Assignment 3-2

Files

root/

config.py: Configuration of the project (variables and hyper-parameters).

preprocessor.py: Preprocess the data set and create train/dev/test sets.

task.py: Desciption: Combine intention module, retrieval module and ranking module for task-oriented dialogue.

data/

Data files. Put the original data files under this directory.

intention/

Data files for the intention module.

retrieval/

Data files for the retrieval module.

ranking/

Data files for the ranking module.

model/

intention/

Models for the intention module.

retrieval/

Models for the retrieval module.

ranking/

Models for the ranking module.

intention/

business.py: Train a fasttext model and use it for classifying the intention of the user's query.

retrieval/

word2vec.py: Train a word2vec model on the JD dataset.

hnsw_faiss.py: HNSW model implemented with Faiss.

hnsw_hnswlib.py: HNSW model implemented with hnswlib.

ranking/

data.py: Data processing for the ranking module.

bm25.py: Train a bm25 model.

train_LM.py: Train tfidf, w2v, fasttext models.

similarity.py: Definition of manual features.

matchnn.py: Definition of matching network using BERT.

matchnn_utils.py: Helper functions for training the matching network.

train_matchnn.py: Train a matching network.

ranker.py: Generating features and train a LightGBM ranker.

result/

Result files.

lib/

bert/

Pretained BERT model, vocab file and config file.

TO-DO list:

辅助文件

1. 排序的训练数据集：ranking_datasets.zip
2. bert的辅助文件：bert.zip

排序模块流程

1. **处理数据**

   数据集在ranking_datasets.zip中给到，其中包含三个文件，数据集是分别是蚂蚁金服提供的花呗客服数据和微众银行提供的微粒贷客服数据，数据都是问题相似度的标注数据，分别包含两个问题，以及表示他们是否相关的标签（0或1）。利用这些数据集我们可以构建一个pointwise的L2R数据集来训练我们的L2R模型（也可以采用负采样的方式来构建pairwise和listwise的数据集）。数据集的描述可查看以下链接：

   - ATEC学习赛：NLP之问题相似度计算
   - CCKS2018 - 微众银行智能客服问句匹配大赛

   数据集需要处理成三个文件train.tsv, dev.tsv和test.tsv，放在data/ranking目录下，每一个文件的格式如下（分隔符使用\t）：

`question1 question2 label`

示例：

`但是没联系我啊 一直在审核中，也没接到电话啊 1`

2. **人工特征**
   - 先运行一下`ranking/bm25.py`和`ranking/train_LM.py`在排序数据集上训练几个模型：TF-IDF、BM25、word2vec、FastText
   - 构建各种相似度特征，详见`ranking/similarity.py`
3. **深度匹配**

   训练一个BERT模型对输入的两个问题做序列相似度的匹配，得到一个相似度的分数。运行`ranking/train_matchnn.py`来训练深度匹配模型。
4. **排序**

   利用前面步骤得到的特征，使用LightGBM来训练一个排序模型。

   运行`ranking/ranker.py`并且参数设置为`RANK(do_train=True)`来训练。

   运行`ranking/ranker.py`并且参数设置为`RANK(do_train=False)`来预测。
5. **整合**

   整合任务型对话模块：先做意图识别，筛选业务性查询；然后对业务性查询进行召回；对召回的结果进行排序。

模块详解: 人工特征

ranking/bm25.py:

任务: 完成BM25类，训练一个bm25模型。

模型保存路径：model/ranking/

ranking/train_LM.py:

任务: 完成Trainer类，训练一个TF-IDF模型、一个word2vec模型和一个FastText模型。

模型保存路径：model/ranking/

ranking/similarity.py:

任务: 完成TextSimilarity类。

目标：实现人工特征的提取。其中各种特征的定义如下，

- Edit Distance
- Longest Commone Subsequence (LCS)
- Euclidean Distance
- Cosine Similarity
- Jaccard Similarity
- Pearson

以上知识点均有在之前的课程中学习过，如果遗忘了可以再复习一下。

主要使用工具：Gensim的models模块

TF-IDF

Word2Vec

FastText

测试：

`python3 bm25.py`和 `python3 train_LM.py`

模块详解: 深度匹配特征

ranking/matchnn.py:

任务: 定义深度匹配模型。

目标：实现3个类 - BertModelTrain, BertModelPredict, MatchingNN。

主要使用工具：huggingface的transformers，请使用BertForNextSentencePrediction

ranking/train_matchnn.py:

任务：利用前面定义的BertModelTrain来训练深度匹配模型。

模型保存路径：model/ranking/

辅助文件所在路径：lib/bert/

辅助文件在bert.zip中给到。

`python3 train_matchnn.py`

模块详解: 排序

ranking/ranker.py:

任务: 完成RANK类。

目标：提取各类特征后训练一个LightGBM，保存，并用来预测。

主要使用工具：

LightGBM

测试：

`python3 ranker.py`

模块详解: 整合

ranking/task.py:

