1.9——何雨菁：全文检索实现（表实现）

想要用postgresql的全文检索，但是只支持英文，中文分词需要安装插件，安装pg_jieba时出现了一些问题：

CMAKE出错问题（ubuntu）

首先参考：https://itdashu.com/docs/sqlbase/aff6b/sqlpractice6.html
https://github.com/jaiminpan/pg_jieba#install

- **CMAKE编译问题：sudo apt-get update**

```
No CMAKE_CXX_COMPILER could be found.
```

解决：

```
sudo apt-get update
sudo apt-get install gcc

sudo apt-get install g++（假如装了gcc还是不行就装g++）
```

- **找不到postgresql：**

```
Could NOT find PostgreSQL (missing: PostgreSQL_LIBRARY
  PostgreSQL_INCLUDE_DIR)
```

解决：

```
sudo apt-get install libpq-dev postgresql-server-dev-all
```

安装这个就好了。
参考：https://zhuanlan.zhihu.com/p/98450699

- **找不到postgres.h：**

```
/mnt/d/postgresql/pg_jieba/pg_jieba.c:14:10: fatal error: postgres.h: No such
file or directory
   14 | #include "postgres.h"
```

解决：
首先打开

CMakeLists文件：因为是找不到这个postgres.h，所以可能是目录连接出了问题，

```
include_directories(${PostgreSQL_INCLUDE_DIRS})
link_directories(${PostgreSQL_LIBRARY_DIRS})
```

找到所有include_directories和link_directories的地方，输出和pg_config里的目录不同

```
message(STATUS "PostgreSQL_INCLUDE_DIRS is " ${PostgreSQL_INCLUDE_DIRS})
message(STATUS "PostgreSQL_LIBRARY_DIRS is " ${PostgreSQL_LIBRARY_DIRS})
```

如下输入pg_config发现postgres.h应该在/usr/include/postgresql/12/server。把这个目录连接进去就行啦。

```
~:pg_config
BINDIR = /usr/lib/postgresql/12/bin
DOCDIR = /usr/share/doc/postgresql-doc-12
HTMLDIR = /usr/share/doc/postgresql-doc-12
INCLUDEDIR = /usr/include/postgresql
PKGINCLUDEDIR = /usr/include/postgresql
INCLUDEDIR-SERVER = /usr/include/postgresql/12/server 这里
LIBDIR = /usr/lib/x86_64-linux-gnu
PKGLIBDIR = /usr/lib/postgresql/12/lib
LOCALEDIR = /usr/share/locale
MANDIR = /usr/share/postgresql/12/man
SHAREDIR = /usr/share/postgresql/12
SYSCONFDIR = /etc/postgresql-common
PGXS = /usr/lib/postgresql/12/lib/pgxs/src/makefiles/pgxs.mk
```

在CMakeLists文件中加入：

```
include_directories(/usr/include/postgresql/12/server)
link_directories(/usr/include/postgresql/12/server)
```

再cmake，再make就好啦！`[100%] Built target pg_jieba`
昨天安装zhparser也有这个问题，如下这样可以解决！！！（安装jieba也可以试试这个，直接添加到环境变量中）
`export PATH=$PATH:/usr/include/postgresql/12/server`

索引实现

1、安装了postgresql中的分词扩展工具：pg_jieba和zhparser，实验下来觉得pg_jieba分词感觉比zhparser更好一些。

2、数据库配置用orm框架有些麻烦，直接用SQL语句。

3、索引有权重：title权重最大，然后依次是作者等

```
# 在架书籍，这个表用于全文检索
class Book_Onsale(base):
    __tablename__ = 'book_onsale'
    store_id = Column('store_id', Text, primary_key=True)
    book_id = Column('book_id', Text, primary_key=True)
    title = Column('title', Text, nullable=False)
    author = Column('author', Text)
    publisher = Column('publisher', Text)
    # original_title = Column('original_title', Text)
    translator = Column('translator', Text)
    pub_year = Column('pub_year', Text)
    pages = Column('pages', Integer)
    price = Column('price', Integer)
    # currency_unit = Column('currency_unit', Text)
    binding = Column('binding', Text)
    isbn = Column('isbn', Text)
    author_intro = Column('author_intro', Text)
    book_intro = Column('book_intro', Text)
    content = Column('content', Text)
    tags = Column('tags', Text)
    picture = Column('picture', LargeBinary)
```

```
base.metadata.create_all(engine)
# 添加一个新的字段用于建立倒排索引
Session.execute('ALTER TABLE book_onsale ADD COLUMN posting tsvector;')
# 将需要查询的column分词后插入新列中,A-F为重要顺序,A最重要
Session.execute("UPDATE book_onsale SET posting =
setweight(to_tsvector('public.jiebacfg', coalesce(title,''),'A') || "
                "setweight(to_tsvector('public.jiebacfg',
coalesce(title,''),'A')|| "
                "setweight(to_tsvector('public.jiebacfg',
coalesce(author,''),'B')|| "
                "setweight(to_tsvector('public.jiebacfg',
coalesce(translator,''),'D')|| "
                "setweight(to_tsvector('public.jiebacfg',
coalesce(book_intro,''),'E')|| "
                "setweight(to_tsvector('public.jiebacfg',
coalesce(content,''),'F')|| "
                "setweight(to_tsvector('public.jiebacfg',
coalesce(tags,''),'C'));")
# 建立倒排索引（GIN）
Session.execute('CREATE INDEX gin_index ON book_onsale USING GIN(posting);')
# 创建一个分词触发器
Session.execute("CREATE TRIGGER trigger_posting "
                "BEFORE INSERT OR UPDATE ON book_onsale "
                "FOR EACH ROW EXECUTE PROCEDURE "
                "tsvector_update_trigger(posting, 'public.jiebacfg',
title,author,translator,book_intro,content,tags);")
Session.commit()
Session.close()
```

注意查询格式：参考https://itdashu.com/docs/sqlbase/aff6b/sqlpractice6.html

```
postgres=# select title from book_onsale WHERE posting @@ to_tsquery('public.jiebacfg','我的世界');
      title
-----------------
 我的世界
 三毛流浪记全集
 一个狗娘养的自白
 学习的革命
 呐喊
 靠自己去成功
 马桥词典
 故宫史话
 退步集
 长袜子皮皮
 阿修羅
 红楼梦
 窗边的小豆豆
 水浒传（全二册）
 毛毛
 马桥词典
 撒哈拉的故事
(17 rows)
```

可以看到posting这一列的内容：就是一个分词过后的倒排索引

```
' ':92,94,96,99,101,103,105 '':6 '1994':70 '30':46 'isbn':89 '':9 '—':29,30 '一个':26 '一生':33 '中':19 '人物':104 '传':83 '传记':18,91 '作者':21 '使用':87 '再现':24 '博弈论':63,100 '同一':88 '困扰':51 '大师':106 '天才':28 '奇迹般地':55 '奠基性':66 '娜':10 '工作':67 '年':47,71 '年轻':60 '幽灵':80 '康复':56 '心灵':2,98 '感人至深':15 '打断':41 '数学':27,95 '方面':64 '时':61 '普林斯顿':78 '本书':74 '毁灭性':48 '王尔山':12 '生动':17 '生涯':37 '竟':54 '第一版':75 '精神分裂症':39 '精神疾病':50 '纳什':31,82,93 '经受':45 '经济学':102 '维娅':8 '美':4 '美丽':1,97 '获得':69 '萨':11 '西尔':7 '译作':76 '诺贝尔经济学奖':72 '这本':14 '逼真':22
```

picture bytea	posting tsvector
[binary data]	
[binary data]	'
[binary data]	'