```
import pandas as pd
import re
import sqlite3



input_file = '/content/Input_sheet.xlsx'
spec_file = '/content/Specification_File.xlsx'
output_file = '/content/Output_sheet.xlsx'


input_df = pd.read_excel(input_file)
spec_df = pd.read_excel(spec_file)
output_df = pd.read_excel(output_file)



print(input_df.shape)
print(spec_df.shape)
print(output_df.shape)
```

```
(113554, 3)
(3611, 2)
(113554, 22)
```

```
input_df.head()
```

	Publication Number	Application Number	Disclosure-Information
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...
4	EP3516908A4	EP2017890105A	ISLD : ISLD-201705-026 Disc : 8 Project : 3GPP...

```
input_df.columns
```

```
Index(['Publication Number', 'Application Number', 'Disclosure-Information '], dtype='object')
```

```
spec_df.head()
```

	Formatted Specs	Responsible Working Group
0	TR 00.01U	SMG5
1	TR 00.02	-
2	TR 00.02U	SMG5
3	TR 01.00	SP
4	TS 01.01	SP

```
spec_df.columns
```

```
Index(['Formatted Specs', 'Responsible Working Group'], dtype='object')
```

```
output_df.head()
```

	Publication Number	Application Number	Disclosure-Information	Specifications	Responsible Working Group	RAN 1	RAN 2	RAN 3
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 36.331 | TS 38.304 | TS 36.304 | TS 38.331 ...	RAN 2 | RAN 2 | RAN 2 | RAN 2 | RAN 2 | RAN 2 ...	0	8	0
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.214 | TS 38.211 | TS 38.213 | TS 38.331	RAN 1 | RAN 1 | RAN 1 | RAN 2	3	1	0
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.214 | TS 38.211 | TS 38.213	RAN 1 | RAN 1 | RAN 1	3	0	0
			ISLD : ISLD-201705-022	TS 38.214 | TS	RAN 1 | RAN			

```
output_df.columns
```

```
Index(['Publication Number', 'Application Number', 'Disclosure-Information ',
       'Specifications', 'Responsible Working Group', 'RAN 1', 'RAN 2',
       'RAN 3', 'RAN 4', 'RAN 5', 'RAN 6', 'CT WG 1', 'CT WG 3', 'CT WG 4',
       'CT WG 5', 'CT WG 6', 'SA 1', 'SA 2', 'SA 3', 'SA 4', 'SA 5', 'SA 6'],
      dtype='object')
```

```
def extract_specification_info(text):
    if isinstance(text, str):
        start_tag = "Standard :"
        end_tag = "Version :"

        start_pos = text.find(start_tag)
        end_pos = text.find(end_tag, start_pos)

        if start_pos != -1 and end_pos != -1:
            specification_info = text[start_pos + len(start_tag):end_pos].strip()
            return specification_info
        else:
            return None
    else:
        return None

input_df['Specification'] = input_df['Disclosure-Information '].apply(extract_specification_info)
```

```
input_df
```

	Publication Number	Application Number	Disclosure-Information	Specification
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 38.331
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.213|TS 38.331|TS 38.214|TS 38.211
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.213|TS 38.214|TS 38.211
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...	TS 38.213|TS 38.214
4	EP3516908A4	EP2017890105A	ISLD : ISLD-201705-026 Disc : 8 Project : 3GPP...	TS 38.213|TS 38.214|TS 38.321
...	...	...	...	...
113549	201862670247	201862670247	NaN	None
113550	201862673799	201862673799	NaN	None
113551	201862670549	201862670549	NaN	None

```
input_df = input_df[input_df['Specification'].notna()]
input_df
```

	Publication Number	Application Number	Disclosure-Information	Specification
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 38.331
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.213|TS 38.331|TS 38.214|TS 38.211
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.213|TS 38.214|TS 38.211
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...	TS 38.213|TS 38.214
4	EP3516908A4	EP2017890105A	ISLD : ISLD-201705-026 Disc : 8 Project : 3GPP...	TS 38.213|TS 38.214|TS 38.321
...	...	...	...	...

```
duplicate_rows = input_df[input_df.duplicated()]
print("Duplicate rows:")
duplicate_rows
```

```
Duplicate rows:
```