任务: 完成任务型对话的整合。

目标：如排序模块流程5所述

输出文件：

1. result/retrieved.csv

   保存召回结果，即retrieve模块的输出，每个query对应k个召回结果。

   示例：

	query	retrieved
1	我收到商品不知道怎么使用	商家给我说用不了八代CPU
2	我收到商品不知道怎么使用	[SEP]现在不知道什么情况
3	我收到商品不知道怎么使用	看产品以后不知道怎么买，
4	我收到商品不知道怎么使用	拿，还不知道是哪个商品，
5	我收到商品不知道怎么使用	客户哪了解的这么清楚呢?
6	我买的数据线充不进去电	为的手机数据线插不进去呀
7	我买的数据线充不进去电	器数据线都不用寄过去是吗
8	我买的数据线充不进去电	不冲不了电[SEP]接触不良
9	我买的数据线充不进去电	原线的那根一块进主机是吧
10	我买的数据线充不进去电	的数据线[SEP]可以换货不
11	我现在在学校里，地址有变	址是哪里[SEP]无锡市的吗
12	我现在在学校里，地址有变	家[SEP]可以更换成学校么
13	我现在在学校里，地址有变	没写清楚[SEP]现在能补吗
14	我现在在学校里，地址有变	地址的，忘了改[SEP]学校
15	我现在在学校里，地址有变	但是我在公司地址是家里的
16	四川省***	四川省甘孜州九龙县
17	四川省***	兴镇川兴中学对面天天饰品
18	四川省***	都是江苏省
19	四川省***	香格里拉县虎跳峡镇红桥村
20	四川省***	吉林省伊通县
21	[数字x]能用吗	[数字x]G能用吗
22	[数字x]能用吗	数字x]的卷，明天才能用吗
23	[数字x]能用吗	用[数字x]登陆吗
24	[数字x]能用吗	字x]-[数字x]岁的孩子能用吗
25	[数字x]能用吗	数字x]-[数字x]升就可以用吧
26	什么时候活动	什么时候有活动呀
27	什么时候活动	什么时候活动结束啊
28	什么时候活动	什么时候还有活动么
29	什么时候活动	大概什么时候有这个活动
30	什么时候活动	是什么活动

2. result/ranked.csv

   保存排序结果，即利用ranking模块，对result/retrieved.csv中每个query-retrieved pair进行评分，并记录在一个新的column中。

   示例：

	question1	question2	rank_score
1	我收到商品不知道怎么使用	商家给我说用不了八代CPU	0.205844838210126
2	我收到商品不知道怎么使用	[SEP]现在不知道什么情况	-0.22547842341148905
3	我收到商品不知道怎么使用	看产品以后不知道怎么买，	1.3045892287309195
4	我收到商品不知道怎么使用	拿，还不知道是哪个商品，	0.6570986171140716
5	我收到商品不知道怎么使用	客户哪了解的这么清楚呢?	1.3047506771121131
6	我买的数据线充不进去电	为的手机数据线插不进去呀	0.9176868867716358
7	我买的数据线充不进去电	器数据线都不用寄过去是吗	0.3870626220109501
8	我买的数据线充不进去电	不冲不了电[SEP]接触不良	0.21105191736730944
9	我买的数据线充不进去电	原线的那根一块进主机是吧	0.5938415870405287
10	我买的数据线充不进去电	的数据线[SEP]可以换货不	1.1670164508637444
11	我现在在学校里，地址有变	址是哪里[SEP]无锡市的吗	0.019119337828397122
12	我现在在学校里，地址有变	家[SEP]可以更换成学校么	-0.01921356290144695
13	我现在在学校里，地址有变	没写清楚[SEP]现在能补吗	-0.15986953651907948
14	我现在在学校里，地址有变	地址的，忘了改[SEP]学校	0.7326250833300175
15	我现在在学校里，地址有变	但是我在公司地址是家里的	0.02428284465995139
16	四川省***	四川省甘孜州九龙县	0.20477884349152153
17	四川省***	兴镇川兴中学对面天天饰品	-0.46032937701588866
18	四川省***	都是江苏省	-0.3746081946139088
19	四川省***	香格里拉县虎跳峡镇红桥村	-0.8239866831798799
20	四川省***	吉林省伊通县	0.340010289806757
21	[数字x]能用吗	[数字x]G能用吗	3.014067417670856
22	[数字x]能用吗	数字x]的卷，明天才能用吗	0.23276342744621342
23	[数字x]能用吗	用[数字x]登陆吗	2.3463223324644624
24	[数字x]能用吗	字x]-[数字x]岁的孩子能用吗	1.3141794172337609
25	[数字x]能用吗	数字x]-[数字x]升就可以用吧	0.2185670792413913
26	什么时候活动	什么时候有活动呀	1.969625271779632
27	什么时候活动	什么时候活动结束啊	1.741989210815487
28	什么时候活动	什么时候还有活动么	1.427171535767241
29	什么时候活动	大概什么时候有这个活动	1.3502727488702886
30	什么时候活动	是什么活动	0.8193633903964047