实验一描述文档

实验目的

本实验要求以给定的财经新闻数据集为基础，实现一个新闻搜索引擎。对于给定的查询，能够以精确查询或模糊语义匹配的方式返回最相关的一系列新闻文档。

任务要求

本实验要求同时实现`bool检索`和`语义检索`。首先将每一条新闻视作一个文档，进行分词、词根化、去停用词处理。词根化的过程与停用词库不作硬性要求，可以以你认为合适的方式任意选取。完成上述初始化步骤后，你需要：

- 1. 对于经过预处理的文档集合 $D=\{D_1,D_2,...,D_N\}$ ，根据倒排索引算法建立倒排索引表S ，并以合适的方式存储生成的倒排索引文件。
  2. 对于给定的 bool 查询 Q_{bool} （ Q_{bool}的书写规则以上课内容为准），根据你生成的倒排索引表S，返回符合查询规则 Q_{bool} 的文档集合 $D_{\text{bool}} = \{D_1^{\text{bool}}, D_2^{\text{bool}}, ..., D_M^{\text{bool}}\}$ 。
  3. 根据文档集合D，计算每个文档的 `tf-idf` 向量 $T=\{v_1,v_2,...,v_N\}$，并将 T 以矩阵的形式存储。
  4. 对于给定的语义查询 $Q_{se} = \{$ word $_1$, word $_2$, ..., word $_n\}$ ，其中每个 $word_i$ 代表一个查询词， 计算 Q_{se} 的 `tf-idf` 向量v_{qse} ，并根据 v_{qse} 与 T 的相似度返回前10个最相关的文档集合 $D_{se}=\{D_1^{se},D_2^{se},...,D_{10}^{se}\}$ 。

除此之外，可选做的内容包括：

- 1. 对你的倒排索引过程进行时间复杂度或者是空间复杂度的优化。
  2. 采用外部知识库 (例如同义词表) 优化你的索引效果。
  3. 采用 `word2vec` 等其他语义表征方式表征你的查询和文档，并选用合适的案例与 `tf-idf` 的结果进行对比分析。对于实现的优化我们将视优化效果给予酌情加分。
  4. 对于给定查询，返回最相关的10张图片（自行设计方案，图片可从 json 文件中 main_image 字段获取，具体见数据集介绍），并选用合适的案例进行展示。

数据集介绍

```
US Financial News Articles 收集了彭博社（Bloomberg News）、美国消费者新闻与商业频道（CNBC），路透社（Reuters），华尔街日报（WSJ），财富报（Fortune）提供的从2018年1月到5月的财经新闻。
```

本地可用的数据介绍：

- 1. 数据内容为：主zip文件里面包含了5个子文件夹，每个文件夹分别代表2018年1月~5月的财经新闻集，共有 306,242 条新闻。
  2. 每一条新闻是一个 json 文件。在该文件中，包含了文章的来源、发表时间、作者详情以及**与每篇文章相关的图片**。如下图所示：

```
{
  "organizations": [],
  "uuid": "e561082aa5f4e223ee42af771764524beaaedf2a",
  "thread": {
    "social": {
      "gplus": {
        "shares": 0
      },
      "pinterest": {
        "shares": 0
      },
      "vk": {
        "shares": 0
      },
      "linkedin": {
        "shares": 0
      },
      "facebook": {
        "likes": 0,
        "shares": 0,
        "comments": 0
      },
      "stumbledupon": {
        "shares": 0
      }
    },
    "site_full": "www.reuters.com",
    "main_image": "https://s4.reutersmedia.net/resources_v2/images/rcom-default.png",
    "site_section": "http://feeds.reuters.com/reuters/technologyNews?format=xml",
    "section_title": "Reuters: Technology News",
    "url": "https://www.reuters.com/article/us-emirates-cyber-darkmatter/emerging-gulf-state-cyber-security-
powerhouse-growing-rapidly-in-size-revenue-idUSKBN1FL451",
    "country": "US",
    "domain_rank": 408,
    "title": "Emerging Gulf State cyber security powerhouse growing rapidly in size, revenue",
    "performance_score": 0,
    "site": "reuters.com",
    "participants_count": 0,
    "title_full": "",
    "spam_score": 0,
    "site_type": "blogs",
    "published": "2018-02-01T09:03:00.000+02:00",
    "replies_count": 0,
    "uuid": "e561082aa5f4e223ee42af771764524beaaedf2a"
  },
  "author": "",
  "url": "https://www.reuters.com/article/us-emirates-cyber-darkmatter/emerging-gulf-state-cyber-security-
powerhouse-growing-rapidly-in-size-revenue-idUSKBN1FL451",
  "ord_in_thread": 0,
  "title": "Emerging Gulf State cyber security powerhouse growing rapidly in size, revenue",
  "locations": [],
  "entities": {
    "persons": [
      {
        "name": "alexander cornwell",
        "sentiment": "none"
      },
      {
        "name": "darkmatter",
        "sentiment": "none"
      },
      {
        "name": "faisal al-bannai",
        "sentiment": "none"
      },
      {
        "name": "bannai",
        "sentiment": "none"
      }
    ],
    "locations": [
      {
        "name": "uae",
        "sentiment": "none"
      },
      {
        "name": "united arab emirates",
        "sentiment": "none"
      },
      {
        "name": "abu dhabi",
        "sentiment": "none"
      }
    ],
    "organizations": [
      {
        "name": "emerging gulf state",
        "sentiment": "negative"
      },
      {
        "name": "nesa",
        "sentiment": "none"
      },
      {
        "name": "ibm",
        "sentiment": "none"
      },
      },
      {
        "name": "uae",
        "sentiment": "none"
      },
      {
        "name": "national electronic security authority",
        "sentiment": "none"
      },
      {
        "name": "axiom telecom",
        "sentiment": "none"
      },
      {
        "name": "min read  abu dhabi",
        "sentiment": "none"
      },
      {
        "name": "reuters",
        "sentiment": "none"
      },
      {
        "name": "lockheed martin",
        "sentiment": "none"
      }
    ]
  },
  "highlightText": "",
  "language": "english",
  "persons": [],
  "text": "February 1, 2018 / 7:05 AM / Updated 21 minutes ago Emerging Gulf State cyber security powerhouse growing rapidly in size, revenue Alexander Cornwell 3 Min Read \nABU DHABI (Reuters) - A little-known cyber security company in the United Arab Emirates (UAE) recruiting executives who have worked for Western intelligence services is turning over hundreds of millions of dollars a year, largely in contracts with the government, according to its chief executive. \nDarkmatter was founded three years ago in Abu Dhabi, the UAE capital, by CEO Faisal al-Bannai, an Emirati entrepreneur known for setting up regional mobile phone retailer Axiom Telecom. \nThe majority of its work, 80 percent, is with the UAE government and related entities, which has included advising federal cyber security agency National Electronic Security Authority (NESA). \nThat has helped the company, with ambitions to globally compete in the cyber sphere with IBM and Lockheed Martin, to double its revenue each year. \n“Today, we’re talking about hundreds of millions of dollars,” Bannai told Reuters at Darkmatter’s Abu Dhabi headquarters. \nA UAE government representative was not available to comment on the claims. \nDarkmatter has gone on a recruiting spree since it started in late 2014, and has more than tripled its workforce to 650. It has hired executives who have worked at major international companies such as Intel Corporation and BlackBerry, but also some with backgrounds in Western military and intelligence agencies including the U.S. National Security Agency (NSA). \nBannai said Darkmatter is profitable and is providing the UAE government with defensive security tools, but not offensive technology. \n“A massive chunk of what we do with government entities here is ‘how do we strengthen their network to be immune from attacks?’ or at least to recover from an attack,” Bannai said. \nCyber-attacks by hostile governments, militant groups, or by cyber criminals could disrupt key infrastructure such as oil and gas supplies, and desalination plants, which the UAE and other Gulf states depend on. \nTraditionally low-key UAE has become more influential in its foreign policy in recent years, potentially increasing the threat of cyber-attacks. \nIt has intervened in the Yemen civil war, and is taking a leading role in a dispute between some Arab states and Qatar. \nDarkmatter’s relationship with the UAE “is a pure commercial transaction,” Bannai said, unique in a country where nearly all major entities are state-owned or controlled. \n“Definitely, they see a value in having a local partner build these capabilities,” he said. Reporting by Alexander Cornwell; Editing by Stephen Coates",
  "external_links": [],
  "published": "2018-02-01T09:03:00.000+02:00",
  "crawled": "2018-02-01T09:14:57.004+02:00",
  "highlightTitle": ""
}
```

