数字等。所有这些字符都是由字构词的基本单元。当然,汉字依然是这个构造单元集合中数量最多的一类字符。优点：它能够平衡地看待词表词和未登录词的识别问题。Point：在字标注过程中,所有的字根据预定义的特征进行词位特性的学习,获得一个概率模型,然后在待分字串上,根据字与字之间的结合紧密程度,得到一个词位的分类结果,最后根据词位定义,直接获得最终的分词结果。

（2）、字的词位分类及其所用的基本特征。最常用的特征是字本身以及词位（状态）转移概率。提及CRF（条件随机场）建模的分词系统，只需考虑字特征。 词位学习中确定字特征的主要参数是上下文窗口的宽度,也就是使用距当前字多远的字来作为当前字分类的依据。相关工作表明,使用前后各两个字(即5个字的窗口宽度)是比较理想的，,5字宽的上下文窗口具备了字和词的双重含义。一个确定有效词位标注集的定量标准—平均加权词长。其定义为: 一个确定有效词位标注集的定量标准—平均加权词长。其定义为: 上式中,L_k是i\geq k时的平均加权词长,从是语料中词长为k的词次数,N_k是语料中出现过的最大词长,N是语料库的总词次数。如果k=l,那么Ll,代表整个语料的平均词长。



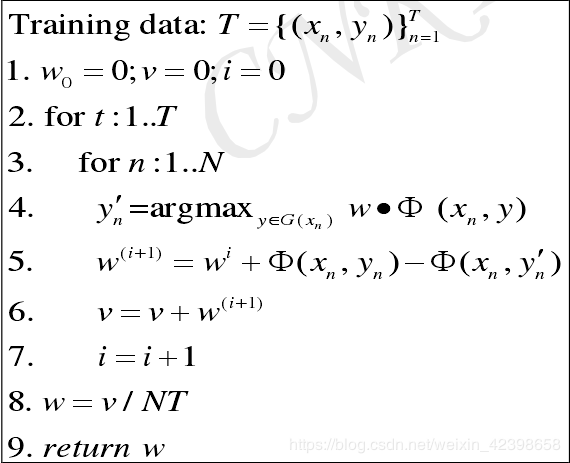
五、基于词感知机算法的汉语分词法： 基本思路：对于任意给定的一个输人句子，解码器每次读一个字，生成所有的候选词。生成候选词的方式有两种：

（1）作为上一个候选词的末尾，与上一个候选词组合成一个新的候选词

（2）作为下一个候选词的开始。

在解码的过程中，解码器维持两个列表：源列表和目标列表。开始时，两个列表都为空。解码器每读人一个字，就与源列表中的每个候选组合生成两个新的候选（合并为一个新的词或者作为下一个词的开始），并将新的候选词放人目标列表。当源列表中的候选都处理完成之后，将目标列表中的所有候选复 制到源列表中，并清空目标列表。然后，读人下一个字，如此循环往复直到句子结束。最后，从源列表中可以获取最终的切分结果。

每次使用一个训练实例对模型参数进行更新，在更新参数时每次将需要更新的参数权重加1或者减1。为了防止模型对数据的过拟合，常对参数进行平均化操作，即Average Perceptron算法。

