06_icnale_clean

March 21, 2017

2/20/17 - smiel

1 Cleaning the ICNALE 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/icnale'
```

The ICNALE data comes to us in a bunch of text files, with a byzantine naming scheme. It also has a file of metadata. We'll use all of this to build an essays csv. I'll document the structure here:

The text files are located in: path / ICNALE_SW_1.1_Texts / ICNALE_SW_1.1_Unmerged Texts / ICNALE_SW_1.1_Text_{spoken (S) or written (W)}_{L1}_{CEFR}_N{file count} / {spoken (S) or written (W)}_{L1}_{prompt id}_{uid}_CEFR}.txt

In the meta data, we have: - Code : {spoken (S) or written (W)}_{L1}_{uid} - CEFR : {CEFR} -- except for English speakers - Age : {Age}

Our approach to collecting the texts will be to walk through the folders, parse the metadata from the filename, and then use that data to look up age in the meta data csv.

One warning: There appears to be at least 1 copy (i.e. ends with "(1).txt"). We will need to ignore those.

```
In [8]: meta = pd.read_csv(os.path.join(path, 'ICNALE_SW_V1.1_Infosheet.csv'), encoding='utf8')
        print(meta.columns)
        print(meta.head())
        print(meta.groupby('Country').size())
```

```
Index([u'Code', u'Country', u'PTJ1 (wds)', u'PTJ2 (wds)', u'SMK1 (wds)',
       u'SMK2 (wds)', u'Self Ev', u'Sex', u'Age', u'Grade/Degree',
       u'Yrs of Stay (< Yrs)', u'ENS Type', u'Major/ Occupation',
       u'Acad. Genre', u'Test', u'Score', u'VST', u'CEFR', u'INTM', u'INSM',
       u'INTM+INSM', u'INTM-INSM', u'Primary', u'Secondary', u'College',
       u'Inschool', u'Outschool', u'Listening', u'Reading', u'Speaking',
       u'Writing', u'NS', u'Pronunciation', u'Presentation', u'EssayW'],
```

```
      dtype='object')
        Code Country  PTJ1 (wds)  PTJ2 (wds)  SMK1 (wds)  SMK2 (wds)  Self Ev  \
0  S_CHN_001     CHN          99       129.0         109       108.0      1.0
1  S_CHN_002     CHN          52        59.0          43        49.0      1.0
2  S_CHN_003     CHN          95       107.0          90        98.0      1.0
3  S_CHN_004     CHN          77        81.0          69        90.0      0.0
4  S_CHN_005     CHN          95       102.0          93       102.0      3.0

  Sex  Age Grade/Degree   ...   Inschool Outschool Listening Reading Speaking  \
0   M   20            1   ...       4.11      3.33      4.50     4.0     3.00
1   M   20            2   ...       4.33      3.56      4.25     4.5     3.50
2   M   19            2   ...       3.00      2.00      3.00     3.0     2.00
3   F   20            3   ...       3.44      3.78      3.50     4.0     3.50
4   M   19            2   ...       2.78      2.78      3.75     3.0     3.25

  Writing  NS Pronunciation  Presentation  EssayW
0    3.75   2             4             4       5
1    4.00   3             4             4       4
2    2.75   1             3             2       4
3    2.75   3             4             4       3
4    1.50   6             2             4       5

[5 rows x 35 columns]
Country
CHN        550
ENS_AUS     33
ENS_CAN     38
ENS_GBR     53
ENS_NIG      1
ENS_NZL     14
ENS_USA    211
HKG        150
IDN        300
JPN        550
KOR        400
PAK        300
PHL        300
SIN        250
THA        450
TWN        300
dtype: int64
```

Ok. Time to go get the essay texts.

