Helena Kazenski - CS333 Final Project

Project Overview:

This hanzi database organizes foundational components of the Chinese language, such as: words, characters (hanzi), radicals, example sentences, semantic fields, and word-to-word relationships. The goal of this project is to model how the Chinese language is structured from the ground up—from characters and the radicals they are built from, to words and the characters they contain, and then connecting those words with meaningful example sentences and semantic fields.

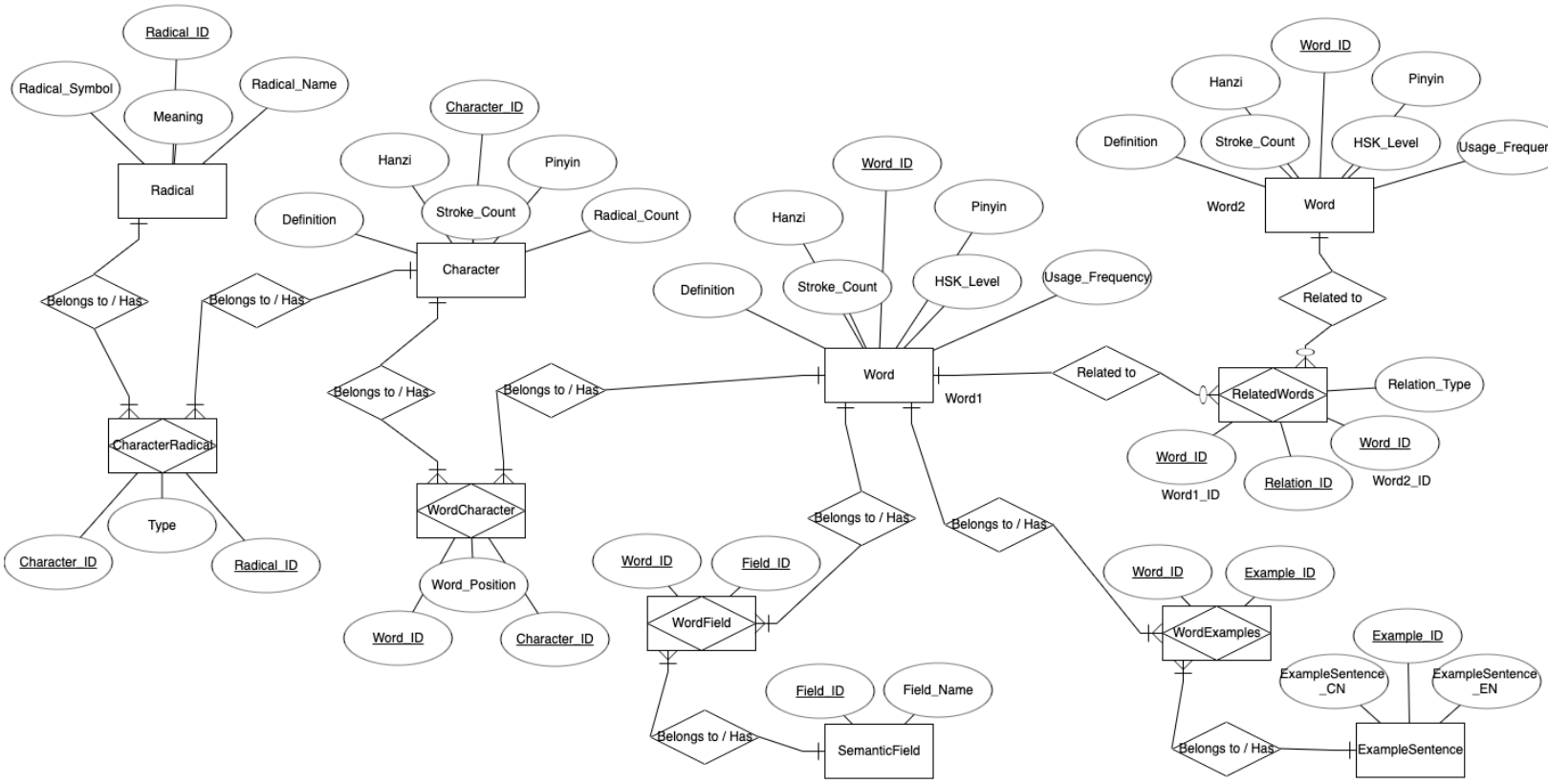
This database has the potential to be useful for language learners, educators, and anyone studying the morphology and semantics of Chinese. It builds the foundations of what could eventually be expanded into a larger-scope project — one that could grow beyond dictionary functions. It could even be a foundation for an educational app that molds itself to best fit the users' needs, or a deeper linguistics analysis tool, or an NLP project for machine translation and/or speech recognition.

