数据库项目——粤语字典

组员：

谢润烁 18340183

杨智博 18340195

李沛航 18340092

项目背景及简介

方言保护是一项很有难度的工作。一方面，我们国家为了全国人民更好地互相沟通，较大力度地推行普通话；另一方面，当前对方言的记载和研究是比较稀缺的。作为使用人群广泛且标准化较好的方言，粤语的资料都达不到非常丰富，易于研究的地步，更不用提其它小众的方言。

我们注意到了，市面上的方言字典，大部分只能通过字来查找读音，或者反过来。这是非常线性的设计思维，其索引跟线性的字典书没有本质区别。我们在学习的过程中，如果能从一个点扩展开来——比方说同音不同字，同偏旁不同音，粤语发音相同但是普通话发音不同，或者反过来——我们会从中发现更多丰富的例子来让我们举一反三地学习，而且这也有利于我们理解方言演变的规律。

项目研究内容

在本项目中，我们旨在设计出一款专攻检索的粤语字典APP。虽然说是粤语字典，但是我们希望该产品能任意扩充到所有方言（包括普通话）。该APP能实现：

1. 最基本的检索方式：输入字查找对应的粤语拼音，以及输入粤语拼音查找对应的字
2. 中级检索方式：对于一个拼音，我们能选定其声母，韵腹，韵尾和声调来进行索引。其中每一项都能选择'所有'，也就是允许丰富的组合方式来检索。比方说，我们想要检索出韵腹为'aa'，其它部分没有限制的所有字，那我们只需要在韵腹这一栏选出'aa'，其它部分默认'所有'，这样我们就能索引出各种字，比方说啊(aa1)，押(aat3)，摆(baai2)等等。
3. 高级检索方式：在中级检索方式的基础上，我们将每一部分的选择从单选变为多选，这就意味着更加丰富的组合。

当然，除此之外我们还有很多想法，例如通过和汉语拼音结合起来共同索引，或者通过部首来进行索引，但限于我们所能获得的数据，我们就暂时做不了这些功能。不过这些可以等今后有了数据再做，因为我们在数据库的设计上已经保证了字典的可扩展性。

系统环境及技术路线

为了使得应用可以在较多场景下应用，我们选择了Android环境进行开发。一方面移动端比PC端更加普遍，便携；另一方面安卓用户比起iOS用户更为广泛。我们采用Java和Kotlin混合开发的方式来开发该产品。

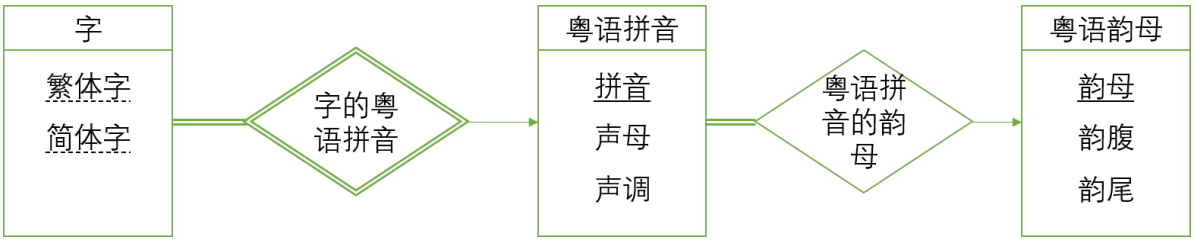
数据处理

我们找到的原始数据来自于开放粤语词典，其数据的原格式是一个Python的List打印出来的文件。其中包含了(繁体字, 简体字, 粤语拼音)的对，粤语拼音可以有1~3个，用斜杠分开。因此我们需要先将该文件处理成比较好读入的格式，然后再做进一步的处理。

除此之外，由于我们需要对拼音进行细分，所以我们还需要将每个字的拼音处理成声母+韵腹+韵尾+声调的形式。具体细节此处不赘述，主要是通过粤语拼音的规范设计。总之，通过适当的数据处理，我们得到了我们所需要的所有原始数据。

数据库设计

为了满足软件的需求，我们首先为数据库设计了下面ER图所表达的ER关系：



图中有3个实体集，分别是字、粤语拼音和粤语韵母，2个联系集，分别是字的粤语拼音和粤语拼音的韵母。转化为关系模式时，联系集字的粤语拼音的模式就省略了，而联系集粤语拼音的韵母的模式并入实体集粤语拼音的模式中，最后得到三个模式：