```
In [29]: import codecs

         records = []
```

```
        # walk the folders, Luke
        text_root = os.path.join(path, 'ICNALE_SW_1.1_Texts', 'ICNALE_SW_1.1_Unmerged Texts')

        folders = os.listdir(text_root)

        for folder in folders:
            elems = folder.split('_')
            if len(elems) == 9:
                _, _, _, _, ws, L1, cefr_1, cefr_2, _ = elems
                cefr = '{}_{}'.format(cefr_1, cefr_2)
            else:
                _, _, _, _, ws, L1, cefr, _ = elems
                cefr = '{}_{}'.format(cefr[:2], cefr[2:])

            if ws == 'S':
                continue

            assert ws == 'W'

            # hilariously they made a typo in the folder names. We have to fix it here
            if L1 == 'PHR':
                L1 = 'PHL'

            for filename in os.listdir(os.path.join(text_root, folder)):
                if '(1)' in filename:
                    continue
                _ws, _L1, prompt_id, lookup_id, _cefr_1, _cefr_2 = os.path.splitext(filename)[0
                _cefr = '{}_{}'.format(_cefr_1, _cefr_2)
                assert ws == _ws, '({}/{}) {} does not equal {}'.format(folder, filename, ws, _
                assert L1 == _L1, '({}/{}) {} does not equal {}'.format(folder, filename, L1, _
                assert cefr == _cefr, '({}/{}) {} does not equal {}'.format(folder, filename, c

                with codecs.open(os.path.join(text_root, folder, filename), 'r', encoding='UTF-
                    text = fin.read().strip()

                records.append({
                    'WS': ws, 'L1': L1, 'CEFR': cefr, 'prompt_id': prompt_id, 'lookup_id': look
                    'essay_id': '{}/{}'.format(folder, filename)
                })

        df_in = pd.DataFrame.from_records(records)
        df_in.to_csv(os.path.join(path, 'all_essays.csv'), encoding='UTF-8', index=False)
```

Now we can look up the age, TOEIC, IELTS, TOEFL of the student

```
In [30]: meta = pd.read_csv(os.path.join(path, 'ICNALE_SW_V1.1_Infosheet.csv'), encoding='utf8')
         df_in = pd.read_csv(os.path.join(path, 'all_essays.csv'), encoding='utf8', dtype={'look
```

```
        df_in['student_id'] = df_in[['WS', 'L1', 'lookup_id']].astype(str).apply(lambda row: '_

        lil_meta = meta.drop([u'PTJ1 (wds)', u'PTJ2 (wds)', u'SMK1 (wds)',
            u'SMK2 (wds)', u'Self Ev', u'Sex', u'Grade/Degree',
            u'Yrs of Stay (< Yrs)', u'ENS Type', u'Major/ Occupation',
            u'Acad. Genre', u'VST', u'CEFR', u'INTM', u'INSM',
            u'INTM+INSM', u'INTM-INSM', u'Primary', u'Secondary', u'College',
            u'Inschool', u'Outschool', u'Listening', u'Reading', u'Speaking',
            u'Writing', u'NS', u'Pronunciation', u'Presentation', u'EssayW'], axis=1)

        df_out = df_in.merge(lil_meta, how='left', left_on='student_id', right_on='Code')

        # convert country codes
        l1_map = {'PHL': 'FIL', 'ENS': 'ENG', 'IDN': 'IND'}
        df_out.L1 = df_out.L1.apply(lambda l: l1_map.get(l, l))

        df_out.to_csv(os.path.join(path, 'all_essays_with_meta.csv'), encoding='UTF-8', index=F
```

In [32]:
```
df_in = pd.read_csv(os.path.join(path, 'all_essays_with_meta.csv'), encoding='utf8')
print(df_in.groupby('L1').size())
df_in.head()
```

```
L1
CHN    800
ENG    400
FIL    400
HKG    200
IND    400
JPN    800
KOR    600
PAK    400
SIN    400
THA    800
TWN    400
dtype: int64
```

Out[32]:

	CEFR	L1	WS	essay_id	lookup_id
0	A2_0	THA	W	ICNALE_SW_1.1_Text_W_THA_A2_0_N238/W_THA_SMK0_...	82
1	A2_0	THA	W	ICNALE_SW_1.1_Text_W_THA_A2_0_N238/W_THA_PTJ0_...	32
2	A2_0	THA	W	ICNALE_SW_1.1_Text_W_THA_A2_0_N238/W_THA_SMK0_...	344
3	A2_0	THA	W	ICNALE_SW_1.1_Text_W_THA_A2_0_N238/W_THA_PTJ0_...	72
4	A2_0	THA	W	ICNALE_SW_1.1_Text_W_THA_A2_0_N238/W_THA_PTJ0_...	169

	prompt_id	text	student_id
0	SMK0	Students should not smoke in the restaurants ...	W_THA_082
1	PTJ0	Yes, I do agree if student needs to do the pa...	W_THA_032

