中文信息检索系统

方案设计

系统设计目的

本实验旨在设计一个中文信息检索系统，主要分为三步：预训练、文本表示和分类器。其中，预训练主要包括中文分词和虚词过滤，文本表示主要包括文本特征选择、特征权重计算和编码成向量矩阵，分类器主要包括相似度计算和排序返值。

中文信息检索系统的研究目的和背景是提高中文信息检索的效率和准确性，为人们提供更加便捷、高效的信息获取方式。随着互联网技术的不断发展，人们面临的信息量越来越大，传统的信息检索方法已经无法满足人们的需求。特别是在中文信息检索领域，由于中文的复杂性和语义歧义性，传统的信息检索方法更加难以应对。因此，研究中文信息检索系统具有非常重要的意义。

中文信息检索系统的重点研究内容包括以下几个方面：中文分词、文本表示、查询扩展和检索排序等。

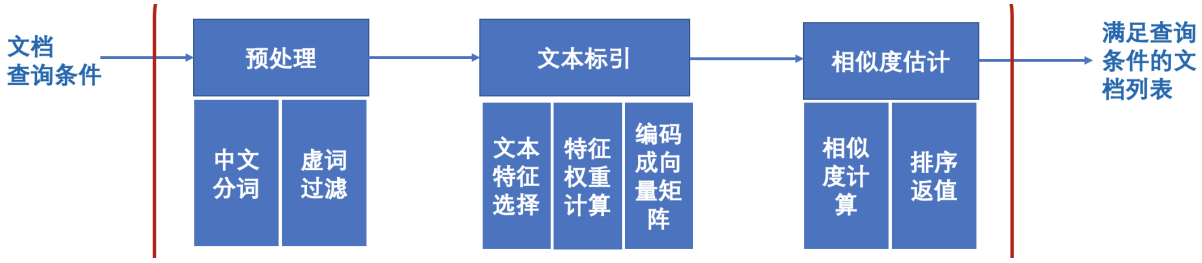
数据集

数据集中包含训练集和测试集，其中训练集26个文件夹共包含4411段信息，每个文件夹下都是某一类的报道，每段信息是一篇中文报道；测试集包含了10句话，需要根据这10句话检索出训练集中最相近的文本。训练集具体情况如下：

```
1：中国女曲能否击败韩国圆梦
2：女垒锦标赛的会徽和吉祥物“丫丫”在南京首次亮相
3：男子十米中国队双保险失灵 加澳抢得冠亚军
4：中国体操在悉尼的辉煌与遗憾
5：棒球联赛第二阶段首回合 沪上闷热击溃猛虎
6：老帅盛赞国奥小将：具备亚洲顶级水平
7：全国山地车冠军赛第二站：马艳萍获女子越野赛桂冠
8：摩托罗拉世界名校赛艇对抗赛
9：网球女单开赛：塞莱斯首战轻取对手
10：NBA：湖人主帅杰克逊支持科比
```

文件夹	主题	文件夹	主题
1	棒球	14	赛艇
2	帆板	15	射击
3	击剑	16	射箭
4	举重	17	手球
5	篮球	18	摔跤
6	垒球	19	跆拳道
7	马术	20	体操
8	排球	21	田径
9	皮划艇	22	网球
10	乒乓球	23	游泳
11	曲棍球	24	羽毛球
12	拳击	25	自行车
13	柔道	26	足球

系统流程



预训练

预训练中主要包括中文分词和虚词过滤

中文分词是将一段中文文本分成一组有意义的词语序列，是中文文本处理的基础，常用的分词工具有jieba、pkuseg、ICTCLAS等。

虚词过滤是去除一些无意义的词语，例如介词、连词、副词等。这样可以减少特征数量，提高模型的训练效率。

经过预处理后，原始文本数据经过清洗、过滤、转换等操作，使得文本可以被计算机更好地理解和处理，提高最后分类的效率和准确性。

文本标引

在文本表示阶段，我们需要将分词结果表示成计算机可处理的形式。具体地，我们需要进行文本特征选择、特征权重计算和编码成向量矩阵三个步骤。对于文本特征选择，我们可以采用TF-IDF算法来选择具有代表性的特征词，从而减少特征维度和计算量。对于特征权重计算，我们可以采用TF-IDF算法或其他算法来计算每个特征词的权重，以便后续的向量编码。最后，我们需要将每个文本的特征向量编码成向量矩阵，以便后续的相似度计算和排序。