Complexity:

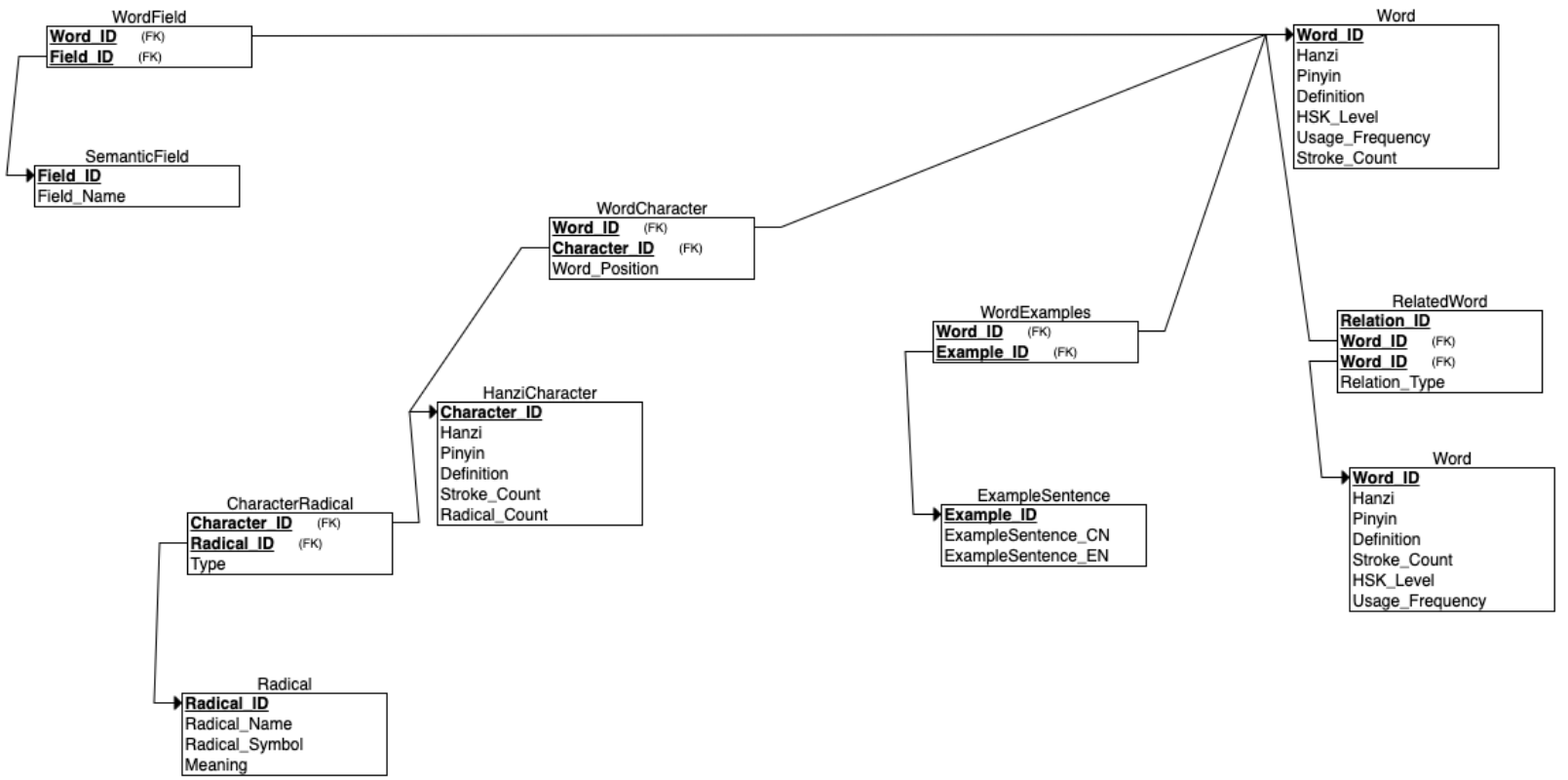
This project has the appropriate technical depth for a 300-level course because it includes multiple connected tables that model real relationships that occur in the Chinese language. Setting up tables and the connecting tables that tackle many-to-many relationships required careful planning to ensure that the data stayed organized and avoided duplication.