	Publication Number	Application Number	Disclosure-Information	Specification
90410	EP996306B1	EP1999120719A	ISLD : ISLD-201809-300 Disc : 1 Project : 5G S...	TS 38.304|TS 38.331|TS 22.011|TS 38.321|TS 38.300
90411	EP996306B1	EP1999120719A	ISLD : ISLD-201809-300 Disc : 1 Project : 5G S...	TS 38.304|TS 38.331|TS 22.011|TS 38.321|TS 38.300
90412	EP1030484B1	EP2000300640A	ISLD : ISLD-201809-308 Disc : 7 Project : 5G S...	TS 37.324|TS 38.323|TS 38.415|TS 38.322
90413	EP1030484B1	EP2000300640A	ISLD : ISLD-201809-308 Disc : 7 Project : 5G S...	TS 37.324|TS 38.323|TS 38.415|TS 38.322
90414	EP1030484B1	EP2000300640A	ISLD : ISLD-201809-308 Disc : 7 Project : 5G S...	TS 37.324|TS 38.323|TS 38.415|TS 38.322
...	...	...	...	...

```
input_df = input_df.drop_duplicates()
input_df
```

	Publication Number	Application Number	Disclosure-Information	Specification
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 38.331
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.213|TS 38.331|TS 38.214|TS 38.211
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.213|TS 38.214|TS 38.211
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...	TS 38.213|TS 38.214
4	EP3516908A4	EP2017890105A	ISLD : ISLD-201705-026 Disc : 8 Project : 3GPP...	TS 38.213|TS 38.214|TS 38.321
...	...	...	...	...

```
unique_entries_count = input_df['Specification'].nunique()
print(f"Number of unique entries in 'Specification' column: {unique_entries_count}")
unique_entries = input_df['Specification'].unique()
print("Unique entries in 'Specification' column:")
for entry in unique_entries:
    print(entry)
```

```
TS 38.213|TS 38.211|TS 38.321
TS 38.213|TS 38.211|TS 38.300|TS 38.214
TS 38.212|TS 38.213|TS 38.211|TS 38.214|TS 38.300
TS 38.322|TS 38.331|TS 38.423|TS 38.321|TS 38.413
TS 29.118|TS 29.118
TS 38.331|TS 38.212|TS 38.331|TS 38.211|TS 38.213
```

```
input_df = input_df.drop_duplicates(subset=['Specification'])
print("DataFrame with unique 'Specification':")
input_df
```

```
DataFrame with unique 'Specification':
```

	Publication Number	Application Number	Disclosure-Information	Specification
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 38.331
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.213|TS 38.331|TS 38.214|TS 38.211
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.213|TS 38.214|TS 38.211
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...	TS 38.213|TS 38.214
4	EP3516908A4	EP2017890105A	ISLD : ISLD-201705-026 Disc : 8 Project : 3GPP...	TS 38.213|TS 38.214|TS 38.321
...	...	...	...	...
83972	US6865262B1	US2001889310A	ISLD : ISLD-201812-005 Disc : 11 Project : 5G ...	TS 29.658|TS 24.647|TS 24.647|TS 29.658
84127	US6816478B1	US2000706132A	ISLD : ISLD-201809-258 Disc : 2 Project : 5G S...	TS 36.213|TS 36.211|TS 36.212

```
spec_df.columns
```

```
Index(['Formatted Specs', 'Responsible Working Group'], dtype='object')
```

```
filtered_df = spec_df[spec_df['Formatted Specs'].str.contains(r'^TS \d', regex=True)]
print("Filtered rows:")
filtered_df
```

```
Filtered rows:
```

	Formatted Specs	Responsible Working Group
4	TS 01.01	SP
6	TS 01.02	SA 1
7	TS 01.03	-
13	TS 01.06	-
14	TS 01.07	-
...	...	...
3599	TS 55.242	SA 3
3600	TS 55.243	SA 3
3601	TS 55.251	SA 3
3602	TS 55.252	SA 3
3603	TS 55.253	SA 3