相似度估计

在相似度估计阶段，我们需要对输入的查询文本和待检索的文本进行相似度计算和排序，以便返回相似度最高的前K个文本。具体地，我们可以采用余弦相似度来计算查询文本和待检索文本之间的相似度，然后将相似度从大到小排序，取前K个文本作为检索结果。这个阶段的关键是如何高效地计算相似度和排序，以便快速地返回检索结果。

结果分析

根据输入的句子，返回最匹配的三个结果。因为用户常常会由上而下的查看检索结果，因此，要评测检索的效果，对越前面的检索结果赋予更大的权重是更合理的。

同时，还可以通过一些指标评判模型的分类效果，如准确率和召回率等。

最后，输出错误案例，分析可能的分类错误原因，进一步改进模型的算法。

方案实现说明

中文分词

在本实验中，使用jieba库实现了中文分词。它采用了基于字典匹配和规则模板的方法实现中文分词，根据语料库和词频构建字典树，使用前向最大匹配分词。

虚词过滤

在本实验中，使用了停用词表对文本进行虚词过滤。其中停用词表是根据中文停用词表 、哈工大停用词表、百度停用词表和四川大学机器智能实验室停用词库构建的。

据此可以将test文本进行预处理，效果如下：

```
['中国', '女曲', '击败', '韩国', '圆梦']
['女垒', '锦标赛', '会徽', '吉祥物', '丫丫', '南京', '首次', '亮相']
['男子', '十米', '中国队', '双保险', '失灵', '加澳', '抢', '冠亚军']
['中国', '体操', '悉尼', '辉煌', '遗憾']
['棒球', '联赛', '第二阶段', '首回合', '沪', '闷热', '击溃', '猛虎']
['老帅', '盛赞', '国奥', '小将', '具备', '亚洲', '顶级', '水平']
['全国', '山地车', '冠军赛', '第二站', '马艳萍', '获', '女子', '越野赛', '桂冠']
['摩托罗拉', '世界', '名校', '赛艇', '对抗赛']
['网球', '女单', '开赛', '塞莱斯', '首战', '轻取', '对手']
['NBA', '湖人', '主帅', '杰克逊', '支持', '科']
```

文本标引

本实验中，使用了TF-IDF值作为特征选择的依据，这么做的原因是TF-IDF（Term Frequency-Inverse Document Frequency）是一种常用的文本表示方法，用于评估一个词语在文本中的重要程度。其中，TF表示词项频率（Term Frequency），用于衡量一个词在当前文本中出现的频率，TF越高，代表该词在文档中越重要。IDF表示逆文档频率（Inverse Document Frequency），用于衡量一个词的普遍重要性，IDF值越大，代表该词语越重要。具体来说，TF-IDF的计算公式如下：

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$tf\text{-}idf(t,d,D) = tf(t,d) * idf(t,D)$$

相似度估计

相似度计算有多种方法，以下是其中几种常见的方法及其原理：

1. 欧几里得距离（Euclidean Distance）：也称为L2距离，是最常见的距离度量方式之一。欧几里得距离是两个向量之间的直线距离，它的计算方法是将两个向量的每个元素进行差值计算，再对所有差值的平方求和，最后再开方。公式如下:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2. 曼哈顿距离（Manhattan Distance）：也称为L1距离，它是两个向量中每个元素差值的绝对值之和。曼哈顿距离适用于数据分布比较规则的情况，计算方法如下：

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

3. 余弦相似度（Cosine Similarity）：是用于衡量两个向量之间相似性的一种方法。余弦相似度是通过将两个向量的点积除以它们的范数的乘积来计算的。公式如下：

$$\cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$$

4. 皮尔逊相关系数（Pearson Correlation Coefficient）：衡量两个变量之间的线性关系。它的取值范围在-1到1之间，取值越接近1或-1，表示两个变量之间的线性相关性越强，取值为0时表示两个变量之间没有线性关系。计算方法如下：

