摘要：

随着计算机的不断普及发展，编程能力已经成为了人们日常生活中，学生老师科研学习工作中一项不可或缺的能力与工具。编写程序能够帮助我们有效的，随心所欲的利用计算机，极大的提高我们工作学习的效率。然而，也不得不承认编写代码对很多人来说是一项相对困难而且耗时的工作，对于那些并非从事计算机相关行业的人来说，往往希望快速得到解决方案，编写自己需要的程序而不耗费额外而不必要的精力。

以JAVA语言为例，JAVA语言中有丰富的函数库API，可以帮我们解决很多复杂的问题，当有问题需要解决时，找到合适的API可以让我们对问题的解决事半功倍甚至直接解决我们当前的问题。因此，当我们有编程困难时，第一件事便是到网上查询API和code片段，帮助我们打开解决问题的思路。

传统的搜索方法通过自然语言来搜索自然语言，不适合检索代码。因此，我们针对Java编程语言，设计了编程助手，通过自然语言找到与之关联性最高的GitHub代码片段，并检索出需要使用的方法和API, 采用Seq2Seq神经网络模型，分别对代码和自然语言进行encoder,并训练自然语言与代码的对应关系，使我们可以通过自然语言，找到与之相关性最高的5组代码及代码使用的API反馈给用户。

Code\_Searcher采用神经网络训练模型,搜索结果相关性高，且虽然使用神经网络，但将后端代码数据训练与前端输入自然语言的训练分离，实现在搜索准确度高的同时搜索速度快。并为其搭建了Web客户端，方便客户使用。

关键词：编程助手，深度神经网络，GitHub，API，Web应用

一：绪论：

（一）问题定义与背景

我们在编程时，常常有遇到问题不知道该如何去解决的困境，比如在第一次OO作业时，我们可能会想要知道怎么样在java语言中实现删除字符串的某一子串，有类似困难时，我们做的第一件事往往是到网上寻找答案，或者是CSDN，又或者是Stack Overflow，在找到类似的代码使用范例时，我们再把它转换成我们想要的格式。在这个过程中，我们实际面对的是一个调用API的问题：选择什么样的API（Application Programming Interface）来实现功能？又怎样用这些API来实现这个功能？显而易见，如果能有一个根据自然语言的请求来自动生成源代码或API的软件的话，我们的开发将会轻松许多。

（二）相关应用

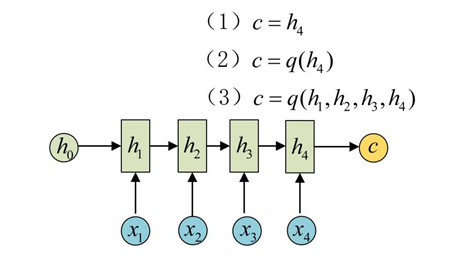
一方面，该应用将方便学习者对于github上的特定工程项目进行搜索，我们可以给搜索者提供所有含有其搜索内容的github上的已知项目的链接。另一方面，该应用将大大降低初学者乃至门外汉使用特定语言编程的门槛，他只需要使用自然语言对自己的需求进行描述，就可以得到该需求在特定语言中的实现方法。最后，该应用也将大大节省开发者的开发时间，提高开发效率，使得开发者可以在最短的时间内找到符合自己需求的代码，并为其提供实现样例。

二：整体设计：

我们的整体设计主要针对代码数据和自然语言数据分别建立神经网络，并将二者分别映射到高维向量空间中去，以达到将两者均转化为数值向量的目的，最后，我们建立两者之间的联系，在抽去的encoder层，即代码层上增加线性映射层，将两个向量集合映射到同一向量空间，并通过对神经网络的训练。具体实现主要分三步：

1. 数据收集：我们从GitHub上获取真实的项目代码，为确保我们代码的有效性，我们对GitHub中低Star的代码进行了筛除，最终得到我们的代码数据库。
2. 得到代码数据库后，为了提取出代码特征，我们采取了seq2seq模型中encoder部分来将代码映射到高维向量空间，以达到获取代码特征的目的。

