DickensAssignmentValidator

MUCEP Task 1 (Dr. Pierre-Paul Bitton)

Author: Shawon Ibn Kamal
Email: sikamal@mun.ca (mailto:sikamal@mun.ca)

Updates to exisiting program

I made a few changes in the existing curator program to work with it efficiently and sorted out a few bugs. Here's the list:

- The files getting read from is renamed to "DataFiles" from "Files".
- The output csv files are being stored in a folder named "OutputFiles".
- Renamed "NotMatched.csv" to "MissingMeta.csv" in order to avoid confusion with "MissingFiles.csv".
- Stored the program in git, currently a private repo to me. I think it is a good way to track updates, we can work on it if you are interested
- Fixed few minor bugs in DickensAssignment.py program

In [140]:

```
import pandas as pd
import numpy as np
```

Run DickensAssignment.py

In [141]:

```
exec(open('DickensAssignment.py').read())
```

```
4905 no. of files
4093 match found
812 match not found
Complete
```

Compare OutputFiles with OutputFiles_2020_07_14

In [142]:

```
# Load old outputs
df_old_result = pd.read_csv('OutputFiles_2020_07_14/Result.csv', engine='python')
df_old_missing_files = pd.read_csv('OutputFiles_2020_07_14/MissingFiles.csv', engine='python')
df_old_missing_meta = pd.read_csv('OutputFiles_2020_07_14/MissingMeta.csv', engine='python')

# Load new outputs
df_new_result = pd.read_csv('OutputFiles/Result.csv', engine='python')
df_new_missing_files = pd.read_csv('OutputFiles/MissingFiles.csv', engine='python')
df_new_missing_meta = pd.read_csv('OutputFiles/MissingMeta.csv', engine='python')

# Load filenames
filenames = [name for path, subdirs, files in os.walk("DataFiles")
             for name in files]

df_data_files = pd.DataFrame({'filename':filenames}).sort_values(by='filename')

# Load template
df_template = pd.read_csv('template.csv', engine='python')

# Sort Result
df_old_result = df_old_result.sort_values(by='FileName')
df_new_result = df_new_result.sort_values(by='FileName')
```

In [143]:

```
df_diff_result = pd.concat([df_old_result.dropna(axis=0),df_new_result.dropna(axis=0)], sort=True).drop_duplicat
df_diff_missing_files = pd.concat([df_old_missing_files,df_new_missing_files], sort=True).drop_duplicates(keep=F
df_diff_missing_meta = pd.concat([df_old_missing_meta,df_new_missing_meta], sort=True).drop_duplicates(keep=Fals

if (df_diff_result.shape[0] == 0):
    print("Results are the same")
else:
    print("Results have ", df_diff_result.shape[0], " differences")

if (df_diff_missing_files.size == 0):
    print("MissingFiles are the same")
else:
    print("MissingFiles have ", df_diff_missing_files.shape[0], " differences")

if (df_new_missing_meta.shape[0] == 0):
    print("NotMatchedFiles are the same")
else:
    print("NotMatchedFiles have ", df_diff_missing_meta.shape[0], " differences")
```

```
Results are the same
MissingFiles are the same
NotMatchedFiles have  0  differences
```

Check to see if MissingMetaData entries are due to typo

In [144]:

```
def includes(fullstring, substrings=[]):
    count = 0
    for each_substring in substrings:
        if fullstring.find(each_substring) != -1:
            count += 1
    return count

# Testing
print(includes("I like data", ["like", "data"]))
```

```
2
```

In [145]:
```python
df_template['key'] = 0
df_new_missing_meta['key'] = 0

# Cartessian product of two dataframes
df_merged_template_and_missing_meta = df_template.merge(df_new_missing_meta, how='outer')
```

In [146]:
```python
df_merged_template_and_missing_meta['similarity'] = df_merged_template_and_missing_meta.apply(lambda row : inclu
```

In [147]:
```python
df_merged_template_and_missing_meta = df_merged_template_and_missing_meta[['institutionCode', 'catalogueNumber',
print(df_merged_template_and_missing_meta['notmatched'].count())

# Export data
df_merged_template_and_missing_meta.to_csv('ValidatorExports/MissingMetaSimilar.csv', index=False)

# Print first 50 data
df_merged_template_and_missing_meta.head(10)
```

```
27
```

Out[147]:

	institutionCode	catalogueNumber	notmatched	similarity
43377	MNRJ	4359	CH.R.MNRJ44359.00000005.csv	2
43376	MNRJ	4359	CH.R.MNRJ44359.00000002.csv	2
23167	CM	72696	AM.U.CM972696.00000005.Master.Transmission	2
23166	CM	72696	AM.U.CM972696.00000004.Master.Transmission	2
23165	CM	72696	AM.U.CM972696.00000003.Master.Transmission	2
23164	CM	72696	AM.U.CM972696.00000002.Master.Transmission	2
23163	CM	72696	AM.U.CM972696.00000001.Master.Transmission	2
23162	CM	72696	AM.T.CM972696.00000005.Master.Transmission	2
23161	CM	72696	AM.T.CM972696.00000004.Master.Transmission	2
23160	CM	72696	AM.T.CM972696.00000003.Master.Transmission	2

