




为你写诗db

数据库原理与应用期末项目报告

汇报人：税家晖 刘彦升 赵宏泽

时间：2021年12月8日





PART
0

成果展示



目录 Content

01

项目背景及成果展示

02

数据预处理

03

核心算法分析-SQL

04

服务器搭建



The slide features a minimalist design with dark blue geometric shapes and squares in the corners. A dark blue square in the center contains the text 'PART 01'. Below it, the title '项目背景' is written in a large, dark blue font. Two thin, parallel diagonal lines are positioned to the right of the central text.

PART
01

项目背景



引言

项目的起源

当下NLP算法快速发展，AI写诗也是火热“出圈”的人工智能应用之一，当下的智能写诗主要是利用机器学习、深度学习算法，我们希望用数据库及SQL语言进行简化版的智能写诗

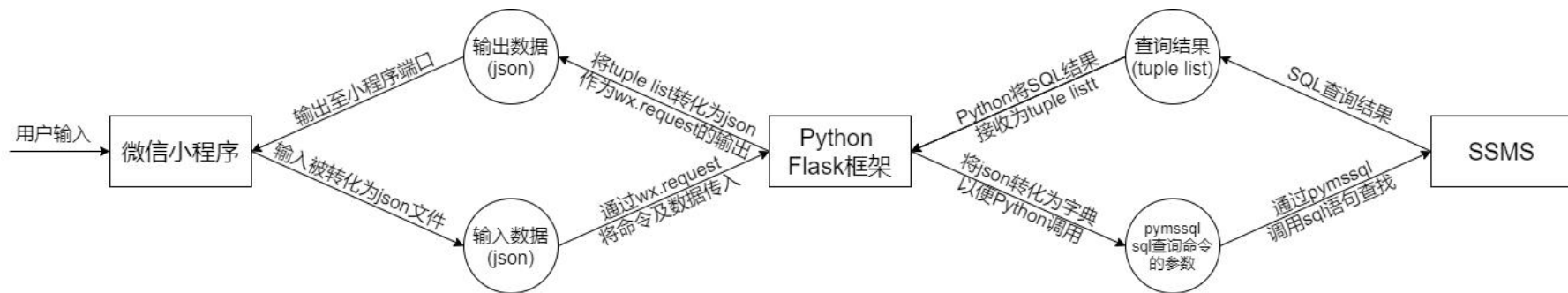
同样，我们受到杨城老师的微信小程序《金点广告词》的启发，我们也希望做出一个大家能够最终访问的，亲身探索的写诗微信小程序。

为尔
写诗





引言



开发环境

为实现数据从微信到SSMS的传输，我们需要使用如下开发环境：

- SQL Server 2019
- Python 3.8+flask+gunicorn+nginx
- Ubuntu 20.04
- 微信开发者工具





数据来源

本次数据来自

<https://github.com/Werneror/Poetry>

该数据集收录了从先秦到现代的共计85万余首古诗词。

诗词数据按朝代分别存入*.csv文件

本次我们选取了其中汉(363), 魏晋(3020), 南北(4586), 隋(1170), 唐(49195), 宋(287114), 元(37375)共计382823条数据

	题目	朝代	作者	内容
0	忆秦娥	唐	冯延巳	风淅淅。夜雨连云黑。滴滴。窗外芭蕉灯下客。除非魂梦到乡国。免被关山隔。忆忆。一句枕前争忘得。
1	送兄	唐	七岁女子	别路云初起，离亭叶正飞。所嗟人异雁，不作一行归。
2	再赠	唐	上元夫人	弄玉有夫皆得道，刘纲兼室尽登仙。君能仔细窥朝露，须逐云车拜洞天。
3	留别	唐	上元夫人	萧郎不顾凤楼人，云涩回车泪脸新。愁想蓬瀛归去路，难窥旧苑碧桃春。
4	赠封陟	唐	上元夫人	谪居蓬岛别瑶池，春媚烟花有所思。为爱君心能洁白，愿操箕帚奉屏帏。
... ..				
49190	菩萨蛮	唐	魏承班	罗裾薄薄秋波染，眉间画得山两点。相见绮筵时，深情暗共知。翠翘云鬓动，敛态弹金凤。宴罢入兰房，...
49191	满宫花	唐	魏承班	雪霏霏，风凛凛。玉郎何处狂饮，醉时想得纵风流，罗帐香帏鸳寝。春朝秋夜思君甚，愁见绣屏孤枕。少...
49192	谒金门	唐	魏承班	烟水阔，人值清明时节。雨细花零莺语切，愁肠千万结。雁去音徽断绝，有恨欲凭谁说。无事伤心犹不彻...
49193	玉楼春	唐	魏承班	寂寂画堂梁上燕，高卷翠帘横数扇。一庭春色恼人来，满地落花红几片。愁倚锦屏低雪面，泪滴绣罗金缕...
49194	李万州挽诗	唐	唐年	仙李来南日，熙宁去国人。一身辞赤芾，四世尚朱轮。桑梓推前辈，葭莩接世姻。伤心挽耆旧，不独为情...