```
2      SMK0  Should the removal of all smoking in restaura...  W_THA_344
3      PTJ0  Part time job is to use free time to benefit ...  W_THA_072
4      PTJ0  It is important for college students to have ...  W_THA_169

         Code Country  Age  Test Score
0  W_THA_082     THA   18  ONET   45
1  W_THA_032     THA   23  ONET   50
2  W_THA_344     THA   20  ONET   30
3  W_THA_072     THA   19  ONET   56
4  W_THA_169     THA   20  ONET   35
```

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

```
True
```

```
In [7]: def isfloat(value):
          try:
            float(value)
            return True
          except ValueError:
            return False

        # load data
        df_in = pd.read_csv(os.path.join(path, 'all_essays_with_meta.csv'), encoding='utf8')
        df_in[df_in.Score.apply(lambda s: not isfloat(s))]
```

```
Out[7]:      CEFR   L1 WS                                              essay_id  \
        330   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        332   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
        336   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
        343   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
        351   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        355   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        359   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
        365   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        368   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        377   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        380   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
        381   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        387   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
        389   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
        400   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
```

```
407   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
411   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
415   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
424   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_PTJ0_...
426   B1_2  HKG  W  ICNALE_SW_1.1_Text_W_HKG_B1_2_N104/W_HKG_SMK0_...
1412  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_SMK0_...
1413  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_PTJ0_...
1414  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_PTJ0_...
1415  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_PTJ0_...
1416  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_SMK0_...
1417  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_SMK0_...
1418  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_PTJ0_...
1419  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_SMK0_...
1420  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_SMK0_...
1421  B1_2  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B1_2_N268/W_SIN_SMK0_...
...    ...  ... ..                                                ...
4428  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4429  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4430  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4431  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4432  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4433  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4434  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4435  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4436  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4437  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4438  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4439  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4440  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4441  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4442  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4443  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4444  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4445  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4446  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4447  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4448  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4449  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4450  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4451  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4452  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4453  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_PTJ0_...
4454  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4455  B2_0  SIN  W  ICNALE_SW_1.1_Text_W_SIN_B2_0_N132/W_SIN_SMK0_...
4920  B1_1  KOR  W  ICNALE_SW_1.1_Text_W_KOR_B1_1_N122/W_KOR_SMK0_...
4935  B1_1  KOR  W  ICNALE_SW_1.1_Text_W_KOR_B1_1_N122/W_KOR_PTJ0_...