$$r = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

在本实验中，使用TF-IDF已经将文本统一成向量，因此可以使用余弦相似度计算两个向量之间的相似度。

结果评测

检索信息	得分	准确率	是否查到原文
1: 中国女曲能否击败韩国圆梦	6	100%	否
2: 女垒锦标赛的会徽和吉祥物“丫丫”在南京首次亮相	6	100%	否
3: 男子十米中国队双保险失灵 加澳抢得冠亚军	5	60%	是
4: 中国体操在悉尼的辉煌与遗憾	6	100%	否
5: 棒球联赛第二阶段首回合 沪上闷热击溃猛虎	5	70%	否
6: 老帅盛赞国奥小将：具备亚洲顶级水平	-	-	是
7: 全国山地车冠军赛第二站：马艳萍获女子越野赛桂冠	6	100%	是
8: 摩托罗拉世界名校赛艇对抗赛	6	100%	是
9: 网球女单开赛：塞莱斯首战轻取对手	6	100%	否
10: NBA：湖人主帅杰克逊支持科比	6	100%	否

注：为了相对简单地统计检索效果，本实验制定简单检索得分规则：如果Top 1 match与输入文本相关得3分，Top 2 match与输入文本相关得2分， Top 3 match与输入文本相关得1分。准确度描述检索结果是否与查询结果主题相同。是否查询到原文描述返回的结果中是否包含了检索信息的原文。

本实验中，使用了一个包含了10条检索信息的test.txt文件用来测试检索的效果，这10条信息都是从训练集中原样提取的。从上述表格中可以看出本实验检索结果能够大致反映出查询结果的相关内容，美中不足的是模型有时不能很好的返回查询信息的原文，这是因为本实验将原始数据转化为TF-IDF向量后，能够较好的优化时间和空间复杂度，但是破坏了原文的上下文连接性，导致模型对于原文句子检索出现一定程度的偏差。详细解决方法见错误分析部分。

错误分析

1. 原文检索

   通过系统内置检索(MacOS)，可以发现训练集中的数据是从文本的标题中提取的。其中，启发性的思想是，要想进一步优化中文信息检索系统，需要在爬取的时候对网页不同位置的信息赋予不同大小的权重，再根据权重进行一定程度的信息保留，从而在检索的时候进行优先级的排布，并且能够提升检索的速度。

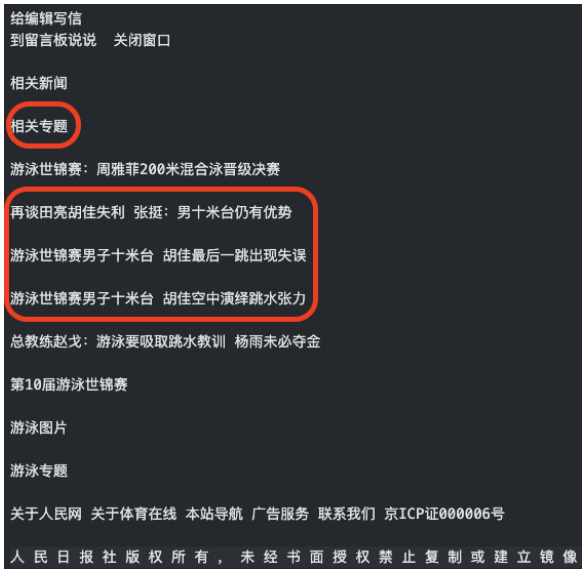
   相比于操作系统级的检索，因为其能够按照特定的方式保存原始数据，所以能够设计出较好的算法，在保留全部原始数据的同时，依然有较快的检索速度和较小的空间存储空间占有；而本实验中，如果将原始数据全部转化为TF-IDF矩阵，则较难查询到原文，比较好的解决方法是在爬取内容的时候，将标题部分等较为重要的部分和其他正文部分分开保存，前者完整保存，后者转化为TF-IDF，在检索的过程中，先将查询内容和重要部分进行匹配，如果出现完全相同的内容，则直接返回完整内容，如果没有完全相同的部分，则使用余弦相似度匹配TF-IDF向量和查询向量。这样做的好处是能够兼顾查询速度、模型存储空间大小和查询准确度。

如下是原始数据和转化为TF-IDF占用内存大小比较：

```
# 使用sys库中的getsizeof()函数获取var_name的存储空间大小
print("原始数据占有空间大小: {} Byte".format(sys.getsizeof(data)))
print("TF-IDF占有空间大小: {} Byte".format(sys.getsizeof(tfidf_matrix)))
✓ 0.0s
原始数据占有空间大小: 38224 Byte
TF-IDF占有空间大小: 64 Byte
```