PART
02

数据预处理



数据预处理

我们首先对数据源进行整合

合并所有诗歌并去除异常值

五言诗与七言诗同理

```
select * into shige_full
from (select * from song
      where contents not like '%□%'
      and tttitle not like '%□%'
      and contents not like '%?%'
      and tttitle not like '%?%'
      union
      .....
)
```

接着我们拆出五言与七言诗歌

```
create view wuyan as
select * from shige_full
where contents like
'[ㄣ-座][ㄣ-座][ㄣ-座][ㄣ-座][ㄣ-座][,。]%'

create view qiyan as
select * from shige_full
where contents like
'[ㄣ-座][ㄣ-座][ㄣ-座][ㄣ-座][ㄣ-座][ㄣ-座][ㄣ-座][,。]%'
```





数据预处理

	ttitle	dynasty	author	contents
1	边上看猎赠元戎	唐	韩偓	绣帘临晓觉新霜，便遣移厨较猎场。燕卒铁衣围汉相，鲁儒戎服从梁
2	边上闻笛三首 其三	唐	杜牧	胡雏吹笛上高台，寒雁惊飞去不回。尽日春风吹不散，只应分付客愁
3	边上作三首 其三	唐	贯休	见说青冢穴，中有白野狐。时时出沙碛，向东而号呼。号呼复号呼，
4	边韶	宋	徐钧	寸晷分阴闲可惜，粪墙朽木责非苛。便便书腹贪眠昼，免得诸生笑诮
5	边州客舍	唐	项斯	开门不成出，麦色遍前坡。自小诗名在，如今白发多。经年无越信，
6	编猿作	宋	姜特立	老来事业无多子，收拾新诗又满编。金玉那能润身后，等閒犹作百年
7	蝙蝠	宋	范成大	伏翼昏飞急，营营定苦饥。聚蚊充口腹，生汝亦奚为。
8	贬朱崖行临高道中买愁村古未有对马上口占	宋	胡铨	北往长思闻喜县，南来怕入买愁村。区区万里天涯路，野草荒烟正断
9	扁舟放流而下	宋	韩淲	小舟回柁下弯埼，烟草云山风雨时。莫夜归来全似梦，石桥灯火见疏
10	弁净人奉辟支佛牙求度	宋	释居简	错认缘生不自知，退牙今属弁沙弥。若还祖弁有灵骨，选得僧成莫学
11	汴堤冬日二首 其一	宋	周紫芝	榆柳风微不动尘，平沙如粉衬车轮。今年汴上三冬暖，已带长安二月





数据预处理

分词

由于sql语言没有很多优秀的扩展，
故无法做到较为精准的分词，
所以我们使用python的thulac库进
行古诗分词，该库参考：

<http://thulac.thunlp.org/demo>

示例：

输入：随机过程随机过

输出：随机_v 过程_n 随机_v 过_u

```
import thulac
import pandas as pd

df = pd.read_csv('wuyan.csv',
encoding = 'gb18030')
thul = thulac.thulac()

for i in range(len(df)):
    text = df['contents'][i]
    text_res = thul.cut(text,
text = True)
    df.loc[i, 'fenci'] = text_res
print(df['fenci'][0])
df.to_csv('wuyan_fenci.csv',
encoding = 'gb18030')
```