      lookup_id prompt_id                                          text  \
```

330	94	SMK0	Should smoking be completely banned in the co...
332	49	PTJ0	Have you heard of the argument for university...
336	71	PTJ0	After getting success in the A-level exam, st...
343	13	PTJ0	Nowadays, university students have to study d...
351	13	SMK0	Smoking has a lot of disadvantages. Smoking j...
355	50	SMK0	I do not agree the following statement which ...
359	50	PTJ0	Recently, number of university students are h...
365	44	SMK0	Smoking is a habit of breathing the smoke of ...
368	49	SMK0	Should we ban smoking at all restaurants? No,...
377	77	SMK0	Have you ever experienced that the one sittin...
380	9	PTJ0	It's not uncommon for a university student to...
381	9	SMK0	On 1st January, 2007, the government announce...
387	94	PTJ0	Since some people need to focus on doing some...
389	78	SMK0	In several counties, smoking has been banned ...
400	77	PTJ0	Nowadays, it is not surprised for us to see t...
407	44	PTJ0	Nowadays, with the evolution of education sys...
411	71	SMK0	Every time when people are having their meals...
415	3	PTJ0	There are many students think that having a p...
424	78	PTJ0	I have been employed as a part-time student h...
426	3	SMK0	Recently, there are more and more people smok...
1412	33	SMK0	Close your eyes and imagine walking into a re...
1413	63	PTJ0	In my opinion, holding a part time job also t...
1414	144	PTJ0	What does part time job means to you? While t...
1415	111	PTJ0	I feel that college students should not get a...
1416	13	SMK0	In my opinion, smoking should be banned in al...
1417	107	SMK0	Since I was young, my parents have brought me...
1418	83	PTJ0	It is important for college students to have ...
1419	110	SMK0	There are two large groups of people in Singa...
1420	55	SMK0	Smoking should be banned at all restaurants i...
1421	160	SMK0	First of all, I am not a smoker. Hence I woul...
...	...	...	...
4428	145	PTJ0	I agree that it is important for college stud...
4429	139	SMK0	Agree or disagree? As a non-smoker, I feel st...
4430	50	SMK0	Considering that restaurants are places where...
4431	185	SMK0	He took a deep puff of his cigarette and exha...
4432	58	PTJ0	In my opinion, it is not that important for s...
4433	14	SMK0	I agree that smoking should be banned in all ...
4434	24	SMK0	As a non-smoker, it would be easy for me to a...
4435	24	PTJ0	I believe that it can be beneficial for colle...
4436	174	SMK0	Smoking is a habit that never ceases to amaze...
4437	149	SMK0	One of the main contributors to one of the mo...
4438	102	SMK0	I am strongly against smoking, so I agree wit...
4439	46	PTJ0	All students must first and foremost recogni...
4440	123	SMK0	To a large extent, I am supportive of a compl...
4441	49	SMK0	Smoking has been a socially acceptable habit ...
4442	94	SMK0	The long-term negative effects of smoking are...
4443	106	PTJ0	I am of the opinion that while it may be fina...
4444	175	PTJ0	I feel that it may be important for college s...

4445	93	PTJ0	I feel that having a part-time job for colleg...
4446	173	PTJ0	I disagree that it is important for college s...
4447	106	SMK0	At the risk of providing an extreme view, I a...
4448	45	SMK0	I strongly agree with this statement. Smoking...
4449	197	SMK0	Firstly, before we begin on banning smoking c...
4450	118	PTJ0	I agree with the assertion that having a part...
4451	166	SMK0	No I do not agree with the statement. While t...
4452	173	SMK0	I agree that smoking should be completely ban...
4453	66	PTJ0	In this day and age, there are many college s...
4454	130	SMK0	Being a non-smoker, the natural stand would t...
4455	132	SMK0	Smoking is a habit that is extremely detrimen...
4920	206	SMK0	I agree that smoking should be completely ban...
4935	206	PTJ0	I agree that it is important for college stud...

	student_id	Code	Country	Age	Test	Score
330	W_HKG_094	W_HKG_094	HKG	21	HKALE	D
332	W_HKG_049	W_HKG_049	HKG	20	HKALE	E
336	W_HKG_071	W_HKG_071	HKG	19	HKALE	D
343	W_HKG_013	W_HKG_013	HKG	18	HKALE	C
351	W_HKG_013	W_HKG_013	HKG	18	HKALE	C
355	W_HKG_050	W_HKG_050	HKG	18	HKALE	E
359	W_HKG_050	W_HKG_050	HKG	18	HKALE	E
365	W_HKG_044	W_HKG_044	HKG	20	HKALE	C
368	W_HKG_049	W_HKG_049	HKG	20	HKALE	E
377	W_HKG_077	W_HKG_077	HKG	21	HKALE	B
380	W_HKG_009	W_HKG_009	HKG	21	HKALE	D
381	W_HKG_009	W_HKG_009	HKG	21	HKALE	D
387	W_HKG_094	W_HKG_094	HKG	21	HKALE	D
389	W_HKG_078	W_HKG_078	HKG	22	Cambridg O Level	C6
400	W_HKG_077	W_HKG_077	HKG	21	HKALE	B
407	W_HKG_044	W_HKG_044	HKG	20	HKALE	C
411	W_HKG_071	W_HKG_071	HKG	19	HKALE	D
415	W_HKG_003	W_HKG_003	HKG	19	HKALE	E
424	W_HKG_078	W_HKG_078	HKG	22	Cambridg O Level	C6
426	W_HKG_003	W_HKG_003	HKG	19	HKALE	E
1412	W_SIN_033	W_SIN_033	SIN	21	A Level (General Paper)	S
1413	W_SIN_063	W_SIN_063	SIN	21	A Level (General Paper)	A
1414	W_SIN_144	W_SIN_144	SIN	19	O Level (Eng Lang)	B4
1415	W_SIN_111	W_SIN_111	SIN	23	A Level (General Paper)	B
1416	W_SIN_013	W_SIN_013	SIN	21	A Level (General Paper)	A
1417	W_SIN_107	W_SIN_107	SIN	19	O Level (Eng Lang)	A1
1418	W_SIN_083	W_SIN_083	SIN	18	O Level (Eng Lang)	C6
1419	W_SIN_110	W_SIN_110	SIN	23	A Level (General Paper)	D
1420	W_SIN_055	W_SIN_055	SIN	23	O Level (Eng Lang)	B4
1421	W_SIN_160	W_SIN_160	SIN	23	A Level (General Paper)	C
...	...	...	...	...	...	...
4428	W_SIN_145	W_SIN_145	SIN	20	A Level (General Paper)	A
4429	W_SIN_139	W_SIN_139	SIN	21	O Level (Eng Lang)	C5

