# 毕业项目开题报告

## 项目背景

“句子相似度匹配”项目属于自然语言处理（NLP）领域。顾名思义，该领域就是利用计算机处理人类语言（自然语言）。所谓“处理”则包括分析、理解自然语言文本，以及理解自然语言文本后生成自然语言文本，这两者分别对应自然语言处理的两大目标，即自然语言理解和自然语言生成。

自然语言处理最早可追溯到1950年代“图灵测试”的提出，直到1980年代，该领域主要利用人为制定的规则来进行自然语言处理。1980年代末期和1990年代，统计模型被引入该领域，大量基于机器学习方法的自然语言处理技术被提出。随后，基本统计模型的方法得到迅速的发展，词袋模型、TF-IDF、SDV 、LDA等先后被提出。在这之后，随着计算能力的大幅提升，需要强大计算能力支持的深度学习模型在自然语言处理上获得广泛且有效的应用，基于词嵌入模型（word-embedding）的深度学习模型（如CNN、LSTM、Seq2Seq等）。

## 问题概述

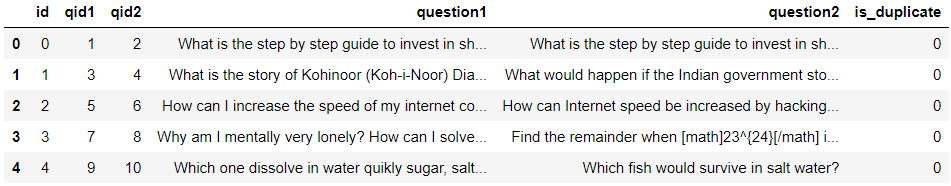
本项目需要解决的问题是判断一对句子的意思是不是相同的。这些句子对来自Quora 2017年在kaggle举办的Quora问题匹配比赛。参赛者需要给出一系列句子对是相同的意思的概率，结果将由log-loss指标评价。

该问题可以看作一个监督学习中的分类问题，通常的解决方案是通过特征工程将文本转变为数值特征，训练出相应的分类器。

在比赛中，参赛队大多采用传统统计模型和深度神经网络结合的方式，例如排名第一的队伍就使用了词袋模型生成的文本向量以及基于文本向量挖掘的特征，输入Siamese 和Attention两种架构的深度神经网络训练分类器。

## 数据与输入

项目的训练数据是约40万对句子对，包含句子对编号，每个句子的编号，句子文本和句子对是否相同的标签。其中，约25万对为不相同句子对、约15万对为相同句子对，相同句子对概率约为40%。前五个句子对数据如下所示：



为了将训练数据输入模型进行训练，需要将文本转换为数值特征。目前的方法有三种，一是是基于词袋模型的TF-IDF、SDV等；二是基于贝叶斯模型的LDA；三是将文本转换为向量的词嵌入方法，例如word2vec、GloVe等。获得文本的数值表示后，可以计算文本的余弦距离、word mover distance等。这些方法得到的特征与句子对标签将作为输入数据，训练分类器。

## 解决方案

在kaggle比赛中，表现优异的队伍最终的模型大多数是大量模型的堆叠，例如XGB、LGBM和神经网络模型堆叠。这样不仅需要对文字特征进行挖掘、还需要训练多个模型，造成很大的工作量。

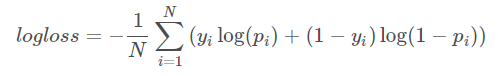
最近两年，一些预训练的深度神经网络语言模型发展迅速，经过微调后在NLP的多个问题都获得了state-of-art的表现。这些预训练语言模型的出现，使得迁移学习在自然语言处理领域有所突破。这样，我们将有可能摆脱繁琐的特征工程并大大较少训练时间。本项目计划采用两种方法。一种是以随机森林方法作为分类器，输入统计类文本特征和使用预训练模型BERT产生的文本向量（Han Xiao开发了bert-as-service，可利用bert将文本转换为固定长度向量）计算的相似度数值。另一种是将句子对输入bert模型并添加分类层，通过微调训练神经网络分类器。

## 基准模型

本项目拟用输入统计类文本特征的随机森林方法作为基准模型。

## 评估指标

项目使用log-loss指标作为评估指标。该指标可用如下方程描述：



(来源: <https://www.kaggle.com/c/bnp-paribas-cardif-claims-management/overview/evaluation>)

其中，N表示数据数量，yi 为i数据的真实标签，pi 为预测的i数据标签为yi 的概率。

该指标对确切的错误分类的惩罚较大。例如，预测所有数据为真实标签的概率为1的log-loss为21.79，预测所有数据为真实标签的概率为0的log-loss为12.75，而预测所有数据为真实标签的概率为0.5的log-loss为0.69）

## 项目设计

项目的工作流程计划如下：

1. 对数据进行预处理，例如分词、大小写替换、lemmatization、去除stop-words等。由于这些处理实际上都丢弃了部分信息，因此需要同时使用未经过这些处理的数据。
2. 利用数据生成特征，包括统计类特征、bert生成的特征。
3. 将统计特征输入随机森林算法训练基准模型M\_bm
4. 将数据输入bert生成相似度分数，同时将统计特征和相似度分数输入随机森林算法训练模型M\_bert
5. 将数据输入bert进行微调，训练神经网络分类器M\_nn
6. 对比三个模型的表现，决定堆叠模型或只使用其中一个。

## 参考资料

1. Natural language processing<https://en.wikipedia.org/wiki/Natural_language_processing>
2. 自然语言处理-简介 <https://www.zhihu.com/topic/19560026/intro>
3. Is That a Duplicate Quora Question? <https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur/>
4. A Gentle Introduction to the Bag-of-Words Model <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>
5. What Are Word Embeddings for Text? <https://machinelearningmastery.com/what-are-word-embeddings/>
6. Gentle Introduction to Statistical Language Modeling and Neural Language Models <https://machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/>
7. 文本主题模型之LDA(一) LDA基础 <https://www.cnblogs.com/pinard/p/6831308.html>
8. 自然语言处理基础与实战（8）- 主题模型LDA理解与应用 <https://www.jianshu.com/p/74ec7d5f6821>
9. 从0到1，了解NLP中的文本相似度 <https://www.jianshu.com/p/22afb6d25e74?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation>
10. 1st place solution <https://www.kaggle.com/c/quora-question-pairs/discussion/34355#latest-572705>
11. BNP Paribas Cardif Claims Management, Evaluation <https://www.kaggle.com/c/bnp-paribas-cardif-claims-management/overview/evaluation>
12. Log Loss Evaluation Section Missing. <https://www.kaggle.com/c/quora-question-pairs/discussion/30605#latest-172575>
13. Manhattan LSTM model for text similarity <https://medium.com/@gautam.karmakar/manhattan-lstm-model-for-text-similarity-2351f80d72f1>
14. BERT <https://github.com/google-research/bert>
15. BERT和ULMFIT embedding比较文本分类结果 <https://www.cnblogs.com/demo-deng/p/10797405.html>
16. Bert-as-service <https://github.com/hanxiao/bert-as-service>
17. Predicting Movie Review Sentiment with BERT on TF Hub <https://github.com/google-research/bert/blob/master/predicting_movie_reviews_with_bert_on_tf_hub.ipynb>