字（繁体字，简体字，拼音）

粤语拼音（拼音，声母，韵母，声调）

粤语韵母（韵母，韵腹，韵尾）

其中模式字的属性拼音是参考模式粤语拼音的外键，模式粤语拼音的属性韵母是参考模式粤语韵母的外键。下面是三个模式对应的三张表的示例：

- 粤语字典表

繁体字	简体字	拼音
諙	诿	wai2

- 粤语拼音表

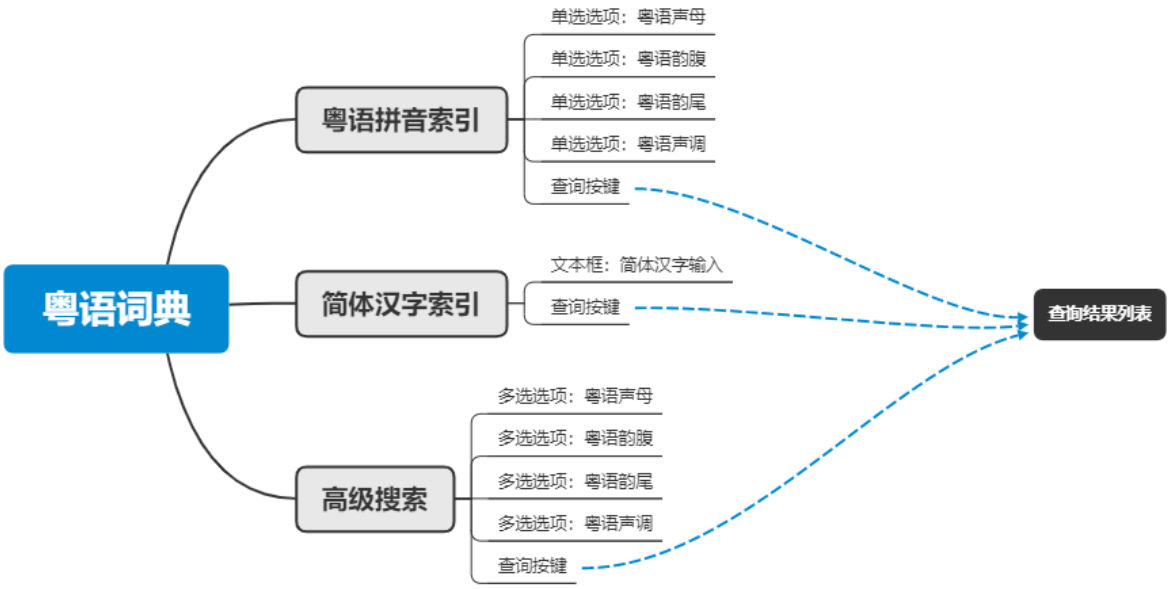
拼音	声母	韵母	声调
seoi3	s	eoi	3

- 粤语韵母表

韵母	韵腹	韵尾
ei	e	i

用户界面设计

根据我们规划的用户需求，设计了如下图所示的UI界面层次：


粤语词典
粤语拼音索引
单选选项：粤语声母
单选选项：粤语韵腹
单选选项：粤语韵尾
单选选项：粤语声调
查询按键
简体汉字索引
文本框：简体汉字输入
查询按键
高级搜索
多选选项：粤语声母
多选选项：粤语韵腹
多选选项：粤语韵尾
多选选项：粤语声调
查询按键
查询结果列表

结果展示

- 初始界面

- 粤拼检索


1:41
LTE
粤拼检索
无
aa
无
音..
丫(丫)aa1
呀(呀)aa1
啊(啊)aa1
鸦(鴉)aa1
哑(啞)aa1
吖(吖)aa1
锕(錒)aa1
哑(啞)aa2
啊(啊)aa2
亚(亞)aa3
阿(阿)aa3
氩(氬)aa3
搜索

- 简字检索


1:45
LTE
简字检索
你好
你(你)nei5
你(妳)nei5
好(好)hou2
好(好)hou3
搜索

- 高级检索


1:46
LTE
高级检索
选择声母
选择韵腹
选择韵尾
选择音调
叭(叭)baa1
疤(疤)baa1
粑(粑)baa1
芭(芭)baa1
笆(笆)baa1
爸(爸)baa1
葩(葩)baa1
巴(巴)baa1
吧(吧)baa1
钯(鈀)baa2
靶(靶)baa2
搜索

总结

虽然完成这次数据库项目的时间比较紧迫，但是由于我们一开始设定的目标不会过大也不会过简单，所以我们最后还是能准时完成该项目。在这次项目合作中，我们学到了很多。不仅是对Android开发方面有了一定的了解，还通过彼此的讨论增强了团队合作的意识和思维。在本次项目中，由于前端和后端需要进行对接，以及数据库的设计与后端逻辑也需要对接，我们必须保证良好的沟通才能实现这一点，而我们也确实做到了。

除此之外，该项目虽然在功能上和技术上比较简单，但是它是比较有用的，因为它填补了市面上高级索引的数据库这一块空白。尽管专门的科研平台可能有自己的数据库，不过由于那些数据库及其相应的应用程序没有开放，普通的音韵研究爱好者以及学习者难以从中获得较好的资源，从而浪费了很多时间。因此，我们坚信我们的应用是有人愿意用的，而且是真的对大家有帮助的。

最后是对该项目的展望。目前为止，该项目虽然不算很成熟，但是也不算差。如果要对词典进行词条，部首和方言扩充，也无需增添什么新的技术。我们希望可以将该应用作为一个平台，凭借公众的力量来完成对该词典数据的完善，并且以此推动我国的汉语方言保护和汉语研究工作。