

```

1  import logging
2  import re
3  import sqlite3
4  from time import time
5  import pandas as pd
6
7  def parse_line(line,
8                ignore_pattern=r'^[%#].*',
9                pattern=r'^(.+)(.+)\[([^\]]+)\](/.+/)'):
10
11     try:
12         match = re.match(pattern, line)
13         if match:
14             traditional = match.group(1)
15             simplified = match.group(2)
16             pinyin = match.group(3)
17             english = match.group(4).strip('/')
18
19             return {
20                 'traditional': traditional,
21                 'simplified': simplified,
22                 'pinyin': pinyin,
23                 'english': english
24             }
25
26         else:
27             raise ValueError(f'Failed to match expected pattern:\n\t {line}')
28     except Exception as e:
29         logging.error(f'{str(e)}')
30         return None
31
32 def load_to_sqlite(filename, table_name, db_file, max_line=-1):
33     con = sqlite3.connect(db_file)
34     cur = con.cursor()
35
36     # cur.execute(f'DROP TABLE IF EXISTS {table_name}')
37     cur.execute(f'delete from {table_name}; ')
38     cur.execute(f'''
39         CREATE TABLE if not exists {table_name} (
40             traditional TEXT,
41             simplified TEXT,
42             pinyin TEXT,
43             english TEXT
44         );
45     ''')
46
47     nline = 0
48     with open(filename, 'r', encoding='utf-8') as f:
49         for line in f:
50             if max_line > 0 and nline > max_line:
51                 break
52
53             if line.startswith('#') or line.startswith('%'):
54                 continue
55
56             entry = parse_line(line)
57             if entry is None:
58                 continue
59
60             nline += 1
61             values = (
62                 entry['traditional'],
63                 entry['simplified'],
64                 entry['pinyin'],
65                 entry['english'] )
66
67             cur.execute(f'INSERT INTO {table_name} VALUES (?, ?, ?, ?);', values)

```

```
68
69
70     con.commit()
71     con.close()
72
73 if __name__ == '__main__':
74     start_ts = time()
75     filename, table_name, db_file = 'cedict_ts.u8', 't_mdbg_dict', 'cc_cedict.db'
76     load_to_sqlite(filename, table_name, db_file, max_line=1000)
77     end_ts = time()
78     print(f"Completed loading file '{filename}' into sqlite db table '{table_name}' in {
79         end_ts-start_ts} sec")
80
81     # verify
82     with sqlite3.connect(db_file) as con:
83         df = pd.read_sql(f"select count(*) from {table_name}", con)
84         print(df.head())
```