数据预处理

	fenci
1 佳，可但莲菊好。富贵本何心，莫以色见我。...	高枕_n 松间石_n，_w 如_v 依_g 未_d 易_a 知_v 。_w 世情_n 祇益_d 睡...
2 问，寒光为底留。拟随风叶去，还似雪花浮。...	此夜_r 一_m 轮_q 满_a，_w 骚情_n 不_d 奈_g 秋_g 。_w 月_n 非_g 人...
3 钓，千茎雪鬓蓬。短衣难掩肝，独棹似浮空。...	江远_np 孤舟_n 小_a，_w 悠悠_a 避世翁_n 。_w 望_v 中_f 留_v 夕照_n...
4 所，剥啄访平生。不是师严钓，应须学尹耕。...	避喧_v 谁_r 氏子_n，_w 远近_n 不_d 知名_a 。_w 踏_v 雨_n 来_v 相_d...
5 屐，馀生付酒缸。新诗眩老眼，细读傍寒窗。...	一_d 别_v 从_p 京洛_ns，_w 相逢_v 向_p 海邦_n 。_w 倦游怜舌在_id ...
6 口，晴晖绊柳腰。迂儒欲何用，只合伴渔樵。...	去_v 去_v 春_g 无_v 语_g，_w 融融_a 物自骄_a 。_w 饷耕怜野老_id，...
7 数，黄花插自羞。青山本无事，敛黛为谁忧。...	秋色_n 澹脩渚_n，_w 夕_g 光明_a 小_a 楼_n 。_w 溪_g 虚_a 烟_n 漠漠...
8 动，雄词骇浪翻。九重应震悼，同性失毛原。...	珠海_ns 光无尽_id，_w 瑶山_n 气_n 自温_v 。_w 鸿基开烈祖_id，_w ...
9 道，分符知几州。终遗搢绅恨，不作富民侯。...	卓荦千人杰_id，_w 飞腾_v 四十_m 秋_g 。_w 才_d 应有_v 馀刃_n，_w ...
10 酒，谁同别后襟。空馀钓矶在，落日见孤岑。...	伯仲_np 皆_d 登第_v，_w 人_n 言薛_v 与_p 林_g 。_w 凄凉_a 五_m 年...
11 淪，雨声閤??。安用天官为，不悟秋序夺。巫...	乘时火_id 初_d 流_v，_w 执热病_n 未_d 脱_v 。_w 踏_v 冰思少苏_id ...





数据预处理

词频统计

使用sql游标，统计分词后每个词出现的频率。

将五言诗和七言诗分类统计

```
declare @str nvarchar(200)

if object_id('tempdb.dbo.#array') is
not null --判断临时表是否已经存在
    drop table #array
create table #array(ch nvarchar(20))

--全部改成qiyan即可创建七言的词频库
declare cur scroll cursor for
    select fenci from wuyan_fenci
```

```
open cur
fetch first from cur into @str
while @@FETCH_STATUS = 0
begin
    insert into #array(ch) select value from
    string_split(@str, ' ')
    where patindex('%[吖-座]%', value)>0
    fetch next from cur into @str
end
close cur
deallocate cur

select ch, count(*) cnt into wuyan_cipin
from #array group by ch order by cnt desc
drop table #array

select * from wuyan_cipin order by cnt desc
```



数据预处理

	ch	cnt
1	不_d	57688
2	有_v	27084
3	来_v	24205
4	无_v	21319
5	一_m	21211
6	我_r	17391
7	人_n	14267
8	未_d	13106
9	去_v	12955
10	中_f	12696
11	已_d	12677
12	时_g	12291





数据预处理

韵脚

我们以每个词的最后一个韵母作为其韵脚
例如头(tou)韵脚为u，与图(tu)是押韵的

```
create view wuyan_ci_full as
select a.*, replace(substring(a.ch, patindex('%[a-z]%', a.ch)-1, len(a.ch)), '_ ', '') cixing,
       replace(substring(a.ch, 1, charindex('_', a.ch)), '_ ', '') ci,
       substring(reverse(rtrim(b.py)), patindex('%[aeiou]%', reverse(rtrim(b.py))), 1) py
from wuyan_cipin a, xhzd b
where b.zi=substring(rtrim(replace(substring(a.ch, 1, charindex('_', a.ch)), '_ ', '')),

                    len(rtrim(replace(substring(a.ch, 1, charindex('_', a.ch)), '_ ', ''))), 1)

select * from wuyan_ci_full
```