The queries show sufficient complexity and showcase skill in using SQL. They use multi-table joins, grouping, aggregates like AVG and COUNT, and string functions like GROUP_CONCAT to analyze patterns in the data, such as shared radicals, average stroke counts, and how words are used in sentences. Because the database supports linguistic breakdowns and semantic analysis, it demonstrates the level of design thinking and SQL skills expected for a 300-level course.

ERD:

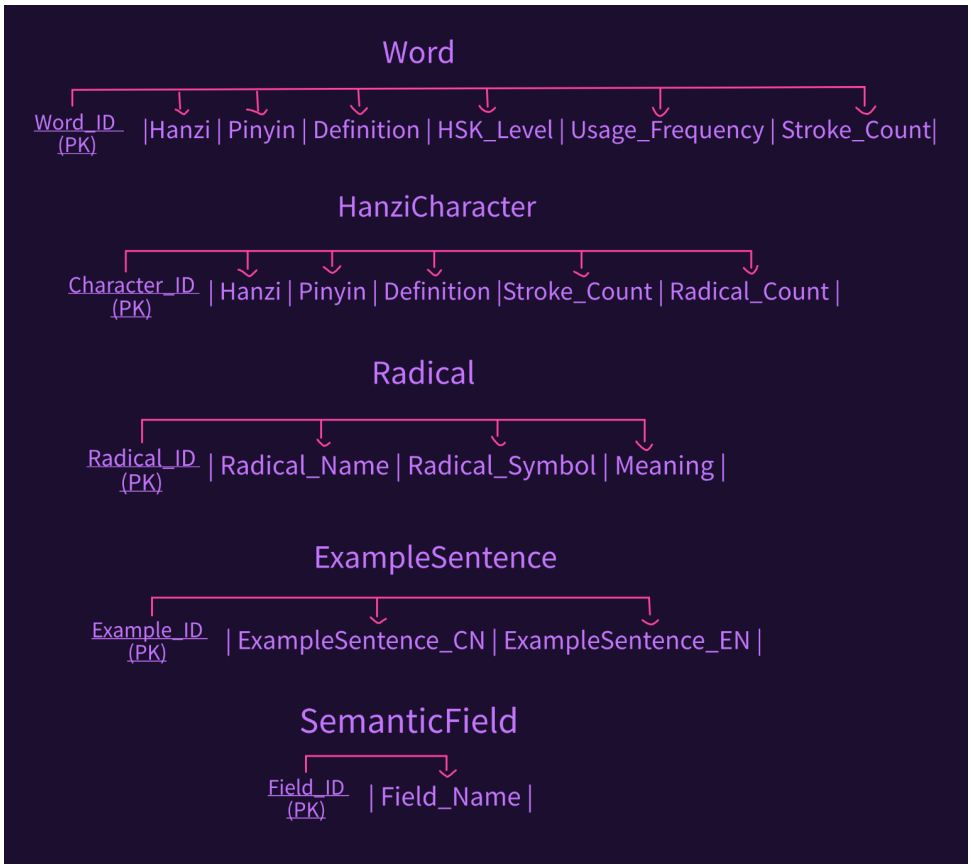

Radical_ID
Radical_Symbol
Radical_Name
Meaning
Radical
Belongs to / Has
Belongs to / Has
Character_ID
Hanzi
Pinyin
Stroke_Count
Definition
Radical_Count
Character
Belongs to / Has
Belongs to / Has
CharacterRadical
Type
Character_ID
Radical_ID
WordCharacter
Word_Position
Word_ID
Character_ID
Word_ID
Hanzi
Pinyin
Definition
Stroke_Count
HSK_Level
Usage_Frequency
Word
Word1
Belongs to / Has
Belongs to / Has
Word_ID
Field_ID
WordField
Field_ID
Field_Name
SemanticField
Belongs to / Has
Related to
Word_ID
Hanzi
Pinyin
Definition
Stroke_Count
HSK_Level
Usage_Frequen
Word2
Word
Related to
RelatedWords
Relation_Type
Word_ID
Word1_ID
Relation_ID
Word_ID
Word2_ID
Word_ID
Example_ID
WordExamples
Example_ID
ExampleSentence_CN
ExampleSentence_EN
Belongs to / Has
ExampleSentence

RS:

