# 机器学习工程师纳米学位毕业项目

## **《[句子相似度匹配](https://github.com/udacity/cn-machine-learning/tree/master/quora-question-duplicate" \t "https://classroom.udacity.com/nanodegrees/nd009-cn-advanced/parts/3a867cbf-2876-427e-8e3f-bed2ac7111f5/modules/06b6f1e8-b65f-42a9-ab09-799a23f069ad/lessons/72c558d9-a4b3-4c20-8f20-8023afad4cb6/concepts/_blank)》**开题报告

陈宁

2019年6月25号

目录

**[1. 背景介绍](#_Toc18658_WPSOffice_Level1)** **[2](#_Toc18658_WPSOffice_Level1)**

**[2. 问题](#_Toc26137_WPSOffice_Level1)** **[2](#_Toc26137_WPSOffice_Level1)**

**[3. 数据集](#_Toc31275_WPSOffice_Level1)** **[2](#_Toc31275_WPSOffice_Level1)**

**[4. 解决方案](#_Toc4724_WPSOffice_Level1)** **[2](#_Toc4724_WPSOffice_Level1)**

**[5. 基准模型](#_Toc22797_WPSOffice_Level1)** **[3](#_Toc22797_WPSOffice_Level1)**

**[6. 评价指标](#_Toc28694_WPSOffice_Level1)** **[3](#_Toc28694_WPSOffice_Level1)**

**[7. 项目设计](#_Toc24308_WPSOffice_Level1)** **[3](#_Toc24308_WPSOffice_Level1)**

[7.1：读入数据，数据如下](#_Toc26137_WPSOffice_Level2) [3](#_Toc26137_WPSOffice_Level2)

[7.2：语料编码](#_Toc31275_WPSOffice_Level2) [4](#_Toc31275_WPSOffice_Level2)

[7.3：词语映射](#_Toc4724_WPSOffice_Level2) [4](#_Toc4724_WPSOffice_Level2)

[7.4：搭建一个单层LSTM+全连接层的网络](#_Toc22797_WPSOffice_Level2) [4](#_Toc22797_WPSOffice_Level2)

[7.5：训练网络](#_Toc28694_WPSOffice_Level2) [4](#_Toc28694_WPSOffice_Level2)

[7.6：预测结果](#_Toc24308_WPSOffice_Level2) [4](#_Toc24308_WPSOffice_Level2)

[7.7：保存预测结果](#_Toc22815_WPSOffice_Level2) [4](#_Toc22815_WPSOffice_Level2)

# 背景介绍

[引用](http://www.sohu.com/a/100532291_157627)

让机器来理解人类语言一直都是人工智能的梦想，最先从词到短语到句子， 再到段落到整篇文章。所有的方式都是将字符串转换为向量，最终从数学的角度来理解语义。

# 问题

[引用](https://www.cnblogs.com/huilixieqi/p/6493089.html)

本项目中提供了已经配对好的句子对，需要用已有的数据集进行训练，最终可以预测两个句子的相似性。这属于监督学习类型。可以用编辑距离类似的方法来解决。可以把每一个句子看成是一个向量，再求两个向量之间的关系。这种关系在每一个句子对中几乎都存在。把这种关系保存下来，当有新的一对未知关系的句子对需要检测时，那么就可以先根据数据模型算出他们的关系，再和训练数据的句子对的关系对比，可以得到新的句子对的相似性。

# 数据集

[Quora Querstion Pairs数据集](https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs)是Quora于2017年公开的句子匹配数据集，其通过给定两个句子的一致性标签标注，从而来判断句子是否一致。

Quora 数据集训练集共包含40K的句子对，且其完全来自于Quora网站自身，Quora在发布数据集的同时，在Kaggle平台，发起了[Quora句子相似度匹配大赛](https://www.kaggle.com/c/quora-question-pairs)，共有3307支队伍参加了本次句子相似度匹配大赛，参赛队伍不仅包括来自麻省理工学院、伦敦大学学院、北京大学、清华大学、中科院计算所等高校研究所，也包括了来自微软、Airbnb、IBM等工业界的人员。