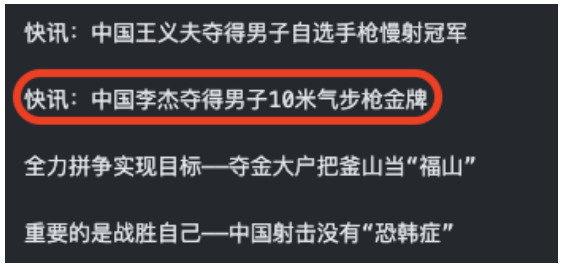
2. 数据清洗

原始数据中存在较多的无关信息，这会对检索结果造成影响，比如当检索“男子十米中国队双保险失灵 加澳抢得冠亚军”，的时候，在正文的结尾部分出现相关专题中含有 “十米台” 的相关信息（如右图所示），但是正文其实与十米台并没有太多的关系，导致检索的结果出现偏差。要解决这个问题，最好是在爬取网页的过程中，自动检索和分离正文部分和正文无关的部分，避免原始数据的污染。



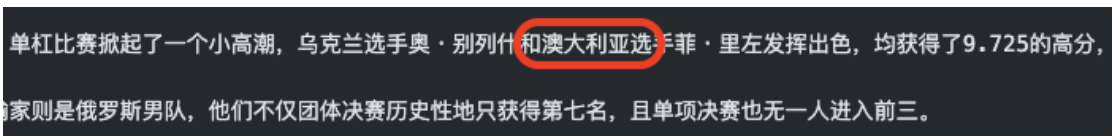
3. 词组歧义

仍然是检索“男子十米中国队双保险失灵 加澳抢得冠亚军”这个例子，这里的十米可以理解成跳水的十米台，也可以理解成射击的10米气步枪，存在词组歧义，要解决这个问题，需要对检索的信息进行进一步的详细说明，提供更多关键词。



4. 鲁棒性

比如当检索“中国体操在悉尼的辉煌与遗憾”的时候，需要在一定程度上考虑地名的包含与被包含关系，当检索“悉尼”的时候，如果返回的结果是澳大利亚，这也是有价值的和可被接受的。同样的，也需要考虑同一性的问题，比如说“京”和“北京”是同一个意思。因此，为了使检索更加精细，可以建立一个知识图谱，提升检索系统的鲁棒性。



5. 专有名词

   在检索的过程中可能会出现一些专有名词，但是在数据预处理的阶段却被错误的分词和过滤，导致最后的检索结果出现偏差。比如，当检索“棒球联赛第二阶段首回合 沪上闷热击溃猛虎”的时候，如果不清楚这里的“猛虎”是指中国棒球联赛队名“北京猛虎”，则很容易在文本分词的时候错误划分。再比如检索“NBA：湖人主帅杰克逊支持科比”的时候，如果不知道“科比”是个人名的话，就很容易出现错误。要解决这个问题，可以在预处理的时候建立专有名词库，在预处理的过程中完整保留专有名词，代码如下：

```
# 添加专有名词
jieba.add_word('人工智能', '自然语言处理', '杭州电子科技大学')
```

小结

在这个信息爆炸的时代，信息检索系统已经成为人们获取信息的重要途径。在本次实验中，本实验成功地设计并实现了一个中文信息检索系统，可以较好地处理各种类型的中文文本，并返回与查询文本相关的文本结果。

为了实现高效的检索，本实验采用了jieba分词库进行中文分词和停用词过滤，使用TF-IDF算法进行文本特征选择和特征权重计算，最终将每个文本的特征向量编码成向量矩阵。通过余弦相似度计算和排序，得到了相似度最高的前K个文本作为检索结果。在测试集上，模型基本能够返回与检索内容相关程度较高的文本，同时针对出现的一些错误案例给出了具体可行的方案。

在未来，随着人工智能和大数据技术的发展，信息检索系统将会变得更加智能化和精准化。本次实验不仅使我们深入理解了信息检索系统的原理和实现方法，还让我们掌握了相关的技术和工具，为我们未来的学习和工作打下了坚实的基础。