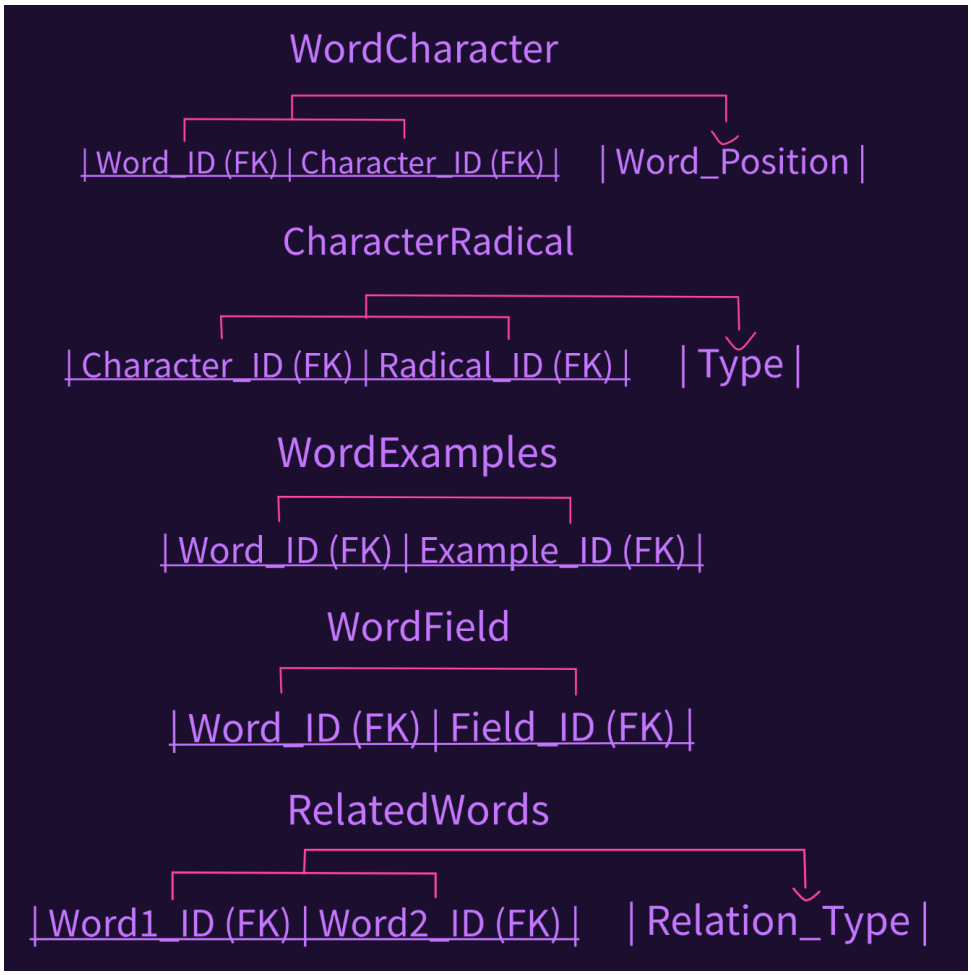

WordField
Word_ID (FK)
Field_ID (FK)
SemanticField
Field_ID
Field_Name
Word
Word_ID
Hanzi
Pinyin
Definition
HSK_Level
Usage_Frequency
Stroke_Count
WordCharacter
Word_ID (FK)
Character_ID (FK)
Word_Position
WordExamples
Word_ID (FK)
Example_ID (FK)
RelatedWord
Relation_ID
Word_ID (FK)
Word_ID (FK)
Relation_Type
HanziCharacter
Character_ID
Hanzi
Pinyin
Definition
Stroke_Count
Radical_Count
CharacterRadical
Character_ID (FK)
Radical_ID (FK)
Type
ExampleSentence
Example_ID
ExampleSentence_CN
ExampleSentence_EN
Word
Word_ID
Hanzi
Pinyin
Definition
Stroke_Count
HSK_Level
Usage_Frequency
Radical
Radical_ID
Radical_Name
Radical_Symbol
Meaning

Dependency Diagrams:

Entities:



Associative Tables:



Dependency:

Yes, my database is 3NF. It doesn't have any repeating groups, so it is 1NF. It doesn't have any partial dependencies either — even in the associative tables, all non-key attributes depend on the full composite key — so it is 2NF. Finally, it doesn't have any transitive dependencies because every attribute in every table depends directly on the table's primary key (or composite key), so it is 3NF.

Queries:

```
27  -- Query 1