```
4430  W_SIN_050  W_SIN_050     SIN   21  A Level (General Paper)     C
4431  W_SIN_185  W_SIN_185     SIN   20  A Level (General Paper)     B
4432  W_SIN_058  W_SIN_058     SIN   21  A Level (General Paper)     B
4433  W_SIN_014  W_SIN_014     SIN   21       O Level (Eng Lang)    B3
4434  W_SIN_024  W_SIN_024     SIN   19  A Level (General Paper)     A
4435  W_SIN_024  W_SIN_024     SIN   19  A Level (General Paper)     A
4436  W_SIN_174  W_SIN_174     SIN   21       O Level (Eng Lang)    B3
4437  W_SIN_149  W_SIN_149     SIN   21  A Level (General Paper)     A
4438  W_SIN_102  W_SIN_102     SIN   22  A Level (General Paper)     A
4439  W_SIN_046  W_SIN_046     SIN   24  A Level (General Paper)     A
4440  W_SIN_123  W_SIN_123     SIN   20  A Level (General Paper)     B
4441  W_SIN_049  W_SIN_049     SIN   22       O Level (Eng Lang)    A2
4442  W_SIN_094  W_SIN_094     SIN   22  A Level (General Paper)     A
4443  W_SIN_106  W_SIN_106     SIN   22  A Level (General Paper)     A
4444  W_SIN_175  W_SIN_175     SIN   22  A Level (General Paper)     A
4445  W_SIN_093  W_SIN_093     SIN   21  A Level (General Paper)     A
4446  W_SIN_173  W_SIN_173     SIN   22       O Level (Eng Lang)    A1
4447  W_SIN_106  W_SIN_106     SIN   22  A Level (General Paper)     A
4448  W_SIN_045  W_SIN_045     SIN   20       O Level (Eng Lang)    A2
4449  W_SIN_197  W_SIN_197     SIN   19  A Level (General Paper)     B
4450  W_SIN_118  W_SIN_118     SIN   21  A Level (General Paper)     B
4451  W_SIN_166  W_SIN_166     SIN   22  A Level (General Paper)     A
4452  W_SIN_173  W_SIN_173     SIN   22       O Level (Eng Lang)    A1
4453  W_SIN_066  W_SIN_066     SIN   19       O Level (Eng Lang)    B3
4454  W_SIN_130  W_SIN_130     SIN   22  A Level (General Paper)     B
4455  W_SIN_132  W_SIN_132     SIN   22  A Level (General Paper)     B
4920  W_KOR_206  W_KOR_206     KOR   21                Cambridge   PET
4935  W_KOR_206  W_KOR_206     KOR   21                Cambridge   PET

