07_moecs_clean

March 21, 2017

2/24/17 - smiel

1 Cleaning the MOECS essay data.

```
In [1]: %matplotlib inline
        # run me when first starting this notebook
        import os

        import numpy as np
        import pandas as pd

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

The MOECS data comes to us in a bunch of text files, in L1 folders. We'll use these to create an essays csv.

```
In [5]: import codecs
        recs = []
        l1_map = {'Japanse_L2_English_final data_for windows': 'JPN', 'English_L1_final data': '

        for lang, l1 in l1_map.items():
            lang_path = os.path.join(path, lang)
            for filename in os.listdir(lang_path):
                if filename.startswith('._'):
                    continue

                rec = {'dataset': 'MOECS', 'essay_id': filename}

                rec['prompt_id'] = 'MOECS'
                rec['L1'] = l1

                with codecs.open(os.path.join(lang_path, filename), 'r', encoding='ISO-8859-2')
                    text = fin.read().strip().replace('<P>', '').replace('</P>', '')

                # for some reason, a few of the files have the id in them
                eid = filename[:-4]
                text = text.replace(eid, '')
```

```
            rec['text'] = text
            recs.append(rec)

    df = pd.DataFrame.from_records(recs)
    df.to_csv(os.path.join(path, 'all_essays.csv'), encoding='ISO-8859-2', index=False)
    df.head()
```

Out[5]:

	L1	dataset	essay_id	prompt_id	text
0	JPN	MOECS	JPE042en.txt	MOECS	The importance of information\r\nI agree the o...
1	JPN	MOECS	JPE079en.txt	MOECS	Convinience and responssibillity\nThere are bo...
2	JPN	MOECS	JPE003en.txt	MOECS	Disadvantege of Internet\nMany people uses int...
3	JPN	MOECS	JPE071en.txt	MOECS	Get the news from paper is good\nI donâĂŹt bel...
4	JPN	MOECS	JPE075en.txt	MOECS	Necessary of newspapers\nNewspapers are need f...

```
In [6]: df[df.essay_id == 'JPE071en.txt'].text.values[0]
```

```
Out[6]: u'Get the news from paper is good\nI don\xe2\x80\x99t believe it doesn\xe2\x80\x99t need
```

1.1 From Essays to Sentences

Now let's start building the sentences data frame. For unidecode to work properly, the following should print "True":

```
In [11]: import sys
         print(sys.maxunicode > 0xffff)
```

```
True
```

```
In [8]: from utilitybelt.text import get_sentences
        import copy
        from unidecode import unidecode
        import numpy as np

        # load data
        df_in = pd.read_csv(os.path.join(path, 'all_essays.csv'), encoding='ISO-8859-2')

        # convert text to ascii
        print('Converting to ASCII')
        df_in['ascii_text'] = df_in.text.apply(lambda t: unidecode(t))

        # normalize line endings
```

```
df_in.ascii_text = df_in.ascii_text.str.replace('\r\n', '\n')
df_in.ascii_text = df_in.ascii_text.str.replace('\r', '\n')

# use space instead of tab
df_in.ascii_text = df_in.ascii_text.str.replace('\t', ' ')

# now remove any non-printable ascii char
df_in.ascii_text = df_in.ascii_text.str.replace(r'[^ -~\n]', '')

# # make sure all is printable
# for i, t in enumerate(df_in.ascii_text.values):
#     for ci, c in enumerate(t):
#         if (32 <= ord(c) <= 126) or c in '\n\t':
#             continue
#         else:
#             print u"Unprintable character {} in {} at char {}:\n\n{}\n================
#                 ord(c), i, ci, t, df_in.iloc[i].clean_text
#             )
#             raise ValueError

# shush the utilitybelt sentence splitter logging
import logging
logger = logging.getLogger()
logger.setLevel(logging.INFO)

def asfloat(x):
    try:
        return float(x)
    except:
        print '{} not a float'.format(x)
        return np.nan

print('Splitting sentences')
# create records for every sentence
records = []
for i, row in df_in.iterrows():
    rec = {
        'dataset': row.dataset, 'prompt_id': row.prompt_id, 'essay_id': row.essay_id,
        'L1': row.L1, 'score': np.nan, 'score_type': '', 'age': np.nan,
    }
    prev_end = 0
    text = row.ascii_text
    si = 0
    for start, end, sentence in zip(*get_sentences(text)):
        srec = {}
        srec.update(rec)
        srec['text'] = sentence
        srec['sentence_id'] = si
```

```
                srec['trailing_whitespace'] = text[prev_end:start]
                si += 1
                prev_end = end
                records.append(srec)

            if i % 1000 == 0:
                print('{} of {}'.format(i, len(df_in)))

        print('Creating data frame')
        df_out = pd.DataFrame.from_records(records)
        df_out['uid'] = df_out[['dataset', 'essay_id', 'sentence_id']].astype(unicode).apply(lam

        print('{} sentences'.format(len(df_out)))
        print('Saving data frame')
        df_out.to_csv(os.path.join(path, 'MOECS_sentences.csv'), encoding='utf8', index=False)
        df_out.head()
```