28  -- Select all characters with pinyin ending in 'i'
29  SELECT *
30  FROM HanziCharacter
31  WHERE Pinyin LIKE '%i'
32  ;
```

Character_ID	Hanzi	Pinyin	Definition	Stroke_Count	Radical_Count
1	你	nǐ	you	7	2
4	是	shì	to be	9	3
13	谁	shéi	who	10	2
18	几	jǐ	how many	2	1
21	一	yī	one	1	1
24	四	sì	four	5	2
27	七	qī	seven	2	1
30	十	shí	ten	2	1
32	岁	suì	years of age	6	2
35	块	kuài	piece; lump; money	7	3
37	没	méi	not; don't have	7	2
39	太	tài	too; very	4	2
42	在	zài	at; in; exist	6	2
47	喂	wèi	hello (on the phone)	12	3
54	医	yī	medicine	7	2
58	北	běi	north	5	2
65	里	lǐ	inside	7	2
75	期	qī	period; week	12	3
79	时	shí	time	10	3
83	子	zǐ	child	3	1
85	师	shī	teacher	10	3
90	衣	yī	clothes	6	2
92	米	mǐ	rice	6	2
96	杯	bēi	cup	8	2
98	飞	fēi	to fly	3	1
99	机	jī	machine; airplane	6	2
104	视	shì	vision	11	2
107	气	qì	air; gas	4	2
112	西	xī	west; thing	6	3
114	字	zì	character; word	6	2
119	椅	yǐ	chair	12	4
122	再	zài	again	6	2
125	对	duì	correct; toward	5	2
126	起	qǐ	to rise	10	3
134	来	lái	to come	7	1
135	回	huí	to return	6	2
137	吃	chī	to eat	6	2
139	睡	shuì	to sleep	13	3
142	买	mǎi	to buy	6	2
143	开	kāi	to open; drive	4	1
146	爱	ài	love	10	3
147	喜	xǐ	happy	12	3
151	识	shí	knowledge	7	2
152	会	huì	can; able to	6	2
162	丽	lì	beautiful	7	2