[436 rows x 13 columns]
```

In [2]:
```
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_with_meta.csv'), encoding='utf8')

# 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)

print('Splitting sentences')
# create records for every sentence
records = []
for i, row in df_in.iterrows():
    rec = {
        'dataset': 'ICNALE', 'prompt_id': row.prompt_id, 'essay_id': row.essay_id, 'L1':
        'score': np.nan, 'score_type': '', 'age': row.Age
    }

    if not row.CEFR.startswith('XX'):
        rec['student_level_CEFR'] = row.CEFR.split('_')[0]

    if not pd.isnull(row.Score):
        test = row.Test.strip()
        test_map = {
            'A Level (General Paper)': 'A_Level',
            'Cambridg O Level': 'O_Level',
            'O Level (Eng Lang)': 'O_Level',
            'TOEFL (PBT)': 'TOEFL',
            'TOEFL (iBT)': 'TOEFL'
        }
        test = test_map.get(test, test)

        # not enough of these to be useful
        if test not in ['UPCAT', 'SAT', 'NCAT', 'NCAE', 'NAT', 'IEPT', 'Cambridge']:
            rec['student_level_{}'.format(test)] = row.Score
```

```
            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, 'ICNALE_sentences.csv'), encoding='utf8', index=False)
```

```
Converting to ASCII
Splitting sentences
0 of 5600
1000 of 5600
2000 of 5600
3000 of 5600
4000 of 5600
5000 of 5600
Creating data frame
80171 sentences
Saving data frame
```

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

```
In [3]: df = pd.read_csv(os.path.join(path, 'ICNALE_sentences.csv'), encoding='utf8')

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


In [5]: df.columns

Out[5]: Index([u'L1', u'age', u'dataset', u'essay_id', u'prompt_id', u'score',
               u'score_type', u'sentence_id', u'student_level_A_Level',
```

```
       u'student_level_CEE', u'student_level_CEFR', u'student_level_CSEPT',
       u'student_level_HKALE', u'student_level_IELTS', u'student_level_NMET',
       u'student_level_ONET', u'student_level_O_Level', u'student_level_TEPS',
       u'student_level_TOEFL', u'student_level_TOEIC', u'text',
       u'trailing_whitespace', u'uid'],
      dtype='object')
```

```
In [12]: age = df.groupby('age').size()
         print(age)
         print('{} sentences with age data'.format(pd.notnull(df.age).sum()))
```

```
age
15      138
16     1588
17     3172
18    16779
19    20487
20    14405
21     9256
22     5321
23     2770
24     2048
25      980
26      700
27      324
28      536
29      342
30       66
31      112
32       96
33       67
34       82
35       47
36       83
37       97
38       76
39       32
40      144
41       74
42       33
43       34
44       17
45       81
46       19
50       15
51       21
52       58
54       18
```

```
57       12
59       41
dtype: int64
80171 sentences with age data


In [13]: score = df.groupby('score').size()
         print(score)

Series([], dtype: int64)


In [14]: df.text.apply(len).describe()

Out[14]: count    80171.000000
         mean        89.012349
         std         54.462845
         min          1.000000
         25%         53.000000
         50%         78.000000
         75%        112.000000
         max       1329.000000
         Name: text, dtype: float64

In [3]: df.groupby('student_level_CEFR').size()

Out[3]: student_level_CEFR
        A2    14732
        B1    55345
        B2     6371
        dtype: int64

In [4]: for col in df.columns:
            if not col.startswith('student_level_'):
                continue
            print(df.groupby(col).size())

student_level_A_Level
A    1139
B    1316
C     512
D     253
E      51
S      27
dtype: int64
student_level_CEE
58.0     64
59.0     60
60.0     63
```