本项目使用Kaggle端的[数据集](https://www.kaggle.com/c/quora-question-pairs)，其由Train,Test两部分构成，Train数据集上可以进行验证集划分、建模，Test数据集上可以进行测试。

# 解决方案

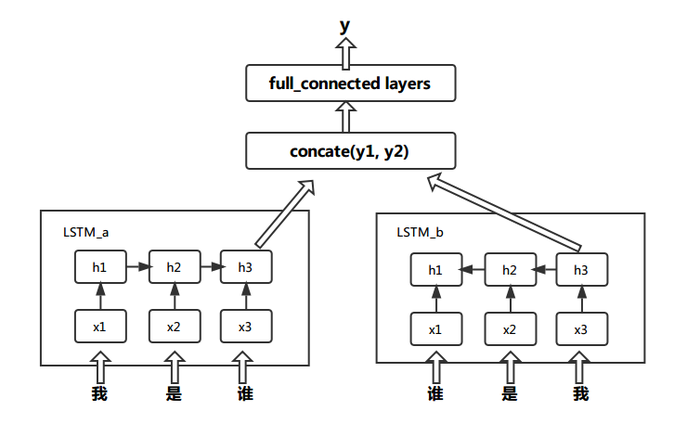
[引用](https://www.meiwen.com.cn/subject/falkrxtx.html)

每一个句子是由单词组成的，可以把所有句子中的单词列出来，进行编码，再用单词的编码对句子进行编码。最终得到是一个个的向量，可以用深度学习的方法从这些向量中找出规律，即模型。该模型就可以用来预测句子的相似性。

# 基准模型

[引用1](https://www.meiwen.com.cn/subject/falkrxtx.html" \l "r1)

本项目中使用的是简单的单层LSTM+全连接层对数据进行训练。如图：



LSTM可以用来解决复杂的数据输入，不会导致梯度消失。

# 评价指标

训练数据集中是一个句子对以及它们是否相似的标签， 相似则值为1，否则为0，所以可以有以下两种方式：

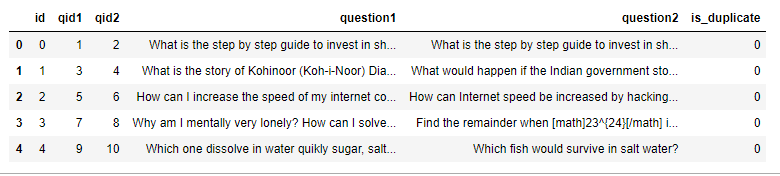
第一种，用训练好的模型去预测训练的所有数据，把得到的值和原来的标签作对比，统计出正确统计的标签的占比，占比赿高，则效时赿好。

第二种，预测所有的测试数据，把预测结果保存到一个csv文件中提交到kaggle, 得到损失度，值赿低，说明效果赿好。

# 项目设计

[引用](https://www.meiwen.com.cn/subject/falkrxtx.html" \l "r1)

7.1：读入数据，数据如下



7.2：语料编码

把训练数据中的所有文本放到一个列表中，输入分词器Tokenizer,

texts = []

tokenizer = Tokenizer(num\_words=MAX\_WORDS, lower=False)

tokenizer.fit\_on\_texts(texts)

7.3：词语映射

用分词器tokenizer把所有单词和它的编码联系起来。、

7.4：搭建一个单层LSTM+全连接层的网络

7.5：训练网络

model.fit(question1\_list, question2\_list, ...)

7.6：预测结果

model.predict

先预测训练集上的数据，和标签作对比，直到准确率达到一个比较满意 的值（0.8以上）时，再预测测试集的数据

7.7：保存预测结果

打印训练集上的准确率，把测试集上的预测结果保到一个csv文件中， 提交到kaggle查看损失度，如果结果不满意，重新优化模型，再次训

练并预测。