```
Converting to ASCII
Splitting sentences
0 of 199
Creating data frame
4523 sentences
Saving data frame
```

Out[8]:

	L1	age	dataset	essay_id	prompt_id	score	score_type	sentence_id	\
0	JPN	NaN	MOECS	JPE042en.txt	MOECS	NaN		0	
1	JPN	NaN	MOECS	JPE042en.txt	MOECS	NaN		1	
2	JPN	NaN	MOECS	JPE042en.txt	MOECS	NaN		2	
3	JPN	NaN	MOECS	JPE042en.txt	MOECS	NaN		3	
4	JPN	NaN	MOECS	JPE042en.txt	MOECS	NaN		4	

	text	trailing_whitespace	\
0	The importance of information\nI agree the opi...		
1	Recently we can use Internet freely and many p...	\n	
2	Internet give us many information, so we can k...		
3	Internet society have good effect for us and o...	\n	
4	Although we should think more about Internet s...		

	uid
0	MOECS_JPE042en.txt_0
1	MOECS_JPE042en.txt_1
2	MOECS_JPE042en.txt_2
3	MOECS_JPE042en.txt_3
4	MOECS_JPE042en.txt_4

Let's do a little descriptive analysis to make sure we got what we want.

```
In [9]: df = pd.read_csv(os.path.join(path, 'MOECS_sentences.csv'), encoding='utf8')
```

```
In [10]: age = df.groupby('age').size()
         print(age)
         print('{} sentences with age data'.format(pd.notnull(df.age).sum()))
```

```
Series([], dtype: int64)
0 sentences with age data
```

```
In [11]: score = df.groupby('score').size()
         print(score)
```

```
Series([], dtype: int64)
```

```
In [12]: df.text.apply(len).describe()
```

```
Out[12]: count    4523.000000
         mean      107.296485
         std        58.455817
         min         4.000000
         25%        65.000000
         50%        97.000000
         75%       137.000000
         max       464.000000
         Name: text, dtype: float64
```

```
In [13]: df.trailing_whitespace = df.trailing_whitespace.fillna('')
         essay1_id = df.essay_id.values[0]
         essay1 = df[df.essay_id == essay1_id]
         essay1['text_plus'] = essay1.trailing_whitespace + essay1.text
         text = ''.join(essay1.text_plus.values)
         print(text)
         print(essay1_id)
```

```
The importance of information
I agree the opinion that necessary newspaper or magazine now and than.
Recently we can use Internet freely and many people can look the news on the Internet. Internet
Internet society have good effect for us and our life is comfortable than the society of not Int
Firstly the information is most important our life. Our society was moved by the information so
If donat have newspaper or magazine, only Internet user can get information therefor non-Interne
Secondly, Internet have numerous information, so we confuse, if the world only have internet inf
Newspaper or Magazine are published by relief company, so we can get truth information and can c
Finally not only the Internet information society is to make equally. It is because not all the
These are my opinion that newspaper or magazine are needed our society. Internet equals informat
Therefor we necessary newspaper or magazine.
JPE042en.txt
```

```
/home/smiel/.venvs/rivendell/lib/python2.7/site-packages/ipykernel/__main__.py:4: SettingWithCop
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: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
```

```
In [14]: df.trailing_whitespace = df.trailing_whitespace.fillna('')
         df.trailing_whitespace = df.trailing_whitespace.str.replace(r'\n  $', '\n')
         essay1_id = df.essay_id.values[0]
         essay1 = df[df.essay_id == essay1_id]
         essay1['text_plus'] = essay1.trailing_whitespace + essay1.text
         text = ''.join(essay1.text_plus.values)
         print(text)
         print(essay1_id)
```

```
The importance of information
I agree the opinion that necessary newspaper or magazine now and than.
Recently we can use Internet freely and many people can look the news on the Internet. Internet
Internet society have good effect for us and our life is comfortable than the society of not Int
Firstly the information is most important our life. Our society was moved by the information so
If donat have newspaper or magazine, only Internet user can get information therefor non-Interne
Secondly, Internet have numerous information, so we confuse, if the world only have internet inf
Newspaper or Magazine are published by relief company, so we can get truth information and can c
Finally not only the Internet information society is to make equally. It is because not all the
These are my opinion that newspaper or magazine are needed our society. Internet equals informat
Therefor we necessary newspaper or magazine.
JPE042en.txt
```

```
/home/smiel/.venvs/rivendell/lib/python2.7/site-packages/ipykernel/__main__.py:5: SettingWithCop
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: http://pandas.pydata.org/pandas-docs/stable/indexing.html#
```