2272 rows × 2 columns

```
def map_spec_to_group(spec_list, filtered_df):
    groups = []
    for spec in spec_list.split('|'):
        group = filtered_df.loc[filtered_df['Formatted Specs'] == spec.strip(), 'Responsible Working Group'].values
        if len(group) > 0:
            groups.append(group[0])
        else:
            groups.append(None)
    return '|'.join([g for g in groups if g])
```

```
input_df['Responsible Working Group'] = input_df['Specification'].apply(lambda x: map_spec_to_group(x, filtered_df))
print("Updated input_df:")
input_df
```

```
Updated input_df:
<ipython-input-27-da5ee829404d>:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_
  input_df['Responsible Working Group'] = input_df['Specification'].apply(lambda x: map_
```

	Publication Number	Application Number	Disclosure-Information	Specification	Responsible Working Group
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 38.331	RAN 2
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.213|TS 38.331|TS 38.214|TS 38.211	RAN 1|RAN 2|RAN 1|RAN 1
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.213|TS 38.214|TS 38.211	RAN 1|RAN 1|RAN 1
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...	TS 38.213|TS 38.214	RAN 1|RAN 1

```
unique_elements = set(input_df['Responsible Working Group'].str.split('|').explode())
unique_elements.discard('')

print("Unique elements in Responsible Working Group:")
for element in unique_elements:
    print(element)
```

```
Unique elements in Responsible Working Group:
CT WG 4
SA 2
RAN 3
RAN 2
RAN 5
RAN 6
SA 6
SA 5
CT WG 3
CT WG 1
RAN 4
SA 1
RAN 1
SA 3
CT WG 6
SA 4
```

```
unique_elements = set(input_df['Responsible Working Group'].str.split('|').explode())

unique_elements.discard('')

for element in unique_elements:
    input_df[element] = 0

for idx, row in input_df.iterrows():
    elements = row['Responsible Working Group'].split('|')
    for element in elements:
        if element in unique_elements:
            input_df.at[idx, element] = elements.count(element)

print("Updated input_df with occurrence counts:")
input_df
```

```
<ipython-input-30-b152770ca812>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cc
  input_df[element] = 0
Updated input_df with occurrence counts:
```

	Publication Number	Application Number	Disclosure-Information	Specification	Responsible Working Group	CT WG 4	SA 2	RAN 3	RAN 2	RAN 5	...	SA 6	SA 5	CT WG 3	CT WG 1	RAN 4	SA 1	RAN 1	SA 3
0	EP3277024B1	EP2017189488A	ISLD : ISLD-201805-014 Disc : 92 Project : 3GP...	TS 38.331	RAN 2	0	0	0	1	0	...	0	0	0	0	0	0	0	0
1	EP3544198A1	EP2019172955A	ISLD : ISLD-201704-009 Disc : 5 Project : 3GPP...	TS 38.213|TS 38.331|TS 38.214|TS 38.211	RAN 1|RAN 2|RAN 1|RAN 1	0	0	0	1	0	...	0	0	0	0	0	0	3	0
2	EP3544360A1	EP2019167098A	ISLD : ISLD-201705-021 Disc : 19 Project : 3GP...	TS 38.213|TS 38.214|TS 38.211	RAN 1|RAN 1|RAN 1	0	0	0	0	0	...	0	0	0	0	0	0	3	0
3	EP3189648B1	EP2016847578A	ISLD : ISLD-201705-022 Disc : 27 Project : 3GP...	TS 38.213|TS 38.214	RAN 1|RAN 1	0	0	0	0	0	...	0	0	0	0	0	0	2	0
4	EP3516908A4	EP2017890105A	ISLD : ISLD-201705-026 Disc : 8 Project : 3GPP...	TS 38.213|TS 38.214|TS 38.321	RAN 1|RAN 1|RAN 2	0	0	0	1	0	...	0	0	0	0	0	0	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
			ISLD : ISLD-201812-005	TS 29.658|TS	CT WG 3|CT														