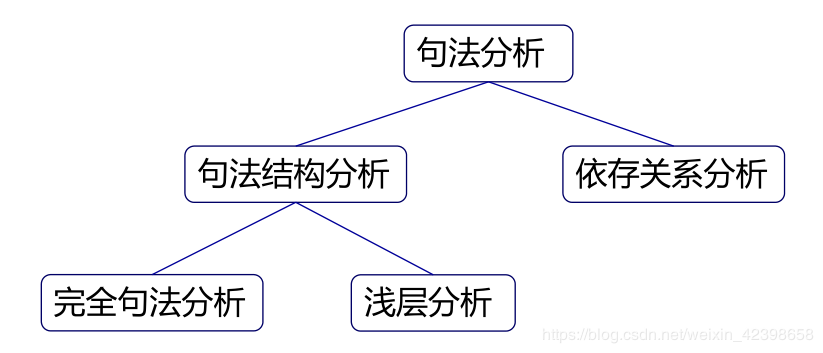


6、句法分析（句子结构）不详细：<https://blog.csdn.net/weixin_42398658/article/details/85223365>

一、问题：（1）语句歧义；

（2）汉语搜索量太大；

二、句法结构：

不同的汉语句法结构，会有不同的句法计算方式。短语结构语法和依存关系语法是现在常见的两种语法关系。

1. 短语结构语法：短语结构语法拥有不同的层级，他们之间都是阐述了语法、语言和自动机之间的关系。短语结构语法呈现一个树分类关系，句法根据一定的规则进行转换分析。每一个词的转换都是需要按照设定的树值规则进行目的性的转换。
2. 依存语法（从属关系语法）：依存的语法句式结构表示的一个依存的关系，并不是一个句法树。每一个依存都是由句子中的支配词和从属词构成，没有规定每一个语句的定义，也没有明确的规则来对依存的关系进行进一步的确定标记。

PCFG（基于概率的上下文无关语法）：可以直接统计语言学中词与词、词与词组以及词组与词组规约信息，并且可以统计由语法规则生成给定句子的概率、一个给定句子最可能的分析、以及由语法规则生成前缀和后缀的概率等等

PCFG三个假设：位置无关性假设（也就是说标注（如NP）只与其管辖的词（狗）有关，而与该词在句子中所处的位置无关。）；上下文无关假设（子结点概率与不受子结点管辖的其他符号串无关，即



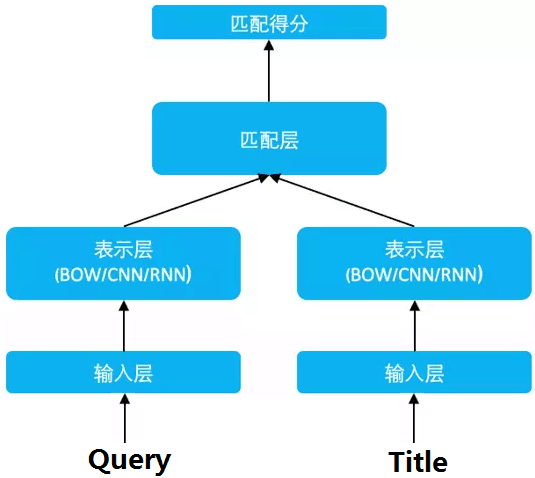
）；祖先节点无关性假设（子结点概率与导出该结点的所有祖先结点的概率无关。）。

HNC理论？

7、深度学习：https://blog.csdn.net/u013074302/article/details/76422551

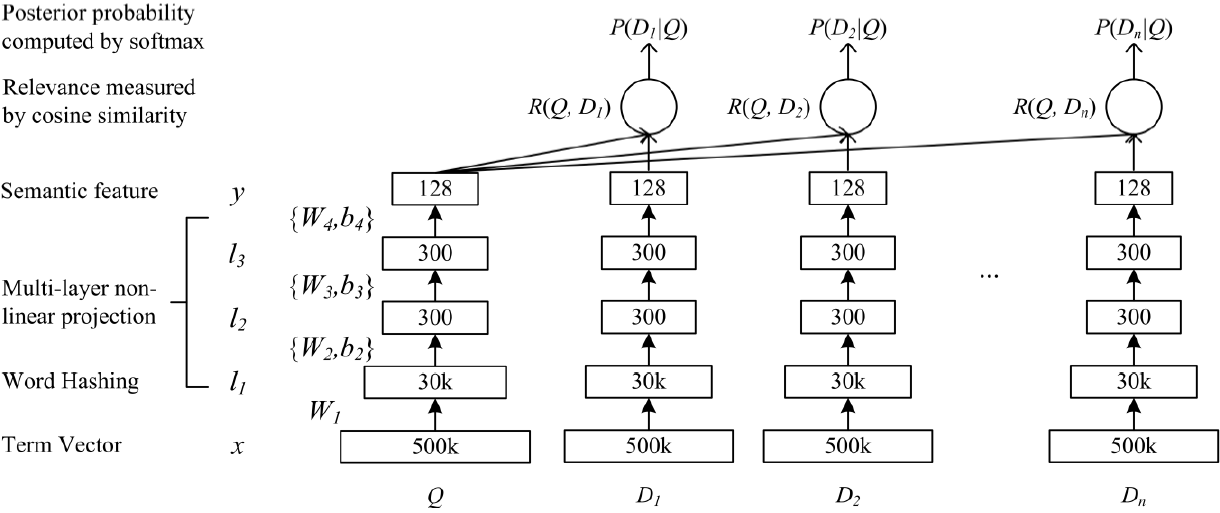
一、DSSM:

Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]// ACM International Conference on Conference on Information & Knowledge Management. ACM, 2013:2333-2338.

