21_fix_fce_prompt_ids

March 21, 2017

02/26/17 - smiel ### I messed up the prompt IDs in FCE when cleaning that data set. It's easier to fix it here than it is to re-run everything.

```
In [1]: import pandas as pd
        import os

        path = '/research/ella/rivendell'
```

```
In [7]: fce = pd.read_csv(os.path.join(path, 'fce', 'all_essays.csv'), encoding='utf8')
        fce['fixed_prompt_id'] = fce.apply(lambda row: '{}.{}'.format(row['exam_id'], row['promp
        fce.index = fce.essay_id
        fce.head()
```

```
Out[7]:                       age           essay_id     exam_id language prompt_id  \
        essay_id
        0101_2000_6_837_1   16-20   0101_2000_6_837_1  0101_2000_6    Greek         1
        0101_2000_6_837_2   16-20   0101_2000_6_837_2  0101_2000_6    Greek         3
        0101_2000_6_1156_1    <16  0101_2000_6_1156_1  0101_2000_6    Greek         1
        0101_2000_6_1156_2    <16  0101_2000_6_1156_2  0101_2000_6    Greek         3
        0101_2000_6_751_1   16-20   0101_2000_6_751_1  0101_2000_6    Greek         1

                            score score_type  student_id  \
        essay_id
        0101_2000_6_837_1    22.0        FCE         837
        0101_2000_6_837_2    22.0        FCE         837
        0101_2000_6_1156_1   27.0        FCE        1156
        0101_2000_6_1156_2   27.0        FCE        1156
        0101_2000_6_751_1    25.0        FCE         751

                                                                         text   L1  \
        essay_id
        0101_2000_6_837_1   Dear Mrs Brown,\nI am writing to give you info...  GRC
        0101_2000_6_837_2   It was Friday morning when I saw John and said...  GRC
        0101_2000_6_1156_1  Dear Mrs Brown,\nI am one of your husband's st...  GRC
        0101_2000_6_1156_2  John said he had some good news to tell me. I ...  GRC
        0101_2000_6_751_1   Dear Mrs Brown,\nIt would be a pleasure to us ...  GRC

                           fixed_prompt_id
```

```
essay_id
0101_2000_6_837_1     0101_2000_6.1
0101_2000_6_837_2     0101_2000_6.3
0101_2000_6_1156_1    0101_2000_6.1
0101_2000_6_1156_2    0101_2000_6.3
0101_2000_6_751_1     0101_2000_6.1
```

```
In [11]: # if original versions exist, use those
         orig = '.orig_21'

         # essays
         essays = os.path.join(path, 'all_essays.csv')
         if os.path.exists(essays + orig):
             df = pd.read_csv(essays + orig, encoding='utf8')
         else:
             df = pd.read_csv(essays, encoding='utf8')

         print('{} essays'.format(len(df)))

         # make a copy
         df.to_csv(essays + orig, encoding='utf8', index=False)

         # fix prompt id
         df.prompt_id = df.apply(
             lambda row: row.prompt_id if row.dataset != 'FCE' else fce.loc[row.essay_id, 'fixed
             axis=1
         )

         # save
         df.to_csv(essays, encoding='utf8', index=False)

         print('{} essays'.format(len(df)))

         # sentences
         sentences = os.path.join(path, 'all_sentences.csv')
         if os.path.exists(sentences + orig):
             df = pd.read_csv(sentences + orig, encoding='utf8')
         else:
             df = pd.read_csv(sentences, encoding='utf8')

         print('{} sentences'.format(len(df)))

         # make a copy
         df.to_csv(sentences + orig, encoding='utf8', index=False)

         # fix prompt id
         df.prompt_id = df.apply(
             lambda row: row.prompt_id if row.dataset != 'FCE' else fce.loc[row.essay_id, 'fixed
             axis=1
         )

         # save
         df.to_csv(sentences, encoding='utf8', index=False)

         print('{} sentences'.format(len(df)))
```

```
/home/smiel/.venvs/rivendell/local/lib/python2.7/site-packages/IPython/core/interactiveshell.py:
  interactivity=interactivity, compiler=compiler, result=result)


111755 essays
111755 essays


/home/smiel/.venvs/rivendell/local/lib/python2.7/site-packages/IPython/core/interactiveshell.py:
  interactivity=interactivity, compiler=compiler, result=result)


1870084 sentences
1870084 sentences
```

```
In [12]: # check work
         df = pd.read_csv(essays, encoding='utf8')
         df[df.dataset == 'FCE'].head()
```

```
Out[12]:         L1  age  alt_score alt_score_type dataset             essay_id  \
         21884  CAT  NaN        NaN            NaN     FCE  0100_2000_12_1000_1
         21885  CAT  NaN        NaN            NaN     FCE  0100_2000_12_1000_2
         21886  KOR  NaN        NaN            NaN     FCE  0100_2000_12_1002_1
         21887  KOR  NaN        NaN            NaN     FCE  0100_2000_12_1002_2
         21888  PRT  NaN        NaN            NaN     FCE  0100_2000_12_1018_1

                     prompt_id  score score_type student_level_A_Level  \
         21884  0100_2000_12.1   28.0        FCE                   NaN
         21885  0100_2000_12.2   28.0        FCE                   NaN
         21886  0100_2000_12.1   32.0        FCE                   NaN
         21887  0100_2000_12.3   32.0        FCE                   NaN
         21888  0100_2000_12.1   21.0        FCE                   NaN

                          ...              student_level_HKALE student_level_IELTS  \
         21884            ...                              NaN                 NaN
         21885            ...                              NaN                 NaN
         21886            ...                              NaN                 NaN
         21887            ...                              NaN                 NaN
         21888            ...                              NaN                 NaN
```

```
       student_level_NMET student_level_ONET  student_level_O_Level  \
21884                 NaN                NaN                    NaN   
21885                 NaN                NaN                    NaN   
21886                 NaN                NaN                    NaN   
21887                 NaN                NaN                    NaN   
21888                 NaN                NaN                    NaN   

       student_level_TEPS  student_level_TOEFL student_level_TOEIC  \
21884                 NaN                  NaN                 NaN   
21885                 NaN                  NaN                 NaN   
21886                 NaN                  NaN                 NaN   
21887                 NaN                  NaN                 NaN   
21888                 NaN                  NaN                 NaN   

                                                    text  \
21884  DECEMBER 12TH PRINCIPAL MR. ROBERTSON DEAR SIR...   
21885  (FAMOUS PEOPLE, SUCH AS POLITICIANS AND FILM S...   
21886  To Mr. Robertson I am writing to tell you some...   
21887  In a country like the UK, we are all bound to ...   
21888  Dear Mr Robertson, Regarding the programme you...   

                             uid  
21884  FCE_0100_2000_12_1000_1_0  
21885  FCE_0100_2000_12_1000_2_0  
21886  FCE_0100_2000_12_1002_1_0  
21887  FCE_0100_2000_12_1002_2_0  
21888  FCE_0100_2000_12_1018_1_0  

[5 rows x 23 columns]
```