45 rows in set (0.00 sec)

```sql
-- Query 2
-- Count the number of characters with more than two radicals
SELECT Count(Character_ID) AS Num_Over_2_Radicals
FROM HanziCharacter
WHERE Radical_Count > 2
;
```

```
+---------------------+
| Num_Over_2_Radicals |
+---------------------+
|                  46 |
+---------------------+
1 row in set (0.00 sec)
```

```sql
-- Query 3
-- Select the first 20 words and all the radicals used in each character (w/ no duplicate radicals in a word)
SELECT Word.Word_ID, Word.Hanzi AS Word_Hanzi,
    GROUP_CONCAT(DISTINCT Radical_Symbol ORDER BY Radical_Symbol) AS Radicals_Used
FROM Word
JOIN WordCharacter ON Word.Word_ID = WordCharacter.Word_ID
JOIN CharacterRadical ON WordCharacter.Character_ID = CharacterRadical.Character_ID
JOIN Radical ON CharacterRadical.Radical_ID = Radical.Radical_ID
GROUP BY Word.Word_ID, Word.Hanzi
ORDER BY Word.Word_ID
LIMIT 20
;
```

```
+---------+------------+-------------------------+
| Word_ID | Word_Hanzi | Radicals_Used           |
+---------+------------+-------------------------+
|       1 | 你         | 亻,刀,小               |
|       2 | 我         | 戈,扌                   |
|       3 | 你们       | 亻,刀,小,门            |
|       4 | 他         | 丨,乚,亻               |
|       5 | 她         | 丨,乚,女               |
|       6 | 它         | 匕,宀                   |
|       7 | 我们       | 亻,戈,扌,门            |
|       8 | 他们       | 丨,乚,亻,门            |
|       9 | 她们       | 丨,乚,亻,女,门        |
|      10 | 什么       | 丿,二,亻,厶           |
|      11 | 谁         | 讠,隹                  |
|      12 | 哪儿       | 丨,二,儿,口,阝        |
|      13 | 怎么       | 一,丨,丿,二,厶,心     |
|      14 | 多少       | 丿,夕,小               |
|      15 | 几         | 几                      |
|      16 | 这         | 文,辶                  |
|      17 | 那         | 丨,二,阝               |
|      18 | 一         | 一                      |
|      19 | 二         | 二                      |
|      20 | 三         | 一,二                   |
+---------+------------+-------------------------+
20 rows in set (0.01 sec)
```

```sql
-- Query 4
-- Count the number of example sentences that 20 words appear in
SELECT Word.Word_ID, Word.Hanzi, COUNT(WordExamples.Example_ID) AS Sentence_Count
FROM Word
JOIN WordExamples ON Word.Word_ID = WordExamples.Word_ID
GROUP BY Word.Word_ID, Word.Hanzi
ORDER BY Sentence_Count DESC
LIMIT 20
;
```

Word_ID	Hanzi	Sentence_Count
2	我	9
30	是	6
4	他	5
1	你	5
28	个	4
5	她	4
7	我们	4
8	他们	3
29	岁	2
20	三	2
16	这	2
11	谁	1
17	那	1
6	它	1
9	她们	1
22	五	1
12	哪儿	1
14	多少	1
10	什么	1
25	八	1

20 rows in set (0.00 sec)

```sql
-- Query 5
-- Select all characters containing the selected 20 radicals
SELECT Radical.Radical_Symbol,
    GROUP_CONCAT(HanziCharacter.Hanzi ORDER BY HanziCharacter.Hanzi) AS Characters
FROM Radical
JOIN CharacterRadical ON Radical.Radical_ID = CharacterRadical.Radical_ID
JOIN HanziCharacter ON CharacterRadical.Character_ID = HanziCharacter.Character_ID
GROUP BY Radical.Radical_Symbol
ORDER BY Radical.Radical_Symbol
LIMIT 20
;
```

Radical_Symbol	Characters
一	一,七,三,不,喂,怎,是,本
丨	个,他,哪,块,她,怎,那
丶	太,的
丷	商
丿	不,么,九,在,少,怎
乚	七,他,喂,她
亅	了
二	三,二,五,些,什,哪,怎,那,院
亠	六,商,校
人	个,块
亻	什,他,们,你,在
儿	儿,四,院
八	八,六,商,校
冂	商
宀	学
几	几
刀	你
力	五
勹	的
匕	些,呢,它