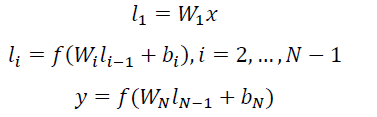
用 DNN 把 Query 和 Title 表达为低纬语义向量，并通过 cosine 距离来计算两个语义向量的距离，最终训练出语义相似度模型。该模型既可以用来**预测两个句子的语义相似度**，又可以**获得某句子的低纬语义向量表达**。分为三层

1输入层：输入，处理文本，分词，转为字向量/词向量输入

2表示层:DSSM中使用词袋（Bag of words）会损失位置信息。后接一个多层的神经网络。



用 Wi 表示第 i 层的权值矩阵，bi 表示第 i 层的 bias 项。则第一隐层向量 l1（300 维），第 i 个隐层向量 li（300 维），输出向量 y（128 维）可以分别表示为：接一个tanh作为隐藏层和输出层之间的激活函数。

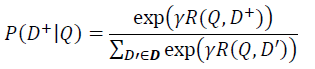


最终输出一个 128 维的低纬语义向量。

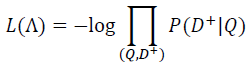
3匹配层。相似性用语义向量的consine距离表示。



通过**softmax 函数**可以把**Query 与正样本 Doc 的语义相似性**转化为一个**后验概率**：其中 r 为 softmax 的平滑因子，D 为 Query 下的正样本，D-为 Query 下的负样本（采取随机负采样），D 为 Query 下的整个样本空间



在训练阶段，通过**极大似然估计**，我们**最小化损失函数**：



**残差**会在表示层的 DNN 中**反向传播**，最终通过**随机梯度下降（SGD）**使模型收敛，得到各网络层的参数{Wi,bi}。

优缺点：用字向量作为输入可以减少切词的依赖和提高模型的范化能力，因为每个汉字所能表达的语义是可以复用的。另一方面，传统的输入层是用 Embedding 的方式（如 Word2Vec 的词向量）或者主题模型的方式（如 LDA 的主题向量）来直接做词的映射，再把各个词的向量累加或者拼接起来，由于Word2Vec 和 LDA 都是无监督的训练，这样会给整个模型引入误差，DSSM 采用统一的有监督训练，不需要在中间过程做无监督模型的映射，因此精准度会比较高。

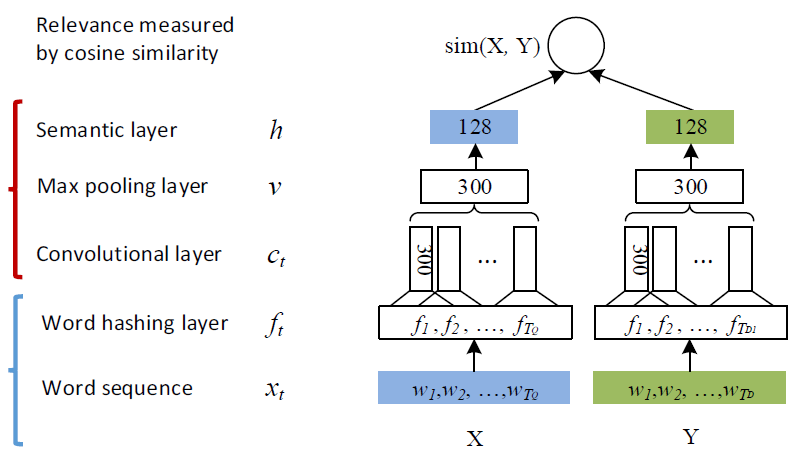
1. CNN-DSSM，与DSSM差别在输入层（中文无差别，还是字向量）和表示层

相关文献：Shen, Yelong, et al. "A latent semantic model with convolutional-pooling structure for information retrieval." Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014

