《互联网数据挖掘》项目作业

题目:

基于文档的中文自动问答评测

任务:

基于文档的问答，针对问题从给定文档中选择包含答案的句子。该任务需要建立一个模型，其可以从问题的给定文档句子判定是否可能为答案。我们将为每个问题提供一组文档句子。允许使用其他资源用于训练必要的模型，例如句子匹配模型。

数据:

我们将提供训练集（264416 条数据）、验证集（39997 条数据）和一个测试工具。最终的测试集将在最后一段时间内开放。训练集中的一个例子如下所示:

```
俄罗斯贝加尔湖的面积有多大？ \t 贝加尔湖，中国古代称为北海，位于俄罗斯西伯利亚的南部。 \t 0
俄罗斯贝加尔湖的面积有多大？ \t 贝加尔湖是世界上最深，容量最大的淡水湖。 \t 0
俄罗斯贝加尔湖的面积有多大？ \t 贝加尔湖贝加尔湖是世界上最深和蓄水量最大的淡水湖。 \t 0
俄罗斯贝加尔湖的面积有多大？ \t 它位于布里亚特共和国 (Buryatiya) 和伊尔库茨克州 (Irkutsk) 境内。 \t 0
俄罗斯贝加尔湖的面积有多大？ \t 湖型狭长弯曲，宛如一弯新月，所以又有“月亮湖”之称。 \t 0
俄罗斯贝加尔湖的面积有多大？ \t 贝加尔湖长 636 公里，平均宽 48 公里，最宽 79.4 公里，面积 3.15 万平方公里。 \t 1
俄罗斯贝加尔湖的面积有多大？ \t 贝加尔湖湖水澄澈清冽，且稳定透明（透明度达 40.8 米），为世界第二。 \t 0
```

其数据格式是:

训练集/验证集: 提供问题（第 1 列），文档句子（第 2 列）和他

们的标准答案（第 3 列）。 这三列将用符号'\ t'分隔。对于标准答案，如果文档句子是问题的正确答案，则其为 1，否则其答案为 0。

测试集：在验证集中，将仅提供问题及其文档句子，需要同学们用自己的模型为每个句子预测答案（以小数表示，范围为[0，1], 1 表示最有可能包含答案）。测试数据会在作业截止日期前发布，各小组运行出结果发送到作业邮箱。

要求:

1. 自由分组，建议且最多四人一组。
    a) 没能成功组队的同学将个人信息发给助教
2. 方法要求
    a) 不限制开发环境、算法
    b) 不得人工修改计算/标注结果
    c) 会根据历届学生代码/网上代码查重，严禁使用以往学生的代码，严禁照搬开源代码。
    d) 鼓励大家多讨论交流，多思考，大开脑洞，通过这个作业提高自身能力。成果突出/模型创新型强的有可能建议发 paper。
3. 评测
    a) Mean Reciprocal Rank(MAP)

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AveP(C_i, A_i)$$

$AveP(C,A) = \frac{\sum_{k=1}^{n}(P(k)\cdot rel(k))}{min(m,n)}$ denotes the average precision. k is the rank in the sequence of retrieved answer sentences. m is the number of correct answer sentences. n is the number of retrieved answer sentences. If $min(m,n)$ is 0, $AveP(C,A)$ is set to 0. $P(k)$ is the precision at cut-off k in the list. $rel(k)$ is an indicator function equaling 1 if the item at rank k is an answer sentence, and 0 otherwise.

b) Mean Reciprocal Rank (MRR)

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

$|Q|$ denotes the total number of questions in the evaluation set, $rank_i$ denotes the position of the first correct answer in the generated answer set C_i for the i^{th} question Q_i. If C_i doesn't overlap with the golden answers A_i for Q_i, $\frac{1}{rank_i}$ is set to 0.

重要日期:

1. 上报组队信息： 2018.12.04-24:00
2. 测试集发布时间：2018.12.21-12:00
3. 提交截至时间：2018.12.23-24:00

"以上以北京时间为准。"

提交材料:

1. 测试集上的结果：

   提交结果文件应遵循以下格式：每行仅包含一个分数，表示问题与同一行的文档句子之间的相关性分数。 这些分数将用于评估工具包对给定问题的所有答案句子进行排名。

```
0.2343556
0.3434554
0.5634232
0.2324467
0.1283477
1.2384834
0.4754545
```

请仔细检查，提交结果文件中的行数应与测试集文件中的行数相同。

2. 源代码
3. 说明文档
   a) 作者信息
   b) 分工情况
   c) 编译/运行环境
   d) 系统架构/关键技术
   e) 使用的方法/资源
      i. 给出必要的计算公式
   f) 在验证集上的 MAP,MRR 结果。
   g) 参考文献
   h) A4：5-7 页

提交方式:

1. 所有材料打包发送至邮箱：webdatamining18@sina.com
2. 提交文件命名格式为：

   姓名+学号+第三次作业.rar|zip

作业提示:

1. 可以使用任何数据资源来训练必要的模型, 如 paraphrasing model, sentence matching model 等。
2. 可以参考关于 DBQA 的已有算法。