而在从代码到向量的过程中，我们采用的是深度神经网络中的seq2seq模型的encoder部分。seq2seq是RNN模型的一个变种，RNN（Recurrent Neural Network）是一类专门用于处理序列数据的神经网络，而seq2seq模型的特点则是输入与输出可以不等长，这也就方便了我们实现从代码到向量的转变。



如图所示，h0是我们的输入层，我们将代码从此层输入，经过h1h2h3h4等隐层的处理之后从c层输出，得到的就是我们需要的向量。

1. 根据前馈神经网络模型改进得到skip-gram的神经网络模型，给模型输入一个单词，模型的输出是的上下文wO,1,...,wO,C，上下文window的大小为C，对于输出的这些单词，采用独热码编码方式。下图为神经网络模型：其中向量x代表单词的独热码，对应输出变量{y1,…,yC}，对于在Input layer和Hidden layer之间的权重矩阵W的第i行代表词汇表中的第i个单词的权重需要学习的目标即该矩阵W，其中包含了词汇表内全部单词的权重信息，在模型中，输出的每个单词向量也有一个N×V维输出向量W′。模型中包含了N个节点的隐藏层，隐藏层结点的输入即是输入层所有输入加权后求和，而因为输入向量x采用独热码编码方式，则只有当向量中的元素为非零项时才可以对输入层进行输入，因此隐藏层的输出仅仅与权重矩阵的某一行有关，对于输入向量x，其中存在并且，，存在数学证明：

(1)

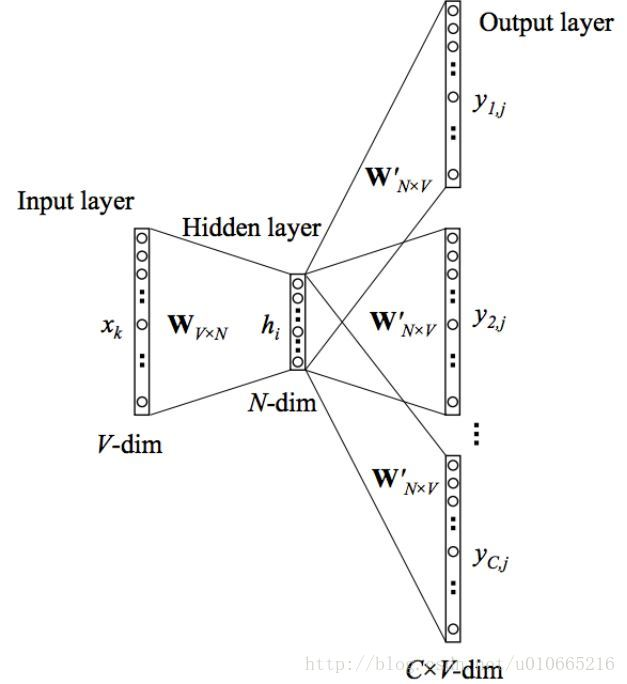
输入使用的独热码，所以此处无需使用函数，模型输出节点C×V的输入由对应输入节点的加权求和计算得出：

uc,j=v′Twjh (2)

上图中我们可以得出输出层中的每个单词的权重都是共享的，则，最终使用softmax函数可得到第C个单词的分布，

 (3)

这就是第C个单词第j个节点的概率。是一堆浮点值，而不是独热码向量。



1. 在第二部抽取的encoder层之上增加线性映射层,通过训练这个网络，最终实现从code向量空间到自然语言向量空间的映射。得到该网络后，我们便将所有代码进行映射并将映射的结果储存到数据库中，最终，当用户进行检索之后，我们可以通过寻找在空间中距离用户搜索内容所对应向量距离最近的结果，并将结果返还给用户。从而实现我们的目的。

三．