(1)输入层：一样

（2）表示层：由一个卷积神经网络构成。

2.1、（输入/字向量+字哈希+）卷积层+最大池化层+全链接层（tanh），输出128维的语义向量。



（3）匹配层：一样

（4）优缺点：优点：CNN-DSSM 通过卷积层提取了滑动窗口下的上下文信息，又通过池化层提取了全局的上下文信息，上下文信息得到较为有效的保留。

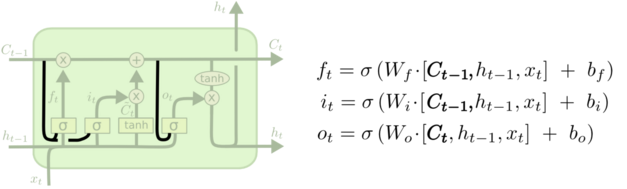
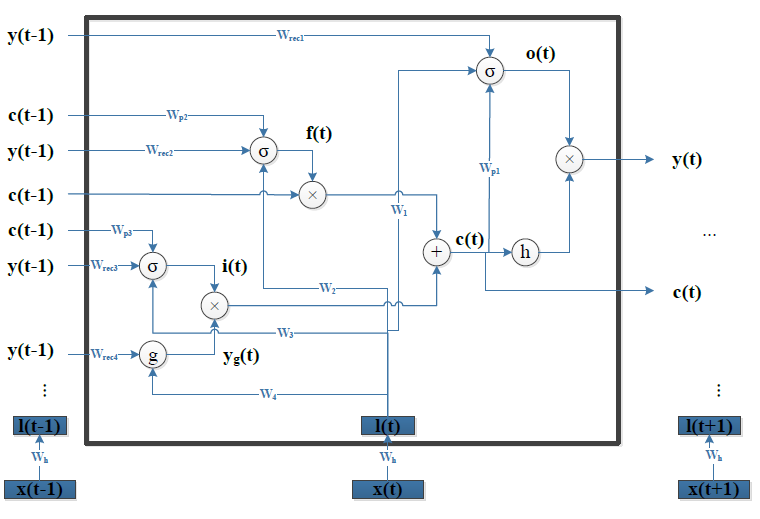
缺点：对于间隔较远的上下文信息，难以有效保留。举个例子，I grew up in France... I speak fluent French，显然 France 和 French 是具有上下文依赖关系的，但是由于 CNN-DSSM 滑动窗口（卷积核）大小的限制，导致无法捕获该上下文信息。（只能俘获较近的上下文信息）

1. LSTM-DSSM（用RNN和LSTM的变形来获取较远的上下文信息）

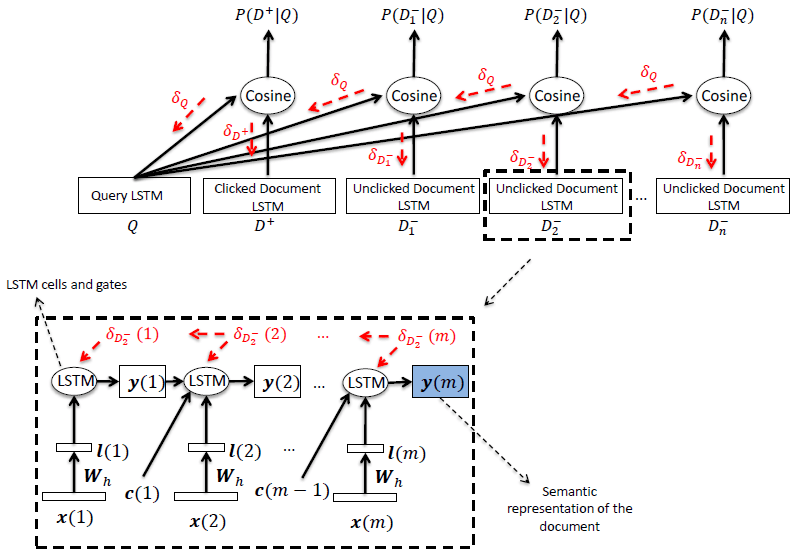
相关文献：Palangi, Hamid, et al. "Semantic modelling with long-short-term memory for information retrieval." arXiv preprint arXiv:1412.6629 2014.

加入peephole的LSTM

传统的 LSTM 中遗忘门、输入门和输出门只用了 h(t-1) 和 xt 来控制门缝的大小，peephole 的意思是说不但要考虑 h(t-1) 和 xt，也要考虑 Ct-1 和 Ct，其中遗忘门和输入门考虑了 Ct-1，而输出门考虑了 Ct



LSTM-DSSM的网络结构（还要再看）



8、文本相似度计算方法汇总博客：https://blog.csdn.net/qq\_16761099/article/details/99288357