数据预处理

	ch	cnt	cixing	ci	py
1	第二十_m	1	m	第二十	i
2	始知登山劳_id	1	id	始知登山劳	o
3	登兹亭_n	1	n	登兹亭	i
4	飞腾障_nz	1	nz	飞腾障	a
5	王国_n	14	n	王国	o
6	缓_a	168	a	缓	a
7	死_v	1556	v	死	i
8	风树_n	37	n	风树	u
9	蝉_n	358	n	蝉	a
10	冀北_ns	12	ns	冀北	i
11	攀折_v	57	v	攀折	e



模板生成

我们可以根据thulac的分词结果来得出古诗词所使用的句子结构，
我们可以统计最经常使用的句子结构作为我们的模板，例如：

高枕_n 松间石_n ， _w 如_v 侬_g 未_d 易_a 知_v 。

的模板为：n n 23 v g d a v 11111

详细的模板函数信息请查看我们的项目报告

右侧是排名前十的模板（不包括id）

	ch	cnt
1	n v n 212	20683
2	n n 23	14030
3	v n 23	12920
4	n np 23	8689
5	n d v 212	8447
6	n n 32	7675
7	v v n 212	6712
8	n v 32	5154
9	np v 32	4984
10	d v n 113	4962





PART
03

功能设计



功能设计

功能1：随机生成诗歌

生成诗歌的条件有：

- 句尾字押韵
- 韵脚统一
- 选用模板且不能一致

故我们以五言的代码为例，
我们的测验以五言绝句为例，
生成结果为：

林间先更争
至道老收声
择奇性刚野
何刻归斯文

```
create proc wuyan_suiji (@m int)
as
begin
    declare @temp nvarchar(30), @num nvarchar(6), @ty nvarchar(10)
    declare @str nvarchar(2), @s nvarchar(10) = ' ', @res nvarchar(200) = ' '
    declare @i int = 0, @n int, @step int = 1, @res_out nvarchar(400) = '',
            @yunjiao nvarchar(100)
    while @step <= @m
    begin
        select @res = ' ', @i = 0
        select top 1 @temp = ch from wuyan_muban
            where cnt > 20
            order by NEWID()
        set @num = substring(@temp, patindex('%[0-9]%', @temp), len(@temp))
        set @num = replace(replace(@num, '_', ''), '-', '')
        set @ty = substring(@temp, 1, patindex('%[0-9]%', @temp) - 1)
        declare cur scroll cursor for
            select value from string_split(@ty, ' ')
```



功能设计

```
open cur
fetch first from cur into @str
while @@FETCH_STATUS=0
begin
    set @i = @i+1
    set @n = cast(substring(@num,@i,1) as int)
    if @i=len(rtrim(ltrim(@num))) and @step>1
    begin
        select top 1 @s=ci from wuyan_ci_full
        where cixing=@str and len(rtrim(ltrim(ci)))=@n
        and cnt>100 and py=@yunjiao
        order by NEWID()
        if len(ltrim(rtrim(@s)))<>@n
        begin
            select top 1 @s=ci from wuyan_ci_full
            where cixing=@str and len(rtrim(ltrim(ci)))=@n
            and cnt>1 and py=@yunjiao
            order by NEWID()
        end
    end
end
else
...
```

```
deallocate global cur
if len(ltrim(rtrim(@res)))=5
begin
    if @step=1
    begin
        select
@yunjiao=substring(reverse(rtrim(py)),patindex('%[aeiou]%',reverse(rtrim(py))),1) from xhzd where
zi=substring(rtrim(@res),len(rtrim(@res)),1)
    end
    set @res_out=concat(@res_out,@res)
    set @step = @step+1
end
end
select @res_out
end

--调用存储过程
exec wuyan_suiji @m=4
```





功能设计

功能2：藏头诗生成

藏头诗生成与功能1类似，仅需再多加一个条件，即为每一句的第一个字和所给词的相应位置相同即可。此时为了计算速度，我们抛弃掉了随机模板，采用n v n 212的模板生成（若找不到，则变为n v n 122），此处代码请查阅我们的报告

功能3：飞花令

我们可以利用SQL选出含有特定字词的诗句，代码如下

```
create function process_feihualing(@str nvarchar(200), @a
nvarchar(10))
returns nvarchar(100)
as
begin
    declare @res nvarchar(100), @i int, @j int, @k int
    set @i = patindex('%'+@a+'%', @str)
    if patindex('%。%', reverse(substring(@str, 1, @i)))=0
    begin
        set @j = 1
    end
    ...
end

select top 10 *, dbo.process_feihualing(contents, '春') from
shige_full
where contents like '%春%'

create proc feihualing(
    @ling nvarchar(10)
)
as
begin
    select top 1 *, dbo.process_feihualing(contents, '%' + @ling + '%')
    juzi
    from shige_full
    where patindex('%'+@ling+'%', contents) > 0
    order by NEWID()
end
```