In [148]:

```
# Find no similarities at all
df_merged_template_and_missing_meta = df_merged_template_and_missing_meta.drop_duplicates('notmatched')
df_missing_meta_nonsimilar = pd.concat([df_new_missing_meta['notmatched'],df_merged_template_and_missing_meta['

# Export non-similar data
df_missing_meta_nonsimilar.to_csv('ValidatorExports/MissingMetaNonSimilar.csv', index=False, header=True)

# Print first 50 data
df_missing_meta_nonsimilar.head(10)
```

Out[148]:

```
0    TE.F.B.LSU180686.00000001.Master.Transmission
1    TE.F.B.LSU180686.00000002.Master.Transmission
2    TE.F.B.LSU180686.00000003.Master.Transmission
3    TE.F.B.LSU180686.00000004.Master.Transmission
4    TE.F.B.LSU180686.00000005.Master.Transmission
5    TE.F.B.LSU180687.00000001.Master.Transmission
6    TE.F.B.LSU180687.00000002.Master.Transmission
7    TE.F.B.LSU180687.00000003.Master.Transmission
8    TE.F.B.LSU180687.00000004.Master.Transmission
9    TE.F.B.LSU180687.00000005.Master.Transmission
Name: notmatched, dtype: object
```

Find similar files for MissingFiles

In [149]:

```
df_data_files['key'] = 0
df_new_missing_files['key'] = 0

# Cartessian product of two dataframes
df_merged_data_files_and_missing_files = df_data_files.merge(df_new_missing_files, how='outer')
df_merged_data_files_and_missing_files.head()
```

Out[149]:

	filename	key	FileName	institutionCode	collectionCode	catalogueNumber	class	order
0	AM.H.AMNH278606.00000001.Master.Transmission	0	NaN	MZUSP	NaN	97287	Aves	Trogoniformes
1	AM.H.AMNH278606.00000001.Master.Transmission	0	NaN	MZUSP	NaN	76792	Aves	Trogoniformes
2	AM.H.AMNH278606.00000001.Master.Transmission	0	NaN	MZUSP	NaN	86474	Aves	Trogoniformes
3	AM.H.AMNH278606.00000001.Master.Transmission	0	NaN	MCZ	NaN	173836	Aves	Trogoniformes
4	AM.H.AMNH278606.00000001.Master.Transmission	0	NaN	MZUSP	NaN	15953	Aves	Trogoniformes

5 rows × 29 columns

In [150]:

```
# Calculate similarity
df_merged_data_files_and_missing_files['similarity'] = df_merged_data_files_and_missing_files.apply(lambda row
```

In [151]:

```
# Sort
df_merged_data_files_and_missing_files = df_merged_data_files_and_missing_files[['institutionCode', 'catalogueNu
print(df_new_missing_files.shape[0])
print("Length of similarities", df_merged_data_files_and_missing_files.shape[0])

# Export data
df_merged_data_files_and_missing_files.to_csv('ValidatorExports/MissingFilesSimilarity.csv', index=False)

# Print first 50 data
df_merged_data_files_and_missing_files.head(10)
```

```
64
Length of similarities 595
```

Out[151]:

	index	institutionCode	catalogueNumber	filename	similarity
0	222894	LSUMNS	114719	SU.S.LSU114719.00000003.Master.Transmission	1
1	243312	LSUMNS	161602	SU.U.LSU161602.00000002.Master.Transmission	1
2	243440	LSUMNS	161602	SU.U.LSU161602.00000004.Master.Transmission	1
3	243504	LSUMNS	161602	SU.U.LSU161602.00000005.Master.Transmission	1
4	243572	LSUMNS	71304	SU.U.LSU71304.00000001.Master.Transmission	1
5	243636	LSUMNS	71304	SU.U.LSU71304.00000002.Master.Transmission	1
6	243700	LSUMNS	71304	SU.U.LSU71304.00000003.Master.Transmission	1
7	243764	LSUMNS	71304	SU.U.LSU71304.00000004.Master.Transmission	1
8	243828	LSUMNS	71304	SU.U.LSU71304.00000005.Master.Transmission	1
9	243883	LSUMNS	87590	SU.U.LSU87590.00000001.Master.Transmission	1

In [152]:

```
# Find no similarities at all
df_merged_data_files_and_missing_files_unique = df_merged_data_files_and_missing_files.drop_duplicates(subset=[
df_missing_files_nonsimilar = pd.concat([df_new_missing_files[['institutionCode', 'catalogueNumber']],df_merged_

# Export non-similar data
df_missing_files_nonsimilar.to_csv('ValidatorExports/MissingFilesNonSimilar.csv', index=False, header=True)

print("There are ", df_missing_files_nonsimilar.shape[0], " meta data with no similarities.")

# Print first 50 data
df_missing_files_nonsimilar.head(10)
```

```
There are  43  meta data with no similarities.
```

Out[152]:

	institutionCode	catalogueNumber
0	MZUSP	97287
1	MZUSP	76792
2	MZUSP	86474
3	MCZ	173836
4	MZUSP	15953
5	MZUSP	44168
6	MZUSP	44172
7	MZUSP	44175
8	MCZ	173842
9	MCZ	173839

In [ ]: