drift 团队代码说明文档

1. 词向量

word2vec 训练。我们采用了训练集的前一亿、最后一亿还有两个测试集训练 word2vec，采用增量训练的方式，每次训练 1000 万或者两千万数据。维度设置为 300，window 设置为 5，iter 为 10。

2. LightGBM 模型

我们三个人分别按照各自思路提取特征训练 lightgbm 模型。

1) 使用最后一亿数据，提取 count、unique、length 等统计特征，提取 embedding 的余弦、欧式相似度等特征，共 25 维。使用 lightgbm 模型训练 25 维特征，在 a 榜测试集上成绩是 0.591+。
2) 使用前 5 千万数据，提取长度、计数等统计特征，query 和 title 的最长子串、杰卡德系数等文本相似度特征，以及余弦距离、曼哈顿距离、canberra 距离、词移距离等距离特征，共 28 维。使用 lightgbm 训练，a 榜成绩 0.590.
3) 训练集使用前 5 千万数据和测试集提取特征，包括文本特征、统计特征、语义特征以及 query-title 交集相关特征共 17 维特征，在 a 榜测试集上成绩是 0.582+。

3. NN 模型

NN 模型共有 3 个，包括：

1) TextCNN 模型，我们分别把 query 和 title 经过共享权重的 TextCNN 网络，然后分别取 Maxpooling 和 AveragePooling，对两部分特征取差和积操作以便模型更快的收敛，最后和人工提取特征合并，最终经过分类网络进行分类。我们训练了 1 轮，在 a 榜测试集成绩为 0.603+；
2) BiLSTM+FastText 模型，我们使用双层 BiLSTM 提取 query 和 title 深层次的语意信息，将 Embedding 层和两个 BiLSTM 层合并起来作为不同层次的文本语意特征，考虑到模型训练速度问题，我们采用了 Fasttext 的思想，直接做 Pooling 操作，然后做差积操作，合并经过 FC 层的人工特征，经过分类网络分类，训练 1 轮，线上成绩大概在 0.603-0.605 之间

（没有线上测过）。

3) BiLSTM+TextCNN 模型。使用 Bilstm 模型提取深层次的语意特征，再使用 TextCNN 模型进一步提取特征，最终联合人工特征，经过分类网络进行二分类。我们训练了 3 轮，线上成绩为 0.610+。

4. 模型融合方案。

我们采用线性加权的融合方式，首先以 442 的比例融合三个 lightgbm 模型，复赛 a 榜成绩大概为 0.594-0.595（没有线上提交）；以 55 比例融合 TextCNN 模型和 BiLSTM+FastText 模型（线上成绩没提交过）；然后以 64 比例融合上面的 NN 融合模型和 lightgbm 融合模型，线上成绩为 0.614+。最后以 64 比例融合 0.614 模型和 BiLSTM+TextCNN 模型，a 榜线上成绩为 0.619+。这是我们 b 榜最好成绩的提交方案，b 榜成绩为 0.645+。

5、若有问题，请随时联系郇帅（18810310683），谢谢。