20 rows in set (0.00 sec)

```sql
-- Query 6
-- Select the top 10 most frequently used words
SELECT Word_ID, Hanzi, Definition, Usage_Frequency
FROM Word
ORDER BY Usage_Frequency DESC
LIMIT 10
;
```

```
+---------+--------+------------+-----------------+
| Word_ID | Hanzi  | Definition | Usage_Frequency |
+---------+--------+------------+-----------------+
|      24 | 七     | seven      |            0.95 |
|      25 | 八     | eight      |            0.95 |
|      21 | 四     | four       |            0.95 |
|      18 | 一     | one        |            0.95 |
|      23 | 六     | six        |            0.95 |
|      19 | 二     | two        |            0.95 |
|      26 | 九     | nine       |            0.95 |
|      22 | 五     | five       |            0.95 |
|      10 | 什么   | what       |            0.95 |
|      20 | 三     | three      |            0.95 |
+---------+--------+------------+-----------------+
10 rows in set (0.00 sec)
```

```sql
-- Query 7
-- Select the average stroke count of all hanzi
SELECT AVG(Stroke_Count) AS Avg_Stroke_Count_Characters
FROM HanziCharacter
;

-- Query 7.5
-- Select the average stroke count of all radicals
SELECT AVG(HanziCharacter.Stroke_Count) AS Avg_Stroke_Count_Radicals
FROM Radical
JOIN CharacterRadical ON Radical.Radical_ID = CharacterRadical.Radical_ID
JOIN HanziCharacter ON CharacterRadical.Character_ID = HanziCharacter.Character_ID
;
```

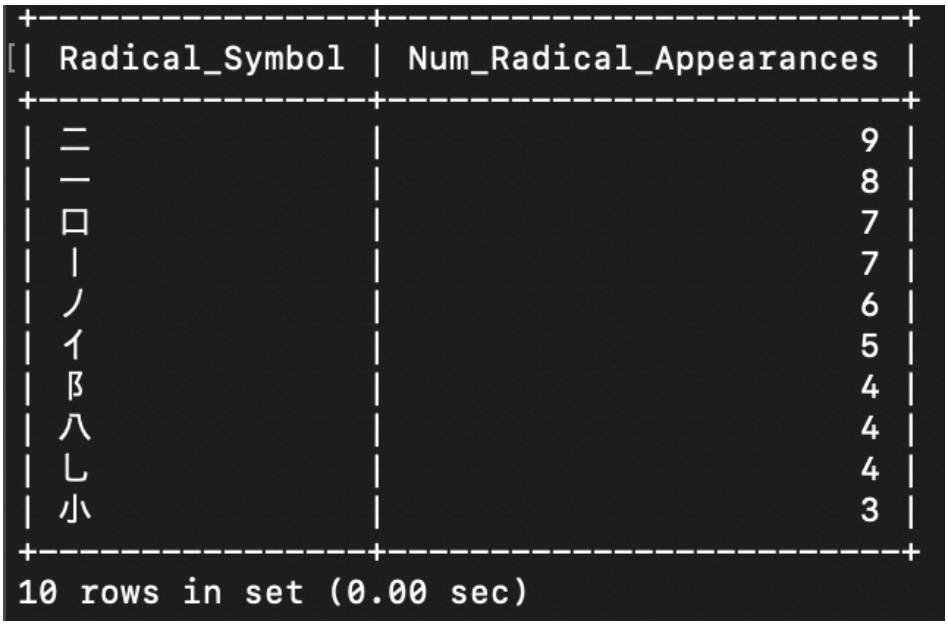
```
+-----------------------------+
| Avg_Stroke_Count_Characters |
+-----------------------------+
|                      6.9136 |
+-----------------------------+
1 row in set (0.00 sec)

+--------------------------+
| Avg_Stroke_Count_Radicals |
+--------------------------+
|                   6.9612 |
+--------------------------+
1 row in set (0.00 sec)
```

```sql
-- Query 8
-- Select pairs of related words and their relationship type
SELECT w1.Hanzi AS Word1, w2.Hanzi AS Word2, RelatedWords.Relation_Type
FROM RelatedWords
JOIN Word w1 ON RelatedWords.Word1_ID = w1.Word_ID
JOIN Word w2 ON RelatedWords.Word2_ID = w2.Word_ID
ORDER BY RelatedWords.Relation_Type
;
```

Word1	Word2	Relation_Type
吃	米饭	Action-Object
喝	茶	Action-Object
买	水果	Action-Object
大	小	Antonym
热	冷	Antonym
看	听	Complementary-Actions
这	那	Contrast
爸爸	妈妈	Family-Pair
儿子	女儿	Family-Pair
是	有	Functional-Pair
做	买	Functional-Pair
有	没	Negation
是	不	Negation-Pair
不	没	Negation-Pair
来	去	Opposite-Actions
吃	喝	Paired-Actions
你	你们	Plural-Form
我	我们	Plural-Form
他	他们	Plural-Form
她	她们	Plural-Form
什么	怎么	Question Words
哪儿	谁	Question Words
多少	几	Question Words
苹果	水果	Type-Of