PART
04

服务器搭建与前端开发



Linux环境配置

环境配置

在Ubuntu上先安装python, virtualenv, 然后安装nginx。创

建python虚拟环境后再pip install gunicorn。

配置完毕后将写好的flask框架放入python虚拟环境中，运行

Gunicorn -w 2 -b 127.0.0.1:5000 manage:app

```
To escape to local shell, press 'Ctrl+Alt+]'.

Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.4.0-47-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

20 updates can be applied immediately.
11 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

*** System restart required ***

Welcome to Alibaba Cloud Elastic Compute Service !

Last login: Fri Dec  3 21:19:25 2021 from 218.89.242.7
root@iZ7j0a80xsaqlDZ:~# cd /root/venv
root@iZ7j0a80xsaqlDZ:~/venv# source bin/activate
(venv) root@iZ7j0a80xsaqlDZ:~/venv# gunicorn -w 2 -b 127.0.0.1:5000 manage:app
```

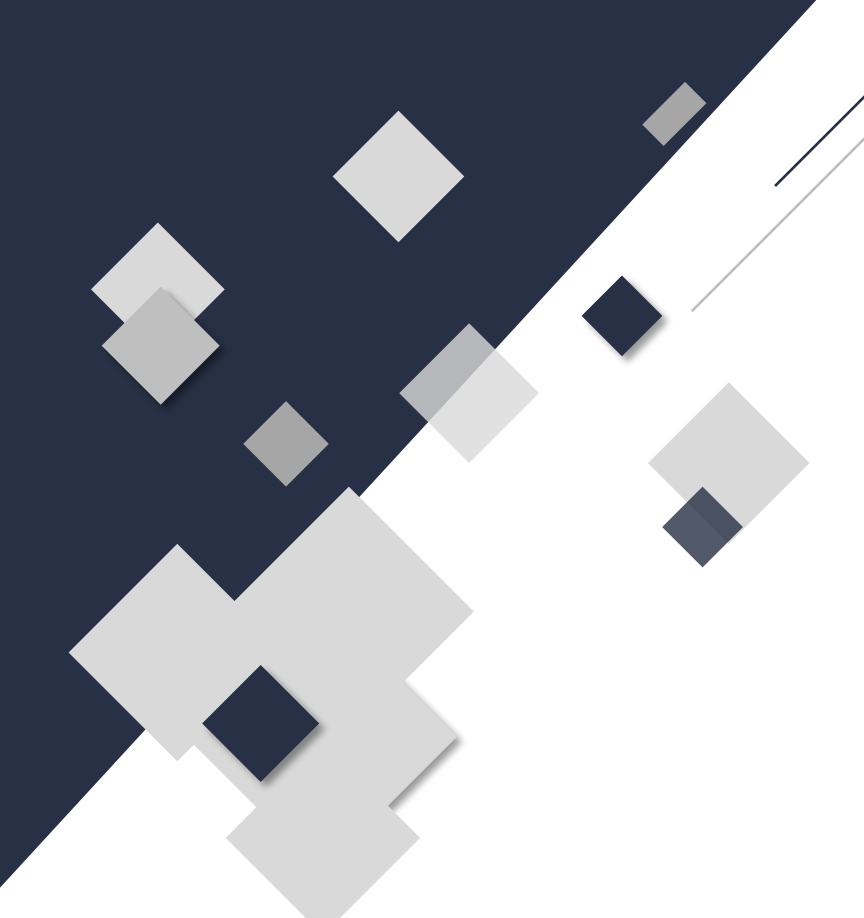




小程序与后端的数据传输

使用微信内置的request函数，post方法，将用户的输入使用json文件传输到服务器端的python文件，再处理由服务器返回json文件，展示到视图层。





为你写诗db 谢谢观看与聆听

Thank you for watching and listening

一鸣菖蒲定间道，九迁丹柿贩女戎。
小谢圭欲寐阶药，金猊午眠五更钟。
人化树层墓古踪，别离凉温饰螯宫。
卷定绝由淡风巢，了事高方畔孤咏。