3. 数据集已被上传到睿客网，下载链接为：https://rec.ustc.edu.cn/share/8a34d6a0-1c59-11ec-990f-3984ef409d4c，密码：exp1
4. 查询词表的下载链接为：https://rec.ustc.edu.cn/share/94289250-1d05-11ec-98e2-dfb4810282a9，密码：exp1

提交要求

请以如下文件目录结构组织相关文件结构：

```
exp1/
|----src/
    |----bool_search.{FileSuffix}
    |----semantic_search.{FileSuffix}
|----dataset/
    |----{your dataset}
|----output/
    |----{your output files}
|----实验报告.pdf
|----README
```

其中，各目录/文件具体要求如下：

- `src` 目录下放置你的源代码文件，其中 `bool_search` 为以 bool 检索方式的源代码文件；`semantic_search` 为以语义检索方式的源代码文件； `.{FileSuffix}` 根据你使用的编程语言的文件后缀而决定。如果存在相应的优化，请直接添加在原文件中。
- `dataset` 目录下放置你的数据集文件。在提交时，可以将此文件夹置空。
- `output` 目录放置你的输出文件，包括生成的倒排表，以及经过所有文档的 `tf-idf` 矩阵。如果采用了其他的语义表征方式，也请生成对应的矩阵并在 `README` 中注明。
- `实验报告.pdf` 应包含对你采用的算法以及所作优化的描述，同时请用相关的实验数据证明你的优化是有效的。同时，请在实验报告中展示你认为最具有代表性的运行示例。如果是多人组队，请在实验报告中注明所有组成员的学号和姓名。
- `README` 文件中包含你的源代码的运行环境、编译运行方式，以及对关键函数的说明。同时，对于所作要求之外的文件，也请在 `README` 中注明这些文件的含义。

提交说明

以 PDF 或 DOC 格式提交，实验报告提交文件及邮件标题命名格式统一为 “学号1_姓名1_学号2_姓名2_web实验一”。

- 例如：“PB19111888_法外狂徒张三_PB19010999_懂法狂魔李四_web实验一”。
- 两人一组，单人进行也可，但无优惠政策，单人请按 “学号_姓名_web实验一” 格式提交。

- 标题须写明小组全部成员学号及姓名，也请在文中注明学号及姓名。
- 因未署名造成统计遗漏责任自行承担。
- 实验报告请务必独立完成，如果发现抄袭按零分处理。
- 迟交作业将不再被接收。

请于 **2021 年 10 月 31 日 23:59 之前**打包成 zip 后提交至课程邮箱 ustcweb2021@163.com ，过期不候。

如有未尽事宜，将对本说明进行进一步更新。