```
61.0     31
62.0     38
63.0    162
64.0     33
65.0    122
66.0     28
67.0    299
68.0    266
69.0    122
70.0    160
71.0    122
72.0    242
73.0    209
74.0     28
75.0    170
76.0    173
77.0    144
78.0     59
79.0    480
80.0    251
81.0    256
82.0    101
83.0    243
84.0    119
85.0    392
86.0    127
87.0    305
88.0    148
89.0    116
90.0    272
91.0     81
92.0     23
93.0     96
94.0     86
95.0     26
97.0     27
dtype: int64
student_level_CEFR
A2    14732
B1    55345
B2     6371
dtype: int64
student_level_CSEPT
120.0    29
150.0    17
219.0    20
237.0    25
dtype: int64
```

```
student_level_HKALE
B     22
C     62
D     74
E    183
F     35
dtype: int64
student_level_IELTS
5.0     93
6.0    123
7.0    210
8.0     31
dtype: int64
student_level_NMET
128.0    29
129.0    24
138.0    32
140.0    26
143.0    28
dtype: int64
student_level_ONET
12.0     27
16.0     24
17.0     42
18.0     32
20.0    103
22.0     23
24.0     65
25.0     54
26.0    101
27.0    317
28.0     77
29.0    214
30.0    416
31.0    181
32.0    129
33.0     94
34.0    162
35.0    340
36.0    183
37.0    188
38.0    227
39.0    224
40.0    242
41.0    260
42.0    176
43.0    135
44.0    102
```

```
45.0    358
46.0     86
47.0    289
       ...
51.0    386
52.0    386
53.0    402
54.0    148
55.0    348
56.0    182
57.0    204
58.0    137
59.0     91
60.0    687
61.0    103
62.0    188
63.0    139
64.0    189
65.0    149
66.0     75
67.0    119
68.0     27
69.0     21
70.0    190
71.0     54
72.0     39
73.0     44
74.0    222
75.0     29
77.0     16
78.0     32
79.0     56
80.0     43
83.0     25
dtype: int64
student_level_O_Level
A1    178
A2    198
B      61
B3    377
B4    180
C5     52
C6    229
D      31
dtype: int64
student_level_TEPS
617.0    35
650.0    25
```

```
664.0    40
692.0    23
703.0    24
710.0    44
730.0    26
735.0    27
750.0    37
756.0    32
788.0    32
829.0    28
841.0    24
857.0    27
877.0    27
911.0    26
930.0    22
dtype: int64
student_level_TOEFL
56.0      39
63.0      26
70.0      31
71.0      32
77.0      37
80.0      29
83.0      31
85.0      48
87.0      28
89.0      53
90.0      25
93.0      38
94.0      46
96.0      61
97.0      28
98.0      71
100.0    110
101.0     53
102.0     57
103.0     61
104.0     30
105.0     90
109.0     58
110.0     56
114.0     36
119.0     30
300.0     19
328.0     11
373.0     34
412.0     35
423.0     37
```

```
425.0     25
442.0     33
457.0     32
460.0     74
473.0     64
487.0     25
488.0     23
490.0     22
493.0     23
497.0     28
507.0     27
510.0     33
527.0     24
530.0    169
550.0     29
560.0     27
dtype: int64
student_level_TOEIC
240.0     42
250.0     27
295.0     32
300.0     48
305.0     13
310.0     73
320.0     27
325.0     25
345.0    122
350.0     58
355.0     72
365.0     40
370.0     36
375.0     51
385.0     29
390.0     39
395.0    107
400.0    151
405.0     28
420.0    170
425.0     25
430.0     21
435.0    216
440.0    117
450.0    435
455.0     44
460.0    100
465.0    296
470.0    248
475.0    123
```

```
        ...
815.0     90
825.0     26
830.0    132
835.0     91
840.0    117
845.0     49
850.0     91
855.0     54
860.0    115
865.0     62
870.0     91
875.0     28
880.0     31
885.0     40
890.0     23
895.0     28
900.0     79
910.0     49
915.0     28
920.0     21
925.0     32
940.0     24
945.0     31
955.0     62
960.0     28
965.0     48
970.0     26
975.0     20
985.0     30
990.0     19
dtype: int64
```