**数据科学基础—期末报告**

**小组信息:** 共 2 人

191250177 杨骏丰 模型训练，相似度计算，演示视频制作

181250014 陈文龙 数据源处理，相似度计算，期末报告编写

邮箱：[1011921795@qq.com](mailto:1011921795@qq.com)

**研究问题:** 文本匹配

**代码地址：** https://github.com/m124065624/data\_science

**研究背景:** 由于不同地区、不同政府部门对于数据的标准不同，也就导致了同一数据有不同的表现形式，例如在公安部门内数据表现为“姓名”，在民政局内的数据表现为“名称”，都表示为同一个意思，以往的工作是通过人工的手段将这些同义字段关联到事项目录管理系统中的标准字段，比如将 “姓名”，“名称”都统一成“姓名”这个标准数据元，但这种方法耗时耗力，而且往往关联某个标准数据元时需 要依靠外部的辅助信息，例如下列的“设计利用储量”关联到标准数据元字段“矿产设计利用储量”就需要利用额外的“自然资源”这个字段作为判断，即我们的场景任务并不简单的是一对一的匹配，有时候需要涉及到多对一的匹配。

**研究方法:** 首先我们对数据源中的重复数据去重，然后对其分词，分解成词向量，再使用基于词向量的word2vec对数据集进行训练。然后我们从键盘获取输入的待匹配字段，对其进行分词，对分词后的结果分别使用训练好的模型进行文本的相似度计算，利用余弦相似度计算，获取出相似度高的3个词向量，再把该词向量来源的数据元整理出来。接着，我们把所有用词向量关键词算出的相似数据，汇聚出所有可能的匹配结果。对这些结果使用Python标准库中的SentenceMatcher计算出输入字段与可能的匹配结果的相似度，对相似度结果集进行排序，找出相似度最高的K个字段，这K个字段就是我们目标的结果。但是这样用Word2Vec算出的结果可能会导致无法直接命中，因为有可能输出字段分词后的关键词就是数据元本身，甚至输入字段本身就是数据元，Word2Vec是找出关键词在模型中的词向量的最相似结果，这时我们需要在可能的匹配结果前用输入字段的分词后结果找出对应的可能的匹配结果，将所有这些结果最终算出的相似度最高的K个结果就是最终符合我们预期的答案。

**案例分析:** 比如我们需要查询语料库里与“有效期起”、“采矿权人地址”、“不动产单元号”相关的内容，我们就可以先把语料库放入程序的文件夹中，然后运行程序，让程序处理原数据，然后我们输入想查询的内容，“有效期起--采矿权人地址-不动产单元号”，程序就会自动搜索语料库里与这些字词相关的字段，并计算出相似度，这样我们就可以很方便地匹配出关联的同义字段。

**附加点：** 对于动态加载，我们的程序每次运行时都会重新训练模型，若果数据元更新了，我们可以不用修改代码即可动态的适配新增的数据，从而实现动态加载。对于多对一的匹配，我们的方法是把输入的多个字短进行分词，然后每个字段都进行相似度计算，结合计算他们的相似度，将最优结果输出。对于算法设置方面，我们结合了两个算法，包括 word2vec 和 SentenceMatcher，务求能实现百分百的准确率，而经过我们的测试，我们的程序精度可以达到95% 以上。

**当前程序截图：**