24 rows in set (0.00 sec)

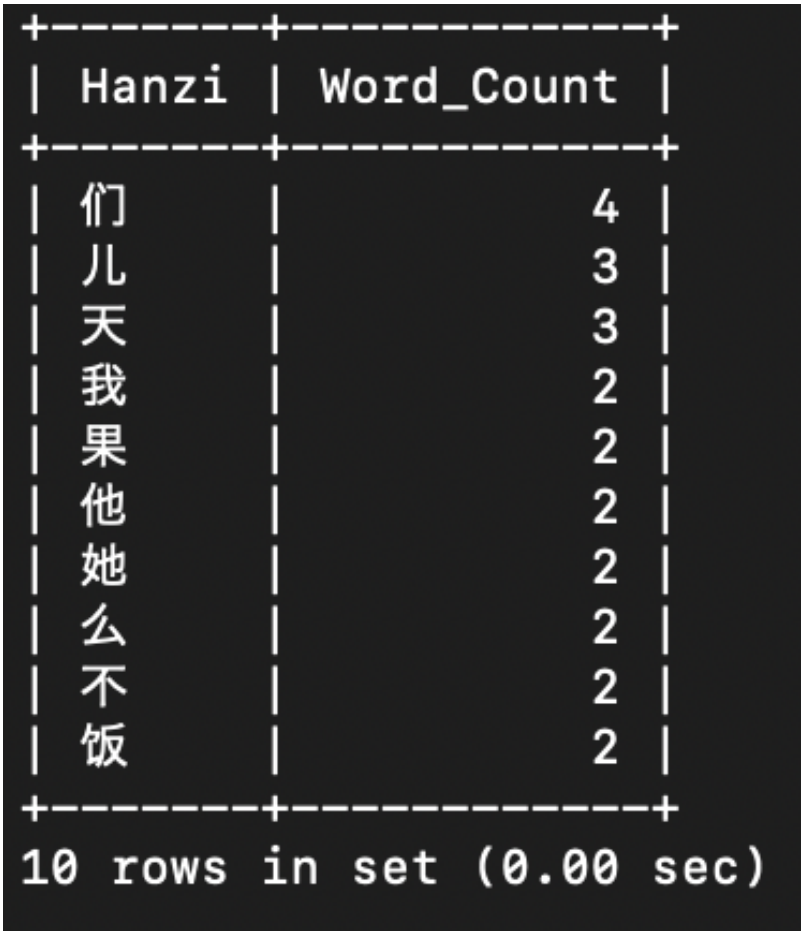
```sql
-- Query 9
-- Select the top 10 radicals used in the most characters
SELECT Radical.Radical_Symbol, COUNT(CharacterRadical.Character_ID) AS Num_Radical_Appearances
FROM Radical
JOIN CharacterRadical ON Radical.Radical_ID = CharacterRadical.Radical_ID
GROUP BY Radical.Radical_Symbol
ORDER BY Num_Radical_Appearances DESC
LIMIT 10
;
```



Radical_Symbol	Num_Radical_Appearances
二	9
一	8
口	7
丨	7
丿	6
亻	5
阝	4
八	4
乚	4
小	3

10 rows in set (0.00 sec)

```sql
-- Query 10
-- Select top 10 characters that appear in more than one word
SELECT HanziCharacter.Hanzi, COUNT(DISTINCT WordCharacter.Word_ID) AS Word_Count
FROM HanziCharacter
JOIN WordCharacter ON HanziCharacter.Character_ID = WordCharacter.Character_ID
GROUP BY HanziCharacter.Character_ID, HanziCharacter.Hanzi
HAVING COUNT(DISTINCT WordCharacter.Word_ID) > 1
ORDER BY Word_Count DESC
LIMIT 10
;
```



Hanzi	Word_Count
们	4
儿	3
天	3
我	2
果	2
他	2
她	2
么	2
不	2
饭	2

10 rows in set (0.00 sec)