股票相关性V4-实验思路

> idears for stock relevance

数据准备

- 数据选择
  1. 筛选备选股票4000支，每支股票选择出高相关的高质量咨询文章100篇作为数据备选
  2. 对于每支股票，提取各个股票的特征用于下一步计算
- 数据预处理
  1. 清洗数据，过滤掉标点符号等无用词汇
  2. 着重处理歧义相关股票涉及的文章，留待之后处理
- 数据标注
  1. 对于每篇文章标注高相关的股票名称作为正样本
  2. 选取百分之二十的文章下做股票负采样，选取不相关的股票进行标注，加入随机噪声，是的模型更加鲁棒

模型结构介绍

- BERT/ERNIE
  1. 加载训练好的模型参数，利用预训练模型做为输入层，将文章题目和内容分别进行tokenize，输入BERT模型进行处理，获得动态的句向量。
  2. 将股票以及股票特征(行业、产品等)拼合为sequence，过BERT得到句向量作为股票的Embedding。

  > BERT原生模型为768维的句向量，maxlen长度最长为512
  >
  > 在Embedding层之后接torch.nn.Linear()，将文章题目和内容的句向量concatenate之后，过线性层进行维度压缩。

- 打分模块

  编写打分模块，对股票和文章Embedding进行相关性打分，推荐最相关的股票及相关性得分。

- 损失函数

  利用搜索的方法和多标签分类损失binary crossEntrophy()作为模型损失进行训练

- 评价指标

  NDCG、hit@5、hit@10